UNIVERSIDADE DE VIGO

DEPARTAMENTO DE ENXEÑERÍA TELEMÁTICA
E.T.S.E. DE TELECOMUNICACIÓN

TESIS DOCTORAL

PROPUESTA METODOLÓGICA PARA EL RAZONAMIENTO SEMÁNTICO EN SISTEMAS DE RECOMENDACIÓN PERSONALIZADA Y AUTOMÁTICA. APLICACIÓN AL CASO DE CONTENIDOS AUDIOVISUALES

Autora: Yolanda Blanco Fernández

Directores: Dr. Alberto Gil Solla
Dr. Manuel Ramos Cabrer

2007

TESIS DOCTORAL

PROPUESTA METODOLÓGICA PARA EL RAZONAMIENTO SEMÁNTICO EN SISTEMAS DE RECOMENDACIÓN PERSONALIZADA Y AUTOMÁTICA. APLICACIÓN AL CASO DE CONTENIDOS AUDIOVISUALES

Autora: Yolanda Blanco Fernández

Directores: Dr. Alberto Gil Solla
Dr. Manuel Ramos Cabrer

TRIBUNAL CALIFICADOR

Presidente: Dr. Carlos Delgado Kloos

Vocales: Dr. Santiago Pavón Gómez

Dr. Juan Ramón Velasco Pérez

Dr. José Juan Pazos Arias

Secretaria: Dra. Rebeca P. Díaz Redondo

CALIFICACIÓN:

Vigo, a ____ de _____ de 2007

A mi padre, que ya no está.

A mi madre y a mi hermana, que siguen estando.

A Juan, sin quien no sabría ni estar ni seguir.

Agradecimientos

No sería justo abandonar el barco sin dar las gracias a todos los que me han acompañado a lo largo de esta travesía, tranquila en ocasiones y un tanto desapacible en otras.

En primer lugar, mi agradecimiento más sincero a Alberto, Manolo y Jose (ya sabéis cómo va eso de la conmutativa) por todo lo que me han enseñado y ayudado a que este momento llegara; por la confianza y el esfuerzo que han puesto en este trabajo sin pedir nada a cambio.

Gracias también al resto de compañeros del Departamento de Enxeñería Telemática y, muy especialmente, a Martín, Manu, Jairo, Pablo y Estela, sin quienes estos últimos años no hubiesen sido lo mismo (ay, si las paredes del 203-B hablaran...). Gracias también por alumbrarme en mis eclipses y por todo lo que nos queda por hacer juntos, ¡ánimo!

En el plano estrictamente personal no puedo olvidarme de los *míos*. Dicen los chinos que la vida te da un golpe primero para después tener una buena excusa para compensarte con ganas; ellos han sido, sin duda, el mejor premio. Gracias a mi padre, porque mis logros serán siempre los suyos, y a mi madre, por sus sacrificios, su entrega, su apoyo incondicional y porque, probablemente sin saberlo, me ha enseñado a seguir caminando en épocas difíciles.

Gracias también a mi hermana, por ejercer de ángel de la guarda en la sombra, por velar por la *peque* y por ser como es, y a mi cuñado, por la conversación y los consejos (sigo sin ser tía, en fin...).

Una vez más gracias a Juan, por su aliento, por su empeño en que vea el vaso medio lleno, por caminar a mi lado "sin cansarse", por compartir cada uno de mis sueños y, en definitiva, por todo lo que sólo él y yo sabemos; ¡ole tú!. Sigue haciendo mucho frío...

Gracias también a mi "hermano" Jorge (porque llamarlo amigo sería un despropósito), por pensar en voz alta conmigo, por la complicidad y por no cambiar ni en el espacio ni en el tiempo; no se puede ser más grande, Xvrxo.

Tampoco quiero olvidarme de mi otra familia —la de sangre vasco-coruñesa— que siempre me hacen sentir como en casa, ni de los amigos de verdad, que están aunque no los vea (ya sabéis quienes sois).

Ésta es también una ocasión perfecta para dar las gracias a los que me acompañaron durante el periplo finlandés: a Carlos y Olga, sin cuyo apoyo las cosas hubiesen sido aún más difíciles; a Jose, por nuestras sobremesas al más puro estilo español; a mi

---

Estos agradecimientos se hacen extensivos al Ministerio de Educación y Ciencia, ya que la financiación recibida a través del proyecto TSI2004-03677 ha permitido realizar este trabajo en las mejores condiciones, tanto en lo referente a medios materiales como a las posibilidades de participación en congresos internacionales.

vecino Mansour, un abuelete entrañable que me ayudó a sobrellevar tres meses muy largos a base de té, arroz y risas, muchas risas (“We’ll meet again; don’t know where; don’t know when; but I know we’ll meet again some sunny day”).

En definitiva, gracias a los que os alegráis tanto como yo de que hayamos llegado al destino de este viaje. Hala... marcho, que teño que marchar.

Resumen

El comercio electrónico, la Televisión Digital o la educación a distancia a través de medios telemáticos (*e-learning*, *t-learning*, *m-learning*) son algunos de los dominios de aplicación en los que actualmente se advierte una sobrecarga de información que dificulta la utilización de los servicios ofrecidos en los mismos. En este escenario, los usuarios deben explorar espacios excesivamente densos, convirtiendo la selección de la información que les interesa (artículos comerciales, contenidos de televisión, cursos educativos, etc.) en una tarea tediosa, difícil de asumir sin la asistencia de herramientas de búsqueda intuitivas y eficientes.

Con el afán de suplir esta carencia, surgen en los años 90 los sistemas recomendadores, capaces de seleccionar, de forma automática y personalizada, los productos que mejor se adaptan a las preferencias o necesidades de cada usuario. En la literatura se han propuesto diversas estrategias de personalización que los investigadores han adoptado de forma generalizada en sus trabajos. Estas estrategias son las encargadas de analizar las preferencias de los usuarios y elaborar sugerencias personalizadas a partir de los productos disponibles en la herramienta. En los enfoques tradicionales, este proceso se basa en técnicas de naturaleza sintáctica, más o menos sofisticadas, que prescinden de gran cantidad de conocimiento útil durante el proceso de personalización, lo que limita la calidad de las recomendaciones seleccionadas.

Limitaciones similares a éstas ya se han manifestado con anterioridad en los motores de búsqueda desarrollados en Internet, encargados de recuperar automáticamente la información solicitada por los usuarios. A pesar de su exitosa implantación, la calidad de los resultados de estas herramientas es cuestionable ya que, al basarse en estrategias de recuperación sintácticas, sólo son capaces de seleccionar la información que coincide *exactamente* con los términos de consulta especificados por los usuarios.

Con el ánimo de superar tal limitación y concebir una red *inteligente*, en los últimos años ha surgido una prometedora línea de investigación bajo el nombre de Web Semántica. Esta iniciativa propone describir los recursos Web mediante metadatos procesables por las máquinas, para que éstas puedan razonar sobre su semántica e inferir relaciones entre ellos, es decir, descubrir nuevo conocimiento a partir del ya conocido. Tales procesos de razonamiento semántico requieren que las colecciones de datos a los que acceden las máquinas estén definidas y estructuradas de una forma adecuada. Para este propósito, y aprovechando la experiencia ganada en el campo de la Inteligencia Artificial, la Web Semántica rescata la noción de ontología, una formalización consensuada y reutilizable en la que se identifican los conceptos y relaciones típicas en un dominio de aplicación.

Son precisamente los resultados obtenidos en este campo los que inspiraron el trabajo presentado en esta tesis. En concreto, este trabajo combate las limitaciones recono-

cidas en los sistemas recomendadores actuales, recurriendo a las técnicas de inferencia de conocimiento tradicionalmente adoptadas en la Web Semántica. Dichas técnicas se incluyen en una metodología, flexible y genérica, que razona sobre las preferencias de los usuarios (modeladas en perfiles personales) y sobre las descripciones semánticas de los productos disponibles en el recomendador (previamente formalizadas en una ontología de dominio). El objetivo es aprovechar la experiencia de las estrategias de personalización propuestas en el estado del arte e incorporar en las mismas este tipo de procesos de razonamiento semántico. Las estrategias resultantes exploran la base de conocimiento del sistema e infieren relaciones semánticas ocultas entre los productos allí representados. Tales relaciones aportan conocimiento adicional sobre las preferencias de los usuarios y, por tanto, favorecen procesos de personalización más precisos. Éstos no sólo permiten recomendar productos que pasarían desapercibidos en los sistemas actuales, sino que además combaten, de forma efectiva, las limitaciones reconocidas en las estrategias de personalización adoptadas en los mismos.

Dado que la metodología desarrollada en esta tesis no está vinculada de forma exclusiva a ningún dominio específico, su capacidad de razonamiento semántico puede ser explotada en aplicaciones de personalización de muy diversa naturaleza. Sin embargo, con el ánimo de describir el enfoque con suficiente claridad, este trabajo incorpora la metodología en el sistema recomendador AVATAR, una herramienta que selecciona, entre la miríada de contenidos audiovisuales que pueden estar disponibles en la Televisión Digital, aquéllos que potencialmente interesan a cada espectador.

Junto a las bases teóricas y los detalles algorítmicos de la metodología, nuestro trabajo también define la arquitectura de AVATAR. En ésta se adoptan tecnologías de libre disposición y formatos normalizados, con el afán de promover un uso generalizado de sus capacidades automáticas de personalización. Asimismo, el trabajo describe las herramientas que se han desarrollado para poder validar la metodología de razonamiento semántico desarrollada en esta tesis.

Palabras clave: Web Semántica, recomendadores automáticos, razonamiento semántico, Televisión Digital

Abstract

Various domains of application, like electronic commerce, Digital Television and distance learning over telematic media (*e-learning*, *t-learning*, *m-learning*) are currently characterized by an information overload that hampers the use of the services available in them. In this scenario, the users are forced into exploring excessively dense spaces, turning the selection of interesting information (e.g. commercial products, TV contents or educational courses) into a cumbersome task, extremely hard to tackle without the assistance of intuitive and efficient searching tools.

In order to overcome this lack, recommender systems arose in the 90s that were able to select, in an automatic and personalized way, the products that best match the preferences or the needs of each user. In literature, one can find multiple personalization strategies that have been massively adopted by researchers in their work. These strategies are responsible for analyzing the users'preferences and, thereupon, making personalized suggestions from the products available in the tool. In the traditional approaches, this process has been based on more or less sophisticated syntactic techniques, which miss a great amount of useful knowledge in the personalization process, which limits the quality of the selected recommendations.

Similar limitations have already shown up in the searching engines developed for the Internet, intended to automatically retrieve the information requested by the users. Despite their successful implantation, the quality of the results achieved by these tools is rather questionable, because, being based on syntactic retrieval strategies, they can only select the information that coincides *exactly* with the query terms specified by the users.

In order to solve this limitation and thus enable an *intelligent* network, a promising research line has been launched in the last few years under the denomination of Semantic Web. This initiative consists of describing the Web resources by means of machine-readable metadata, so that computers can reason about their semantics and infer relationships between them, that is, to allow discovering new knowledge from a set of facts. Such a semantic reasoning process requires the data collections to be defined and structured in a suitable way. For this purpose, and taking advantage of the experience gained in the field of Artificial Intelligence, the Semantic Web rescued the notion of ontology, as an agreed-upon reusable formalization used to identify typical concepts and relationships in a given domain of application.

The works in this field are precisely the ones that inspired this doctoral work. Particularly, this thesis tackles the aforementioned limitations of the current recommender systems by adopting the knowledge inference techniques employed in the Semantic Web. Those techniques are gathered into a flexible and generic methodology that reasons about the user preferences (modeled in personal profiles) and the semantic descrip-

tions of the products available in the recommender (previously formalized in a domain ontology). The goal is to take advantage of the experience gained from the personalization strategies found in the state-of-the-art, and to incorporate this kind of semantic reasoning processes onto them. The resulting strategies explore the system knowledge base and infer hidden semantic relationships between the products represented there. Such relationships provide additional knowledge about the user preferences, and thereby enable more accurate personalization processes. Not only do these processes allow recommending products that would go unnoticed for the current systems, but they also solve the known limitations of the current personalization strategies.

Since the methodology developed in this thesis is not at all linked to any specific domain, its semantic reasoning capabilities can be exploited in personalization applications of a very diverse nature. Nonetheless, in order to describe the approach clearly enough, this work illustrates the methodology within the AVATAR recommender system, a tool that selects potentially interesting contents for each user from among the myriad available in Digital Television.

Together with the theoretical basis and the algorithmic details of the methodology, our work also defines the architecture of AVATAR, which adopts freely-available technologies and normalized formats to promote the generalized use of its automatic personalization capabilities. Likewise, the work describes the tools that have been developed to validate the semantic reasoning methodology developed in this thesis.

Key words: Semantic Web, automatic recommenders, semantic reasoning, Digital Television

Índice general

Índice de figuras XI

Índice de tablas XIII

I Introducción 1

1. Ámbito y objetivos de la tesis 3
1.1. Motivación del problema a resolver 3
1.1.1. Los sistemas recomendadores 4
1.1.2. La Web Semántica 5
1.2. Descripción y objetivos de la tesis 6
1.3. La Televisión Digital 8
1.3.1. Descripción genérica 8
1.3.2. Necesidades de normalización 9
1.4. Organización de la memoria 11

II Estado del arte 13

2. Estado del arte en sistemas recomendadores 15
2.1. Introducción 15
2.2. Representación del perfil del usuario 16
2.2.1. Historiales de consumo 17
2.2.2. Vectores de características 17
2.2.3. Matrices de clasificaciones 17
2.2.4. Características demográficas 18
2.2.5. Redes semánticas ponderadas 19
2.2.6. Redes asociativas ponderadas 19
2.2.7. Modelos basados en clasificadores automáticos 20
2.3. Generación del perfil inicial 26
2.3.1. Reconocimiento automático de preferencias 26

2.3.2. Inicialización manual 26
2.3.3. Estereotipos 26
2.3.4. Conjuntos de entrenamiento 27
2.4. Realimentación de relevancia 27
2.4.1. Sistemas recomendadores sin realimentación 28
2.4.2. Realimentación explícita 28
2.4.3. Realimentación implícita 28
2.4.4. Realimentación híbrida 28
2.5. Técnicas para la adaptación del perfil de usuario 29
2.5.1. Actualización manual por parte de los usuarios 29
2.5.2. Incorporación de nueva información al perfil 29
2.5.3. Función de olvido gradual 29
2.5.4. Selección natural 30
2.6. Métodos de filtrado de información 30
2.6.1. Filtrado demográfico 30
2.6.2. Métodos basados en contenido 31
2.6.3. Filtrado colaborativo 34
2.6.4. Enfoques híbridos 40
2.7. Conclusiones 43

3. Estado del arte en Web Semántica 45
3.1. Introducción 45
3.2. Representación del conocimiento 46
3.2.1. ¿Qué es la representación de conocimiento? 46
3.2.2. Redes semánticas 46
3.2.3. Representación de conocimiento basado en marcos 47
3.2.4. Lógicas Descriptivas 47
3.3. Organización del conocimiento 48
3.3.1. Taxonomías 48
3.3.2. Mapas temáticos 49
3.3.3. Tesauros 50
3.3.4. Ontologías 51
3.4. Arquitectura de la Web Semántica 53
3.4.1. Capa XML 53
3.4.2. Capa RDF y RDF Schema 54
3.4.3. Capa de Ontologías 55
3.4.4. Capa Lógica 60
3.4.5. Capas de Comprobación y Confianza 62
3.5. Asociaciones semánticas complejas 62
3.5.1. Secuencias de propiedades 63

3.5.2. Relaciones entre secuencias de propiedades 63
3.5.3. Tipos de asociaciones semánticas complejas 64
3.6. Valoración de la relevancia de las asociaciones semánticas 65
3.6.1. Métricas semánticas 65
3.6.2. Métricas estadísticas 67
3.6.3. Criterio de clasificación global 69
3.7. SemRank: Clasificación de asociaciones semánticas 70
3.7.1. Métricas empleadas 71
3.7.2. Criterio de clasificación global 75
3.8. Aplicación de las asociaciones semánticas complejas 76
3.9. Técnicas de Propagación de Activación 77
3.9.1. Técnicas puras de propagación de activación 78
3.9.2. Técnicas de propagación de activación restringida 79
3.9.3. El algoritmo Branch-and-Bound 80
3.9.4. El algoritmo Hopfield Net 81
3.10. Búsqueda de información en la Web Semántica 83
3.10.1. Motores de búsqueda semánticos 84
3.10.2. Recuperación de información basada en similitud semántica . 91
3.10.3. Sistemas recomendadores 99
3.11. Sumario 103
3.11.1. Razonamientos semánticos básicos 103
3.11.2. Razonamientos semánticos avanzados 104

III Arquitectura del sistema AVATAR 107

4. Decisiones de diseño 109
4.1. Descripción genérica 109
4.2. Tecnologías adoptadas 111
4.2.1. La especificación MHP 111
4.2.2. La especificación TV-Anytime 112
4.2.3. La especificación OWL 114
4.3. Arquitectura del sistema AVATAR 115
4.3.1. Set-Top Box del usuario 115
4.3.2. Servidor del proveedor de servicios 117

5. Una ontología sobre el dominio de la TV 119
5.1. Lenguaje de implementación 119
5.2. Estructura de la ontología 120
5.2.1. Jerarquía de contenidos de TV 121
5.2.2. Características semánticas de los contenidos de TV 124

5.3. Información temporal 134
5.3.1. La ontología OWL-Time 134
5.3.2. Razonamiento temporal 137
5.4. Información geográfica 139
5.4.1. La ontología GEO 139
5.4.2. Razonamiento geográfico 140

6. Modelado de Usuarios 143
6.1. Perfiles de usuario basados en ontologías 143
6.2. Composición de los *perfiles-ontología* 145
6.2.1. Generación del perfil inicial 145
6.2.2. Realimentación de relevancia 146
6.3. Cálculo de los niveles de interés 150
6.3.1. Nivel de interés de los contenidos de TV 151
6.3.2. Nivel de interés de las características semánticas 152
6.3.3. Nivel de interés de las clases hoja 153
6.3.4. Propagación de los niveles de interés a través de la jerarquía . 155
6.4. Adaptación ante cambios en las preferencias de los usuarios 156
6.4.1. Función de olvido gradual basada en ventanas temporales . . 157
6.4.2. Actualización de los índices de olvido 158
6.4.3. Eliminación de preferencias obsoletas 161

IV Estrategias de recomendación 163

7. Estrategia basada en asociaciones semánticas 165
7.1. Introducción 165
7.1.1. Descripción genérica 165
7.1.2. El agente semántico 167
7.2. Filtrado de secuencias de propiedades 169
7.3. Intensidad semántica 171
7.3.1. Conceptos fundamentales 171
7.3.2. Componentes de la intensidad semántica 173
7.4. Criterio de filtrado: rango del umbral semántico 177
7.5. Asociaciones semánticas inferidas 179
7.5.1. Asociaciones ρ-path 179
7.5.2. Asociaciones ρ-join 180
7.6. Validación semántica 181
7.6.1. Selección del modelo de propagación 182
7.6.2. Selección del algoritmo de propagación 183
7.6.3. La red de propagación de activación 185

7.6.4. Ponderación de los enlaces de la red 186
7.6.5. El algoritmo Hopfield Net 190
7.6.6. Criterio de recomendación: rango del umbral de validación . . 192
7.6.7. Etapas del proceso de validación semántica 194
7.7. Cálculo de la relevancia de los contenidos sugeridos 195
7.8. Ejemplo de aplicación 195
7.8.1. Filtrado de secuencias de propiedades 197
7.8.2. Inferencia de asociaciones semánticas 201
7.8.3. Validación semántica 201
7.8.4. El razonamiento semántico vs. los enfoques convencionales . 204
7.9. Beneficios de la estrategia basada en contenido 204

8. Estrategia híbrida 207
8.1. Introducción 207
8.2. Fase de filtrado basado en contenido 208
8.2.1. Nivel de *matching* semántico 208
8.2.2. Similitud semántica 209
8.2.3. Criterio de recomendación: rango del umbral β_{Match_U} 217
8.3. Fase basada en filtrado colaborativo 218
8.3.1. Selección de usuarios con preferencias similares 219
8.3.2. Selección del tamaño del vecindario 221
8.3.3. Predicción basada en el vecindario del usuario 221
8.3.4. Criterio de recomendación: rango del umbral $\beta_{Pred_{U_c}}$ 222
8.4. Cálculo de la relevancia de los contenidos sugeridos 223
8.5. Técnicas de hibridación empleadas 223
8.6. Ejemplo de aplicación 225
8.6.1. Fase basada en contenido 225
8.6.2. Fase colaborativa 226
8.6.3. El razonamiento semántico vs. los enfoques convencionales . 228
8.7. Beneficios de la estrategia híbrida 228

V Conclusiones 231

9. Conclusiones y líneas futuras 233
9.1. Conclusiones 233
9.1.1. Ontología de TV 234
9.1.2. Perfiles de usuario 235
9.1.3. Estrategias de recomendación 236
9.1.4. Arquitectura 242
9.1.5. Implementación 242

9.2. Líneas de trabajo futuro 244

VI Apéndices 247

A. TV-Anytime 249

B. Implementación 275

B.1. Entorno de ejecución 275

B.2. Herramientas de soporte utilizadas 276

B.2.1. Implementación de la ontología de TV 276

B.2.2. Instanciación de la base de conocimiento 276

B.2.3. Inferencia mediante razonadores lógicos 278

B.2.4. Inspección de los razonamientos semánticos aplicados 278

B.3. Evaluación experimental 279

Índice de figuras

2.1. Funciones de activación utilizadas en las redes neuronales artificiales . 23

3.1. Arquitectura propuesta para la Web Semántica 53
3.2. El concepto de la refracción 74
3.3. Descomposición de matrices mediante la técnica SVD 90

4.1. Visión genérica de la arquitectura de AVATAR 110
4.2. Interfaz MHP entre aplicaciones y receptores digitales 111
4.3. Proceso de resolución de localización utilizado en TV-Anytime 113
4.4. Arquitectura del sistema AVATAR 116

5.1. Extracto de la jerarquía de géneros definida en la ontología OWL (ver Tabla 5.2) 124
5.2. Extracto de las clases e instancias referidas a alertas de contenido en la ontología OWL (ver Tabla 5.3) 127
5.3. Extracto de las clases e instancias referidas al tipo de audiencia de los contenidos en la ontología OWL (ver Tabla 5.7) 131
5.4. Extracto de las clases y propiedades de la ontología OWL-Time . . . 136
5.5. Ejemplo de razonamiento temporal (i) 138
5.6. Ejemplo de razonamiento temporal (ii) 138
5.7. Ejemplo de razonamiento temporal (iii) 139
5.8. Clases y propiedades definidas en la ontología OWL GEO 140
5.9. Ejemplo de razonamiento geográfico 141

6.1. Técnica de modelado de usuarios empleada en AVATAR 147
6.2. Instancias relativas al *perfil-ontología* P_X 149
6.3. Clases relativas al *perfil-ontología* P_X 150
6.4. Ventana temporal utilizada en la función de olvido gradual 157

7.1. Estructura del agente semántico representado en la Fig. 4.4 168
7.2. Secuencia de propiedades $ps_k^{x_i}$ definida en la ontología OWL de TV . 172
7.3. Construcción de una red de propagación a partir de secuencias de propiedades y asociaciones semánticas 187

7.4. El proceso de validación semántica 194
7.5. Perfil P_U 196
7.6. Extracto de la ontología de TV OWL 198
7.7. Secuencias de propiedades seleccionadas por el módulo de filtrado . . 201
7.8. Extracto de la red SA del usuario U 203

8.1. Estructura del agente híbrido representado en la Fig. 4.4 209
8.2. Instancia de unión entre los contenidos a y c_i 212
8.3. Algunas asociaciones semánticas utilizadas en el cálculo de la similitud semántica inferencial 213
8.4. Clase de unión entre los contenidos a y c_i 214
8.5. Ejemplo de razonamiento geográfico 216
8.6. Perfil P_U 225
8.7. Perfil del usuario V_1, vecino de U 227

Índice de tablas

2.1. Criterios empleados para clasificar los sistemas recomendadores . . . 16

3.1. Comparativa entre los algoritmos Branch-and-Bound y Hopfield Net . 83

5.1. Clasificación de géneros definidos en TV-Anytime (I) 122
5.2. Clasificación de géneros definidos en TV-Anytime (II) 123
5.3. Clasificación de alertas de contenido definidas en TV-Anytime 126
5.4. Clasificación de los créditos de los contenidos en TV-Anytime 128
5.5. Clasificación del origen técnico de los contenidos en TV-Anytime . . 128
5.6. Relaciones entre contenidos definidas en TV-Anytime 129
5.7. Clasificación del tipo de audiencia objetivo de un contenido 130
5.8. Clasificación de la intención de un contenido definida en TV-Anytime 132
5.9. Clasificación del formato estructural de un contenido en TV-Anytime 133
5.10. Clasificación de productos ofrecidos en contenidos comerciales definida en TV-Anytime 135

6.1. Niveles de interés de los contenidos definidos en el perfil P_X (Fig. 6.2) 152
6.2. Niveles de interés de las características semánticas incluidas en el perfil P_X 153
6.3. Niveles de interés de las clases hoja definidas en el perfil P_X (Fig. 6.3) 154
6.4. Propagación de niveles de interés en la jerarquía de géneros de P_X (Fig. 6.3) 156
6.5. Efecto de la ventana temporal sobre los índices DOI de las instancias del perfil del usuario 159
6.6. Efecto de la ventana temporal sobre los índices DOI de las clases del perfil 160
6.7. Actualización de los niveles de interés del perfil P_X 161

7.1. Factores de atenuación correspondientes a los enlaces virtuales representados en la Fig. 7.3 190
7.2. Secuencias de propiedades y asociaciones semánticas inferidas para U 202

8.1. Métodos de combinación de estrategias de recomendación empleados por el agente híbrido (ver Fig. 8.1) 224

A.1. Clasificación de géneros definidos en TV-Anytime 249
A.2. Clasificación de alertas de contenido 260
A.3. Clasificación de idiomas en TV-anytime 261
A.4. Clasificación de los créditos que participan en los contenidos 265
A.5. Clasificación del tipo de audiencia objetivo de los contenidos 267
A.6. Clasificación relativa al origen técnico de los contenidos 268
A.7. Clasificación relativa al formato estructural de los contenidos 269
A.8. Clasificación de las características emocionales asociadas a los contenidos 269
A.9. Clasificación de la temática de los contenidos comerciales 271
A.10. Clasificación relativa a la intención de los contenidos 273

Parte I

Introducción

Capítulo 1

Ámbito y objetivos de la tesis

"La Era de la Información es nuestra era. Es un periodo histórico caracterizado por una revolución tecnológica centrada en las tecnologías digitales de información y comunicación, concomitante con la emergencia de una estructura social en red, en todos los ámbitos de la actividad humana, y con la interdependencia global de dicha actividad."

***Manuel Castells**. Catedrático de Sociología en la Universidad de California.*

1.1 Motivación del problema a resolver

La llamada *Sociedad de la Información* es una consecuencia natural del rápido desarrollo que ha experimentado Internet desde su nacimiento en la década de los años 60. Este contexto plantea nuevos retos a los usuarios, quienes deben aprovechar la fuente de conocimiento que genera el nuevo modelo de sociedad y evolucionar de acuerdo a sus exigencias (tecnológicas).

En su lucha por alcanzar tales desafíos, los usuarios tropiezan con la profusa cantidad de información que se gestó inicialmente en la Red, y que ya se ha extendido a dominios de muy diversa naturaleza —véase el comercio electrónico, la Televisión Digital o la educación a distancia mediante medios telemáticos (*e-learning*, *t-learning*, *m-learning*). En este escenario, los usuarios deben explorar espacios excesivamente densos, convirtiendo la selección de la información que les interesa (artículos comerciales, contenidos de televisión, cursos educativos, etc.) en una tarea tediosa, difícil de asumir sin la asistencia de herramientas de búsqueda intuitivas y eficientes.

Precisamente con el afán de suplir esta carencia, surgen en los años 90 los sistemas recomendadores, mecanismos desarrollados en la comunidad de la Inteligencia Artificial que seleccionan, de forma automática y personalizada, los productos que mejor se adaptan a las preferencias o necesidades de cada usuario.

La utilización de los sistemas recomendadores en diversos dominios de aplicación ha revelado ciertas carencias importantes, que limitan la calidad de las sugerencias ofrecidas a los usuarios. El desarrollo de mecanismos efectivos que combatan este tipo de limitaciones es el eje principal que articula el trabajo presentado en esta tesis. A la hora de pergeñar nuestra solución, el enfoque desarrollado aprovecha la experiencia ganada en el campo de la Web Semántica, cuyos resultados son fruto del importante esfuerzo investigador que concentra esta área del conocimiento. Las bases fundamen-

tales de esta última, los beneficios de su utilización en el campo de la personalización y los objetivos planteados en la tesis, serán abordados en las siguientes secciones. Para concluir el índice de contenidos del capítulo, describimos el contexto de aplicación específico que adoptamos para ilustrar nuestro enfoque de personalización. En concreto, nuestro trabajo se centra en el dominio de la Televisión Digital, un medio en el que el proceso de digitalización habilita un incremento en la capacidad de transmisión y, por ende, la disponibilidad de una cantidad ingente de contenidos audiovisuales y servicios interactivos. Tal explosión de contenidos justifica la utilización de un sistema recomendador que asista a los espectadores durante la búsqueda de aquellos programas que, potencialmente, les interesan.

1.1.1 Los sistemas recomendadores

A la hora de valorar la relevancia de los productos disponibles para un usuario dado, los recomendadores tradicionales analizan sus preferencias —convenientemente modeladas en perfiles personales— mediante mecanismos más o menos sofisticados, no exentos de ciertas deficiencias.

- Los enfoques más sencillos establecen simples comparaciones sintácticas entre un conjunto predefinido de palabras claves; por tanto, adolecen de las limitaciones propias de los mecanismos que no razonan sobre el significado de los términos a través de su semántica (palabras con múltiples significados, sinónimos, etc.).
- Junto a estas propuestas, se han definido enfoques más elaborados basados en el empleo de clasificadores automáticos. Tales modelos deciden la relevancia de un determinado producto para el usuario, considerando ciertos patrones de ocurrencia de sus atributos o características en un conjunto de entrenamiento previamente definido.
- Asimismo, la revisión del estado del arte también nos conduce a enfoques que omiten las descripciones semánticas de los productos disponibles en el sistema. Éstos consideran como único criterio de recomendación los niveles de interés asociados a las preferencias del usuario, estableciendo que un producto es relevante para él si los usuarios del sistema que comparten sus preferencias también se han mostrado interesados en el mismo.

Todas las técnicas que hemos mencionado, incluidas en las estrategias de recomendación comúnmente adoptadas en los sistemas existentes [1], prescinden de gran cantidad de conocimiento útil durante el proceso de recomendación. Así, en los enfoques basados en clasificadores automáticos, los patrones detectados entre los atributos de los productos disponibles en el sistema, dependen únicamente de su frecuencia de aparición en el conjunto de entrenamiento y, en consecuencia, ignoran la semántica (es decir, el significado) asociada a los mismos. El potencial de esta semántica también se omite en los enfoques basados en los niveles de interés definidos en los perfiles de los usuarios y, de igual manera, en las propuestas más sencillas que consideran simples comparaciones sintácticas entre palabras clave. Esto no sólo limita la calidad de las sugerencias ofrecidas por los recomendadores actuales, sino que además es el origen de la mayoría de las debilidades identificadas en las estrategias de personalización adoptadas en los mismos.

Limitaciones análogas a las mencionadas en los sistemas recomendadores ya habían sido reconocidas con anterioridad en los motores de búsqueda desarrollados en Internet, encargados de resolver automáticamente las consultas de los usuarios, recuperando información relevante para las mismas. A pesar de la innegable utilidad de estas herramientas, avalada por su adopción generalizada por parte de los usuarios, sus resultados aún pueden ser mejorados de forma considerable. El principal obstáculo que conviene esquivar en este sentido es la naturaleza sintáctica de las estrategias utilizadas, que únicamente recuperan aquella información que coincide *exactamente* con los términos de búsqueda definidos en la consulta del usuario.

Con el objetivo de salvar tal limitación y alcanzar una red *inteligente*, en los últimos años ha surgido una prometedora línea de investigación bajo el nombre de Web Semántica, un término acuñado por el inventor de la Web y director del consorcio W3C Tim Berners-Lee en un artículo publicado en la revista Scientific American [19], en la que el autor indicaba:

"The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

1.1.2 La Web Semántica

La Web Semántica tiene como objetivo crear un medio universal para el intercambio de información, en el que se representa el significado de los recursos de la Web de forma que las máquinas puedan *entenderlo*. Con ello se pretende ampliar la interoperabilidad entre los sistemas informáticos y reducir la mediación de agentes humanos en la realización de tareas excesivamente tediosas.

A tal fin, la Web Semántica propone describir los recursos Web mediante metadatos procesables por las máquinas, para que éstas puedan razonar sobre su semántica e inferir relaciones entre ellos, es decir, descubrir nuevo conocimiento a partir del aprendido. Gracias a estas descripciones semánticas, los motores de búsqueda disponen de conocimiento adicional para relacionar los términos especificados por los usuarios con la información finalmente recuperada, superando con ello las comparaciones sintácticas utilizadas tradicionalmente.

Para poder explotar este tipo de capacidades inferenciales[1], es necesario que las colecciones de datos a las que acceden las máquinas estén definidas y estructuradas de una forma adecuada. Para este propósito, y aprovechando la experiencia ganada en el campo de la Inteligencia Artificial, la Web Semántica rescata la noción de ontología [61]; se trata de una formalización consensuada y reutilizable en la que se identifican los conceptos y relaciones típicas en un dominio de aplicación.

Precisamente para asegurar la compartición de conocimiento, los formatos utilizados a la hora de formalizar un dominio deben estar normalizados. En este sentido, se han definido varios lenguajes que se diferencian fundamentalmente en la capacidad expresiva que proporcionan. En estos lenguajes, los conceptos y relaciones de cada dominio de aplicación concreto se identifican mediante clases y propiedades, respectivamente, y se organizan en estructuras jerárquicas. Una vez definidas las clases y propiedades, es necesario representar también instancias específicas de las mismas,

[1] En este trabajo, se utilizan indistintamente los términos *inferencia* y *razonamiento* en alusión al descubrimiento de nuevo conocimiento.

resultando una base de conocimiento sobre la que es posible aplicar procesos de razonamiento automáticos.

Este tipo de razonamientos, realizados mediante herramientas lógicas, permiten comprobar la consistencia de la ontología, clasificar automáticamente sus instancias o inferir relaciones jerárquicas a partir de las ya formalizadas en la base de conocimiento. Junto a estas relaciones existen otras de naturaleza más compleja, ocultas en la ontología, cuyo descubrimiento requiere capacidades inferenciales más avanzadas que las incluidas en los razonadores lógicos usados actualmente en la Web Semántica.

1.2 Descripción y objetivos de la tesis

Conscientes de las limitaciones de los sistemas recomendadores actuales, nuestro trabajo [28] propone mejorar sus capacidades de personalización recurriendo a las técnicas de inferencia de conocimiento tradicionalmente vinculadas a la Web Semántica.

Para ello, desarrollamos una metodología, flexible y genérica, capaz de elaborar recomendaciones personalizadas y automáticas para los usuarios, basándose en estrategias de razonamiento semántico. El objetivo es descubrir relaciones semánticas entre las preferencias de los usuarios y los productos disponibles en el sistema, previamente formalizados en una ontología de dominio.

Tales relaciones son el aporte de conocimiento que requieren los recomendadores para superar las debilidades que los limitan actualmente (ver Sección 1.1.1). En este escenario, la riqueza de los procesos de razonamiento semántico redunda en la calidad de las sugerencias ofrecidas a los usuarios, de forma que cuanto más nutrido sea el conocimiento que infieren los primeros, mejores serán las recomendaciones elaboradas desde éste. Con el objetivo de garantizar dicha riqueza, las relaciones perseguidas en nuestro trabajo son fruto de una exploración concienzuda del conocimiento de la ontología, proceso que permite establecer vínculos interesantes entre las preferencias de los usuarios y los productos disponibles en el recomendador, omitidos en los enfoques sintácticos actuales.

Por su complejidad, tales relaciones no serían descubiertas mediante los razonadores lógicos usados en la Web Semántica que, lejos de explorar el conocimiento profundo de la ontología, soportan únicamente inferencias (lógicas) sencillas como las apuntadas en la sección previa. Tal situación nos obliga a desarrollar un mecanismo de razonamiento semántico propio, capaz de inferir este tipo de relaciones *complejas* y asegurar, a la vez, que conducen a la recomendación de productos potencialmente interesantes para el usuario.

Esta metodología no sólo descubre productos que, a pesar de su relevancia, pasarían inadvertidos en los sistemas recomendadores actuales —resultado, por otra parte, esperable a la luz de los trabajos realizados en la Web Semántica—, sino que además permite combatir las limitaciones reconocidas en las estrategias que adoptan estos sistemas, problemas para los que aún no se han propuesto soluciones efectivas en la literatura.

Objetivos de la tesis

Identificando los componentes de la metodología propuesta, podemos delimitar claramente los objetivos perseguidos en esta tesis:

- De acuerdo a lo comentado en la Sección 1.1.2, el proceso de razonamiento semántico aplicado en nuestro trabajo obliga al sistema recomendador a disponer de una base de conocimiento formal. Por esta razón, el primer objetivo del trabajo es el desarrollo de una **ontología** de dominio con capacidad expresiva suficiente para: (i) formalizar los conceptos y relaciones típicas en el dominio específico del sistema, y (ii) procesar este conocimiento mediante las capacidades inferenciales incorporadas en la metodología.

- Para satisfacer los requisitos de personalización del sistema recomendador, es necesario disponer de mecanismos adecuados para modelar las preferencias de los usuarios de forma precisa. Con el objetivo de razonar sobre su semántica, la metodología debe representar formalmente dichas preferencias, incluyendo descripciones detalladas sobre los productos que interesan (o no) a los usuarios. Asimismo, la técnica de modelado desarrollada debe almacenar los niveles de interés asociados a estos productos y, a la vez, asegurar que la metodología propuesta es capaz de adaptarse a medida que evolucionan las preferencias de los usuarios.

  Con el ánimo de satisfacer dichos requisitos, el enfoque presentado en este trabajo modela los perfiles de los usuarios de forma dinámica a partir de la ontología de dominio del sistema, rescatando así la capacidad de compartición de conocimiento propia de este tipo de conceptualización formal.

- Una vez formalizados los productos disponibles en el sistema y las preferencias de los usuarios (mediante la ontología de dominio y los perfiles, respectivamente), la metodología debe establecer correspondencias entre ambos elementos a través de **estrategias de personalización**.

  Para desarrollar estas estrategias, nuestro trabajo aprovecha la experiencia de las diferentes técnicas de recomendación adoptadas en los enfoques actuales, e incorpora en las mismas las capacidades de razonamiento semántico apuntadas al comienzo de la sección. Para ello, es necesario combinar dichas técnicas con algunos de los enfoques propuestos en el dominio de la Web Semántica y, más concretamente, en los campos de consulta de ontologías y exploración de grandes redes de conocimiento. Mediante las capacidades definidas en estas propuestas, nuestra metodología: (i) accede a las propiedades y relaciones jerárquicas explícitamente formalizadas en la ontología del sistema, (ii) infiere desde ellas las relaciones ocultas perseguidas en nuestro trabajo, y (iii) una vez representadas dichas relaciones en la ontología, explora de forma eficiente su conocimiento y extrae aquellas instancias (productos) que son relevantes para el usuario.

  Gracias a la ontología de dominio y a las relaciones ocultas inferidas en la misma, la metodología propuesta dispone de conocimiento adicional para vincular las preferencias de los usuarios con los productos del sistema que podrían ser interesantes para ellos. De esta forma, nuestro trabajo enriquece las capacidades de personalización de los sistemas recomendadores actuales, eliminando, a la vez, las limitaciones de las estrategias adoptadas tradicionalmente en los mismos.

- Amén de la labor teórica desarrollada en esta tesis, es necesario realizar una labor de **implementación** para comprobar la validez de la metodología de razonamiento semántico propuesta. Para ello, se implementa un prototipo en el que se incluyen los tres pilares de este trabajo (la ontología de dominio, la técnica de modelado de usuarios y las estrategias de recomendación), además de algunas

herramientas de soporte que permitan entender el tipo de relaciones semánticas inferidas, para así poder valorar su aportación y relevancia en el proceso de personalización.

Dado que la metodología desarrollada en esta tesis no está vinculada a un dominio determinado, su capacidad de razonamiento semántico puede ser explotada en aplicaciones de personalización de muy diversa naturaleza. Para facilitar la comprensión del lector e ilustrar el enfoque con suficiente claridad, en este trabajo se ha adoptado un contexto específico. En concreto, la metodología se aplica en el sistema recomendador AVATAR[2], una herramienta capaz de seleccionar, entre la miríada de contenidos audiovisuales disponibles en la Televisión Digital, aquéllos que potencialmente interesan a cada espectador.

1.3 La Televisión Digital

En los últimos años, la televisión tradicional ha evolucionado hacia un nuevo modelo en el que se transmiten en formato digital tanto contenidos audiovisuales como aplicaciones software. El proceso de transición hacia este nuevo modelo, denominado Televisión Digital (en adelante, DTV - *Digital Television*), viene motivado por el interés por aprovechar de forma más eficiente el espectro actualmente utilizado por la televisión analógica, por ampliar la oferta de canales, y por impulsar los nuevos servicios y facilidades accesibles a través de la tecnología digital. Si bien en España, desde el año 2004, se están simultaneando los dos tipos de transmisión, el reglamento actualmente en vigor (Real Decreto 944/2005 de 29 de Julio) establece que el 3 de abril del año 2010 se produzca el llamado *apagón analógico*, momento en el que las emisiones serán únicamente digitales.

1.3.1 Descripción genérica

Para recibir la señal de DTV, es requisito indispensable que los usuarios dispongan de un receptor externo (llamado *Set-top box* o STB) o bien de un televisor digital integrado. Actualmente, la solución más extendida entre los espectadores es la conexión directa de un STB al televisor analógico convencional, aunque en un futuro próximo, es de esperar que prolifere el uso de televisores integrados.

El proceso de digitalización utilizado en la DTV no sólo reporta beneficios en lo que a la calidad del audio y el vídeo se refiere, sino que también permite un aumento en la capacidad de transmisión gracias al uso de técnicas de compresión y multiplexación muy eficientes. Por una parte, al comprimir la señal de televisión, la tecnología digital permite transmitir entre tres y cinco canales en el mismo ancho de banda que ocuparía un único canal análogico; por otro lado, la naturaleza digital de la información transmitida introduce gran flexibilidad en los contenidos emitidos, siendo posible combinar un número arbitrario de canales de audio, vídeo y datos en una única señal.

El ahorro de ancho de banda es precisamente el origen de la sobrecarga de información acusada en el dominio de la DTV, en el que el incremento en la cantidad de canales ofrecidos se traduce también en un aumento de los contenidos disponibles para

[2] **A**sistente telemático a**VA**nzado para la búsqueda de con**T**enidos **A**udiovisuales mediante **R**azonamiento semántico

el espectador. Gracias a este nutrido número de canales y a la naturaleza temática de los mismos, los usuarios se conviertan en los auténticos programadores de su tiempo de ocio, adaptando los diferentes contenidos de entretenimiento, información o educación a sus necesidades o preferencias personales. En este escenario, los espectadores encuentran en los sistemas recomendadores un aliado perfecto, capaz de enfrentarse al ingente número de programas disponibles en la DTV, seleccionando, de forma automática y personalizada, aquéllos que resultan más interesantes para cada usuario.

Junto a estas ventajas, la irrupción de la DTV conlleva nuevas costumbres y cambios en los hábitos de los espectadores. Éstos dejan de sentarse pasivamente ante el televisor y asumen un rol más participativo [84, 133, 49, 68], gracias a la posibilidad de interactuar con las diferentes aplicaciones que se transmiten en la DTV junto a los contenidos audiovisuales tradicionales. En este nuevo escenario digital, es posible convertir el televisor en un terminal multimedia, de forma que los usuarios pueden recibir servicios como guías electrónicas de programación (EPGs) personalizadas de acuerdo a sus preferencias, vídeo bajo demanda, teletexto avanzado, correo electrónico o, incluso, participar activamente en concursos, encuestas o juegos mediante el mando a distancia.

Este tipo de aplicaciones requieren la existencia de recursos hardware y software en los receptores digitales, capaces de soportar tanto su correcta ejecución como su interacción con los usuarios. Tales facilidades interactivas y computacionales contribuyen, sin duda, a que la nueva televisión juegue un papel crucial como vehículo de acceso a la *Sociedad de la Información*, favoreciendo la adaptación de los usuarios a las exigencias tecnológicas apuntadas al comienzo del capítulo. Con ello, la DTV se perfila como una solución idónea en la lucha contra la preocupante *brecha digital* que se empieza a sentir en los países desarrollados como consecuencia de la limitada penetración de Internet en los hogares. Tal como reflejan los datos recuperados de *InternetWorld Stats*[3], el uso de Internet en este entorno es reducido (en torno al 52 % en la Unión Europea y al 70 % en Estados Unidos) y, lo que aún es peor, su tasa de crecimiento ha disminuido en los últimos años (desde el 29 % en 2001 hasta el 23 % en 2006 en Europa, y desde el 14 % al 9 % en Estados Unidos).

Basándonos en estos datos, así como en la presencia de los receptores de televisión en prácticamente todos los hogares en los países desarrollados, podemos concluir que la DTV es un medio adecuado para que los usuarios se familiaricen con las tecnologías propias del dominio de Internet, a través de un dispositivo de sobra conocido para ellos.

1.3.2 Necesidades de normalización

Con independencia de las ventajas que la tecnología digital ofrece a los espectadores de televisión, la característica más relevante de la evolución hacia el sistema digital es la aparición de un gran número de tecnologías en un campo caracterizado tradicionalmente por su estabilidad. Este mercado estable se enfrenta ahora a cambios significativos, principalmente derivados de la introducción de los servicios y aplicaciones software interactivas mencionadas antes. Es obvio que en este dominio la normalización jugará un papel clave, tratando de promover el uso extendido de productos compatibles que resulten atractivos para los usuarios.

Hoy por hoy, la principal organización reguladora en este campo es el consorcio

[3] Fuente consultada en *http://www.internetworldstats.com*

DVB[4] (*Digital Video Broadcasting*). Desde su creación en 1993, DVB ha definido un conjunto de estándares abiertos, empleados profusamente en todo el mundo, orientados a la difusión de señales de vídeo y servicios interactivos sobre las distintas redes de transmisión, incluyendo el satélite, el cable o la difusión terrena. Más tarde, los objetivos del consorcio DVB se ampliaron hasta alcanzar el desarrollo de una especificación que tratase de combatir la estructura vertical que se había definido en el mercado. En este escenario, la falta de normalización provocó que cada operador adoptase su plataforma de servicios interactivos de forma independiente, comprometiendo así la interoperabilidad de sus aplicaciones con los diferentes receptores digitales existentes.

Precisamente con el afán de suprimir esta implementación propietaria, nace MHP (*Multimedia Home Platform*) [65]. Esta norma, basada en tecnologías de libre disposición, define una interfaz genérica capaz de garantizar la ejecución de las aplicaciones interactivas en cualquier receptor compatible (STBs, PCs multimedia, televisores digitales integrados, etc.), con independencia de sus recursos hardware y software específicos. La estructura horizontal que promueve esta especificación favorece nuevas reglas de negocio que facilitan el impulso y la penetración de la DTV en el mercado actual. Tales reglas conllevan implicaciones importantes para los diversos elementos implicados: por una parte, la proliferación de más y mejores servicios interactivos, junto a la rápida evolución de la tecnología, favorece a los proveedores de servicios, quienes ven cómo se reducen los costes de sus desarrollos gracias a la economía de escala; por otro lado, los usuarios no sólo aprovechan la reducción en el precio de los receptores, sino que también gozan de libertad plena para abonarse a cualquiera de los operadores disponibles en el mercado, gracias a la interoperabilidad conferida a las aplicaciones MHP.

Otra de las iniciativas vinculadas al dominio de la televisión es IPTV (*Internet Protocol Television*), basada en la integración de Internet y la televisión, dos medios que, lejos de posicionarse como rivales, han alcanzado una situación de complementariedad. A modo general, IPTV consiste en distribuir las señales de televisión a través de una conexión de banda ancha sobre el protocolo IP. Esta iniciativa surge como resultado del gran auge de las conexiones de Internet y la evolución tecnológica que ha permitido ofrecer mayor ancho de banda a menor precio. IPTV reporta beneficios cuantiosos tanto para los usuarios como para los operadores que ofrecen los servicios: los primeros acceden a una amplia oferta de contenidos, mientras que los segundos pueden aprovechar las infraestructuras existentes a la hora de brindar sus contenidos, generando así mayores ingresos.

A diferencia de la televisión digital convencional, en IPTV el proveedor no emite sus contenidos a la espera de que el usuario se conecte, sino que éstos llegarán sólo cuando el cliente los demande. La ventaja más atractiva de esta iniciativa es, pues, la personalización del contenido, sobre la que se asientan las bases de una televisión *a la carta* donde el usuario final selecciona los programas que desea ver o descargar para almacenar en su receptor y, de esta manera, poder visualizarlos cuando desee.

Otra de las iniciativas de normalización más importantes en el dominio de la DTV es TV-Anytime [38]. Esta especificación, en la que participan algunas de las organizaciones más relevantes de la industria de creación y difusión de contenidos multimedia, normaliza un formato de datos común para describir contenidos audiovisuales genéricos, instancias concretas de los programas, perfiles de usuario, información de segmentación de contenido y mecanismos para referenciar los programas con independencia

[4] Más información disponible en *www.dvb.org*.

de su localización y hora de emisión.

Para concluir este apartado dedicado a la descripción del panorama tecnológico en el dominio de la DTV, mencionamos los esfuerzos que se están dedicando actualmente al desarrollo de la TV móvil, cuyas capacidades permitirán acceder a este medio en cualquier parte a través de terminales portátiles (en terminología inglesa *handheld*).

Son dos las iniciativas que luchan por desarrollar y desplegar la tecnología necesaria para el establecimiento de esta TV ubicua: DVB-H (*Digital Video Broadcasting - Handheld*) [121], secundada por los proveedores tradicionales de TV y MBMS (*Multimedia Broadcast and Multicast Services*) [146], iniciativa defendida por la industria de las telecomunicaciones móviles. Aunque esta última adolece de limitaciones severas relacionadas con el número de canales de TV disponibles (dado que comparte ancho de banda para voz y datos) representa la aproximación más barata, ya que está basada en una infraestructura existente de redes 3G y, por tanto, no necesita desarrollar una nueva red dedicada como ocurriría en DVB-H. Sin embargo, si el factor decisivo es la disponibilidad inmediata, la especificación DVB-H se perfila como clara ganadora, dado que viene avalada por las múltiples pruebas que ya se han llevado a cabo tanto en Estados Unidos como en el continente europeo (Finlandia, Reino Unido, Alemania, Francia, Italia, España, etc.).

1.4 Organización de la memoria

Esta memoria se organiza con arreglo a la siguiente estructura:

- Parte I: En esta primera parte se incluye el presente capítulo, en el que se avanza el planteamiento genérico del trabajo desarrollado en la tesis.
- Parte II: A continuación, se revisa el estado del arte en el campo de los sistemas recomendadores y en el dominio de la Web Semántica. En concreto, en el capítulo 2 se establece una clasificación de los recomendadores propuestos en la literatura, haciendo especial hincapié en las técnicas más extendidas, sus ventajas y limitaciones. Asimismo, se describen sistemas recomendadores reales en los que dichas técnicas se han implantado con gran éxito, y se analizan los métodos propuestos en el área del modelado de usuarios. En el capítulo 3, ya en el contexto de la Web Semántica, se revisan diferentes propuestas de razonamiento automático, búsqueda y recuperación de información desde ontologías de conocimiento.
- Parte III: Antes de abundar en los detalles de la metodología de razonamiento desarrollada, se describen (en el capítulo 4) las principales decisiones de diseño del sistema AVATAR, así como las tecnologías normalizadas adoptadas en su arquitectura. Ya en el capítulo 5 se presenta la ontología sobre el dominio de la DTV utilizada en este sistema recomendador, en cuya formalización se emplean las descripciones semánticas proporcionadas por la especificación TV-Anytime. En el capítulo 6, se describe la técnica de modelado de usuarios propuesta en este trabajo, analizando tanto la estructura como el proceso de construcción de los perfiles a partir del conocimiento de la ontología de TV.
- Parte IV: Las bases teóricas y los detalles algorítmicos de las estrategias de recomendación incluidas en la metodología se presentan en los capítulos 7 y 8, respectivamente.

- Parte V: Finalmente, en el capítulo 9 se describen las principales conclusiones extraídas del trabajo de esta tesis y se apuntan algunas líneas de investigación futura.
- Parte VI: Antes de enumerar la bibliografía revisada, se incluyen dos apéndices: el primero establece una clasificación de los atributos semánticos definidos en la especificación TV-Anytime; el segundo describe los detalles de implementación de las herramientas utilizadas en la validación de la metodología de razonamiento semántico desarrollada.

Parte II

Estado del arte

Capítulo 2

Estado del arte en sistemas recomendadores

En este capítulo se revisan las propuestas definidas en la literatura en el campo de los sistemas recomendadores. Dicha revisión incluye: (i) las técnicas de modelado de usuarios utilizadas para crear sus perfiles y actualizarlos a partir de la realimentación proporcionada ante una sugerencia del sistema, (ii) los mecanismos para asegurar la adaptación del recomendador a medida que dichas preferencias evolucionan en el tiempo, y (iii) las estrategias de personalización comúnmente adoptadas en este tipo de herramientas.

2.1 Introducción

La necesidad de herramientas de personalización se ha manifestado claramente, desde hace ya mucho tiempo, en numerosos dominios de aplicación, de ahí la gran diversidad de enfoques que han ido surgiendo en el campo de los sistemas recomendadores en los últimos años [1, 163]. En este capítulo establecemos una clasificación completa de los mismos, destacando las técnicas de personalización más relevantes así como sus ventajas y limitaciones. Además, identificamos diferentes sistemas de recomendación reales, implantados con gran éxito en aplicaciones comerciales vinculadas a dominios de naturaleza muy diversa.

En adelante, emplearemos el término *usuarios activos* para referirnos a los individuos que reciben las sugerencias personalizadas del recomendador. Por su parte, aquellos productos que el sistema deba decidir si son o no adecuados para esta comunidad de usuarios, reciben el nombre de *productos objetivo*.

Los perfiles de los usuarios activos son elementos cruciales en el sistema recomendador, dado que permiten conocer las preferencias o necesidades en las que se basan las sugerencias elaboradas. Partiendo de esta premisa, nuestra revisión del estado del arte clasifica los recomendadores existentes atendiendo a dos criterios genéricos que giran en torno a dichos perfiles, y en los que es posible identificar categorías más específicas, tal como se muestra en la Tabla 2.1:

1. En primer lugar, analizamos diferentes métodos tanto para generar los perfiles de los usuarios como para mantenerlos a lo largo del ciclo de vida del recomen-

dador, entre los que destacamos:

a) mecanismos de representación de perfiles de usuarios,
b) métodos para la creación de perfiles iniciales, y
c) técnicas para que los usuarios puedan proporcionar al sistema información de relevancia sobre la precisión de las sugerencias ofrecidas.

2. Por otra parte, nos centraremos también en las estrategias propuestas en la literatura para elaborar sugerencias personalizadas a partir de los perfiles de usuario. Esta tarea está estrechamente relacionada con dos tipos de mecanismos:

   a) los métodos de filtrado existentes para seleccionar los productos más adecuados para cada usuario activo de acuerdo a sus gustos o necesidades, y
   b) los mecanismos necesarios para asegurar que dichos métodos de filtrado se adaptan a los posibles cambios que se produzcan en las preferencias de los usuarios, garantizando con ello la calidad de las sugerencias ofrecidas.

CLASIFICACIÓN DE SISTEMAS RECOMENDADORES	
Generación y mantenimiento del perfil de usuario	Utilización del perfil de usuario
a) Representación del perfil de usuario b) Generación del perfil inicial c) Realimentación de relevancia	a) Métodos de filtrado de información b) Técnicas para la adaptación del perfil

Tabla 2.1: Criterios empleados para clasificar los sistemas recomendadores

2.2 Representación del perfil del usuario

Construir perfiles representativos es una tarea clave para asegurar la precisión en el proceso de personalización, consiguiendo con ello reforzar la confianza del usuario en el sistema recomendador. Han sido muchos los enfoques propuestos en diferentes dominios de aplicación para representar los perfiles de usuario, tales como:

- historiales de consumo,
- vectores que incluyen las características de los productos que identifican las preferencias del usuario,
- matrices que almacenan la valoraciones realizadas por los usuarios (es decir, sus niveles de interés) en relación a los productos disponibles en el sistema,
- características demográficas,
- redes semánticas,
- redes asociativas, y
- modelos basados en clasificadores, como redes Bayesianas, redes neuronales, árboles de decisión y reglas de aprendizaje inductivo.

2.2.1 Historiales de consumo

Son varios los métodos que podemos identificar dentro de esta categoría, dependiendo del dominio de aplicación concreto de cada sistema recomendador. Así, algunas herramientas mantienen registros históricos de las páginas Web visitadas por el usuario recientemente, mientras otras almacenan listas de los productos que ha consumido en el pasado o, incluso, los mensajes de correo electrónico que ha recibido. Además de esta información, estos enfoques almacenan típicamente el nivel de interés del usuario en relación a cada uno de los productos registrados en su perfil personal.

Entre los sistemas reales que han implantado esta técnica de modelado basada en registros históricos, podemos destacar Amazon.com [4] y CDNow.com [44], dos de los recomendadores más exitosos y populares en el campo del comercio electrónico. Otro enfoque similar ha sido adoptado en el sistema WebSell [57], en el que los perfiles están formados por dos listas, una que contiene los productos que el usuario ha comprado y que son relevantes para él, y una segunda lista relativa a aquellos que no le interesan. También en el dominio del filtrado de noticias es posible identificar sistemas basados en históricos de consumo, como ACR News [158]. Por último, destacar el sistema Tapestry [79], encargado de filtrar de forma automática los correos electrónicos recibidos por el usuario activo, basándose en los mensajes que éste ha aceptado antes y en sus anotaciones personales.

2.2.2 Vectores de características

El origen de esta técnica de modelado está vinculado al dominio de la recuperación de información en la Web. En concreto, cada documento consultado por el usuario activo se representa mediante un vector de características, normalmente conceptos o palabras, que tienen un valor asociado. Típicamente, dicho valor es una variable *booleana* o un número real. El valor *booleano* representa la presencia o ausencia de la característica en cuestión en el documento, mientras que el número real se refiere a la frecuencia de aparición o relevancia de dicha característica.

Este enfoque se utiliza en el sistema Webmate [48], una herramienta que recupera aquellos documentos que son relevantes para las consultas formuladas por el usuario activo. El principio fundamental de esta propuesta establece que dos documentos son similares cuando sus respectivos vectores de características son muy parecidos. En concreto, cada componente del vector que representa un documento en Webmate contiene pares de palabras y pesos reales asociados, tal como mencionamos antes.

Enfoques similares han sido adoptados en multitud de sistemas de personalización, como los recomendadores Web Amalthea [168], Let's Browse [137], WebWatcher [9, 109], Letizia [136] y Labour [220], o los sistemas de entrega de noticias personalizadas NewsWeeder [130], NewT [227], Beehive [103] o Anatagonomy [112]. También en el dominio de la TV podemos encontrar recomendadores que optan por modelado de usuarios basados en vectores de características, como los sistemas CBCF [148], Nakif [76], Movielens [150], Recommendation Explorer [69] y Virtualrev's [240].

2.2.3 Matrices de clasificaciones

Algunos sistemas recomendadores representan los perfiles de los usuarios como matrices. En concreto, dichas matrices tienen tantas filas como usuarios haya en el

sistema y tantas columnas como productos puedan ser objeto de clasificación. Así, un elemento genérico (u_i,p_j) puede tomar dos posibles valores dependiendo de si el usuario u_i ha clasificado o no el producto p_j: en caso afirmativo, dicho valor coincide con el nivel de interés de este usuario en relación al producto p_j; en caso contrario, toma un valor nulo debido a que u_i no conoce ese producto.

Entre los sistemas que utilizan este enfoque, destacamos el recomendador de películas Bellcore Video Recommender [94] y los sistemas Ringo [224, 225] y Smart Radio [91, 92], que ofrecen listas de música personalizadas de acuerdo a los gustos musicales de los usuarios activos.

2.2.4 Características demográficas

En algunos sistemas recomendadores, los perfiles de los usuarios no son más que registros que contienen sus datos personales (por ejemplo, edad, sexo, número de hijos, estado civil, profesión, etc.). Estas características demográficas se utilizan para identificar tanto el tipo de usuario que ha llegado al sistema, como las recomendaciones que han interesado a otros usuarios con datos similares a los suyos [118].

La principal ventaja de esta técnica es que no es necesario aplicar ningún mecanismo para aprender conocimiento sobre las preferencias de los usuarios, ya que ellos mismos proporcionan esta información de forma explícita. Sin embargo, esta ventaja también se ha convertido en uno de los principales obstáculos que dificultan la aplicación de este método en un sistema real, ya que los usuarios no siempre están dispuestos a revelar datos de carácter personal, ni a rellenar largos y tediosos formularios para empezar a utilizar el sistema.

Un enfoque basado en esta técnica de representación de perfiles de usuario ha sido adoptado en LifeStyle Finder [125], un sistema que genera perfiles personales *inteligentes* a partir de las características demográficas de varios usuarios. La mencionada capacidad de inteligencia es debida a que, a partir de los datos almacenados, el sistema es capaz de detectar patrones entre las características de varios usuarios y utilizarlos para inferir nueva información sobre los restantes.

En el dominio de la TV personalizada, podemos destacar el sistema Blu-IS descrito en [24]. Esta herramienta modela las preferencias de los usuarios complementando las características demográficas con dos elementos adicionales. El modelo resultante consta de los siguientes componentes:

- Perfil del usuario: captura información personal sobre éste (edad, sexo, profesión, idiomas que conoce, posibles discapacidades, etc.), así como descripciones semánticas de los contenidos que le interesan.
- Información contextual: registra datos como la localización del usuario, la franja horaria a la que prefiere ver televisión, su estado anímico, el número de personas que le acompañan, etc. Esta información permite que el sistema se adapte a la situación actual del usuario durante el proceso de personalización.
- Historial de visionado: contiene una relación de los programas que han interesado al usuario en el pasado.

En lugar de sugerir contenidos de TV personalizados al usuario de una forma automática y sin solicitud previa por parte del mismo, el sistema Blu-IS obliga al espectador

a formular consultas específicas, basadas en palabras clave. En este escenario, la herramienta devuelve un conjunto de contenidos de TV relacionados con estos términos. Con el fin de mejorar la calidad de las sugerencias ofrecidas, Blu-IS refina las consultas de los usuarios añadiendo a los términos que han especificado inicialmente, una serie de conceptos relacionados con éstos, extraídos desde una base de conocimiento.

Una vez refinada dicha consulta, el enfoque propuesto considera la información completa incluida en el modelo de usuario (tanto las preferencias como la información demográfica), para así poder recuperar aquellos contenidos que son más relevantes para el mismo. En primer lugar, la estrategia de recuperación de contenidos establece comparaciones entre los términos (conceptos) incluidos en la consulta y los atributos semánticos de los programas que interesan al espectador (definidos en su perfil personal). Una vez seleccionados estos contenidos, el sistema los ordena de acuerdo a su relevancia para el usuario, considerando tanto sus características demográficas como su contexto actual.

2.2.5 Redes semánticas ponderadas

Las redes semánticas fueron propuestas por Quillian en el año 1968 [199], y desde entonces se han convertido en un elemento clave en el campo de la representación de conocimiento. Según la propia definición de Quillian, las redes semánticas son sistemas que permiten expresar conocimiento, sirviéndose, principalmente, de conceptos, propiedades y relaciones jerárquicas. La descripción de este tipo de estructuras será retomada en el capítulo 3 de la tesis, donde nos centraremos en destacar su aplicación y beneficios en el dominio de la Web Semántica.

Este tipo de enfoque puede ser de gran utilidad en el dominio del modelado de usuarios, de ahí que algunos sistemas recomendadores opten por las redes semánticas como técnica de representación de sus perfiles personales. Para ello, los conceptos que identifican las preferencias de cada usuario activo se representan como nodos en la red semántica, mientras que las relaciones establecidas entre ellos se modelan mediante enlaces convenientemente etiquetados. A estos elementos se asocian también pesos numéricos que identifican el nivel de interés concreto del usuario (de ahí el nombre de redes semánticas ponderadas).

Uno de los dominios donde este tipo de enfoques se ha extendido en mayor medida es el de los recomendadores Web, sistemas que ofrecen sitios a los usuarios en función de las páginas que hayan consultado recientemente. Así, por ejemplo, los perfiles usados por el sistema SiteIF [234] se modelan mediante redes semánticas en las que los nodos son los términos presentes en las páginas visitadas por el usuario, y en los que se establecen enlaces entre dos nodos cuando los términos que representan aparecen en una misma página en su historial de navegación. Tal como dijimos antes, cada nodo tiene asociado un peso numérico, referido al nivel de interés del usuario activo. Un enfoque similar ha sido utilizado en el recomendador ifWeb [153, 10].

2.2.6 Redes asociativas ponderadas

Al igual que las redes semánticas, las redes asociativas representan conceptos e identifican las relaciones existentes entre ellos, sirviéndose de un conjunto de nodos y enlaces, respectivamente. A diferencia de las primeras, los enlaces de una red asociativa no tienen que estar necesariamente etiquetados. De hecho, en algunos casos, esta

etiqueta semántica se sustituye por un peso numérico que cuantifica la *fuerza* de la relación existente entre los dos nodos conectados (de ahí el nombre de redes asociativas ponderadas).

Si bien este tipo de redes están íntimamente ligadas al campo de la recuperación de información, también han sido utilizadas como técnica de modelado de usuarios en los sistemas recomendadores. Así, por ejemplo, el recomendador Web INFOrmer [208] representa los perfiles de los usuarios activos mediante redes asociativas, en las que los nodos identifican los términos presentes en las páginas que éstos han visitado. A la vez, para identificar frases relevantes en las que están incluidos estos términos, este enfoque une entre sí los nodos que los representan a través de enlaces ponderados. Los pesos asignados serán mayores cuanto más significativa sea la relación entre las palabras conectadas, por ejemplo, cuantas más veces aparezcan estos términos a la vez en los documentos consultados por el usuario.

2.2.7 Modelos basados en clasificadores automáticos

Los clasificadores son modelos computacionales capaces de asignar categorías específicas a cada una de las entradas que se les presentan. En el caso de los sistemas de personalización basados en clasificadores, dichas entradas identifican información sobre los productos objetivo y las preferencias de los usuarios activos, mientras que las categorías de salida deciden si estos productos deben ser sugeridos a dichos usuarios, clasificándolos como *interesantes* o *no interesantes*.

A fin de aprovechar tales capacidades predictivas durante el proceso de recomendación, algunos enfoques optan por modelar las preferencias de los usuarios mediante clasificadores automáticos. Éstos están basados en diferentes técnicas propuestas en el campo del aprendizaje de máquinas, tales como redes Bayesianas, árboles de decisión, reglas de aprendizaje inductivo o redes neuronales.

Redes Bayesianas

Una red Bayesiana es un grafo acíclico dirigido [64] en el que los nodos representan variables proposicionales y los arcos identifican dependencias causales entre ellos. Así, un arco entre dos nodos de la red significa que el nodo origen (también llamado nodo padre) tiene un impacto causal sobre el nodo destino (nodo hijo), o lo que es lo mismo, que este último depende del primero. Para poder cuantificar el efecto que los nodos padres tienen sobre un nodo específico de la red, se asocia a cada uno de éstos una tabla de probabilidades condicionales (calculadas mediante el teorema de Bayes), de forma que el valor de cada nodo es función de los valores de sus nodos padres. Además, los nodos hojas representan proposiciones cuyos valores son determinados por observación, a partir de los datos almacenados en un conjunto de entrenamiento del que se extrae la información representada en la red [107].

Dada su utilidad en el campo de la clasificación automática, las redes Bayesianas son herramientas adecuadas tanto para modelar las preferencias de los usuarios, como para utilizarlas posteriormente durante el proceso de recomendación. El perfil de usuario resultante sería un grafo en el que los nodos y arcos identificarían los conceptos/atributos que interesan a los usuarios y las relaciones causales establecidas entre ellos, respectivamente. A la hora de elaborar recomendaciones personalizadas, el clasificador automático decide si un producto objetivo es interesante para el usuario activo,

basándose en las relaciones probabilísticas existentes entre los atributos de dicho producto y las preferencias representadas en su red Bayesiana.

Como principales ventajas, podemos destacar que las redes Bayesianas son modelos de clasificación muy rápidos, sencillos y precisos [35]. Además, las probabilidades condicionales que definen las dependencias causales entre los nodos de la red pueden ser calculadas *off-line*, lo que contribuye a reducir el coste computacional asociado al proceso de recomendación. Para poder aprovechar tales ventajas (computacionales), es necesario que las preferencias de los usuarios no varíen bruscamente a lo largo del tiempo, preservando con ello la validez del modelo calculado inicialmente; es este requisito el que limita la aplicación de las redes Bayesianas en el campo del modelado de usuarios.

Árboles de decisión

Como su propio nombre indica, un árbol de decisión es una estructura en forma de árbol formada por un conjunto de nodos y arcos dirigidos que los conectan entre sí. Como modelo predictivo, su objetivo es clasificar una entrada dada en una categoría específica (típicamente, en los sistemas recomendadores, *interesante* o *no interesante*). Para ello, los nodos internos del árbol identifican preguntas sobre las variables consideradas en el problema de clasificación, mientras que los arcos representan las respuestas a esas preguntas, esto es, todos los posibles valores que pueda tomar cada variable en un conjunto de entrenamiento previamente definido. Por su parte, los nodos hoja se refieren a la clasificación final correspondiente a la variable de entrada considerada.

En la literatura se han propuesto diferentes algoritmos recursivos para construir árboles de decisión a partir de la información almacenada en el conjunto de entrenamiento. De todos ellos, el algoritmo más conocido es ID3 [200], que construye un árbol de decisión de manera descendente, comenzando por identificar el atributo del conjunto de entrenamiento que debe utilizarse como raíz. Para este propósito, ID3 analiza cada atributo utilizando un test estadístico para evaluar la clasificación que establece dicho atributo en relación a los ejemplos de entrenamiento. Tal como describe [171], una manera de cuantificar la relevancia de un atributo en este contexto se basa en medir la cantidad de información que éste proporciona, utilizando la teoría de información propuesta por Claude E. Shannon [223].

Una vez seleccionado el mejor atributo, éste se coloca en la raíz del árbol y se crea un arco para cada uno de los posibles valores del atributo en cuestión. A continuación, los ejemplos de entrenamiento deben repartirse en los nodos en los que terminan los arcos generados, construyendo así el árbol de decisión de forma incremental. Para ello, el algoritmo ID3 considera los valores que toman dichos ejemplos para el atributo raíz y, en función de los mismos, selecciona el atributo que debe ocupar el siguiente nivel del árbol. Este proceso se repite, típicamente, hasta que todos los ejemplos de entrenamiento comparten un mismo valor para el atributo que está siendo evaluado. En ese caso, se crea un nodo hoja en el árbol que decide la clasificación final correspondiente.

Al igual que las redes Bayesianas, los árboles de decisión pueden ser utilizados como técnicas de modelado de usuarios en un sistema de personalización, gracias a su capacidad de clasificación automática. Mediante las capacidades predictivas de este modelo, los recomendadores aprenden a clasificar los productos objetivo (como *interesantes* o *no interesantes*), en función de sus atributos y de las preferencias representadas en el árbol correspondiente a cada usuario activo.

Entre otros, este enfoque ha sido adoptado en los sistemas InfoFinder [126, 127], que recupera documentos para los usuarios basándose en las consultas previas, y en el recomendador Web Syskill & Webert [191, 190], que combina los árboles de decisión con el método basado en vectores de características, descrito en la Sección 2.2.2 de esta misma sección.

Reglas de aprendizaje inductivo

El descubrimiento de reglas de asociación mediante aprendizaje inductivo [158] se ha relacionado tradicionalmente con la realización de estudios de mercado. Esto es debido a que dichas reglas identifican patrones comunes entre los datos contenidos en diferentes transacciones. Mediante estos patrones es posible formular declaraciones del tipo "el 80 % de los usuarios que compran un producto A, también compran B y C", de innegable valor a la hora de conocer la demanda de los usuarios en las transacciones comerciales.

En la literatura se han propuesto algoritmos (como Apriori [2]) que extraen automáticamente las reglas de asociación a partir de un conjunto de entrenamiento, formado por varias transacciones que contienen atributos concretos. Tales reglas, de la forma $X \Rightarrow Y$ (donde X e Y identifican conjuntos de atributos definidos en el conjunto de entrenamiento), tienen dos valores asociados:

- Confianza: Es el porcentaje de las transacciones que contienen el conjunto de atributos X, en las que también está presente Y.
- Soporte: Es el número de transacciones en las que se definen tanto X como Y.

Con el propósito de evitar el descubrimiento de reglas poco significativas, el algoritmo Apriori utiliza dos parámetros de configuración, referidos a los valores mínimos de confianza y soporte exigidos a las reglas de asociación extraídas.

Conscientes de la eficacia del aprendizaje inductivo en el descubrimiento de conocimiento, son varios los sistemas de personalización que se basan en este enfoque a la hora de aprender y representar las preferencias de sus usuarios. En este escenario, los perfiles simplemente almacenan una relación de los atributos que resultan interesantes para los usuarios. Éstos son descubiertos mediante reglas de asociación extraídas por Apriori a partir de bases de datos donde se registran las preferencias de un elevado número de usuarios.

Entre los sistemas que adoptan este enfoque destacan el recomendador WebSift [53], en el dominio de la Web, y PTVPlus [229, 184], Movielens [150] y Recommender [16], en el dominio de la TV personalizada.

Redes neuronales

Las redes neuronales artificiales (RNA) son sistemas de aprendizaje y procesamiento automático de la información, inspirados en el funcionamiento del sistema nervioso de los animales [7]. Están formadas por un gran número de unidades de procesamiento, llamadas neuronas, organizadas en varias capas. Cada neurona recibe una serie de entradas a través de enlaces de comunicación, denominados sinapsis, y emite una salida. Estos enlaces tienen asociado un peso numérico, en el que se codifica el conocimiento que la RNA tiene acerca de un determinado problema. La salida producida por la neurona viene dada por dos funciones:

- una *función de propagación* que, por lo general, es la suma de cada entrada multiplicada por el peso de su enlace o sinapsis (aunque también puede ser no lineal [164]), y
- una *función de activación* que, opcionalmente, modifica la anterior determinando el valor de la salida. Tal como muestra la Fig. 2.1, en la literatura se han propuesto diversos tipos de funciones de activación: funciones lineales, en las que la salida es proporcional a la entrada; funciones umbrales, donde la salida es un valor discreto que depende de si la estimulación total supera o no un determinado umbral; y funciones gaussianas no lineales [117].

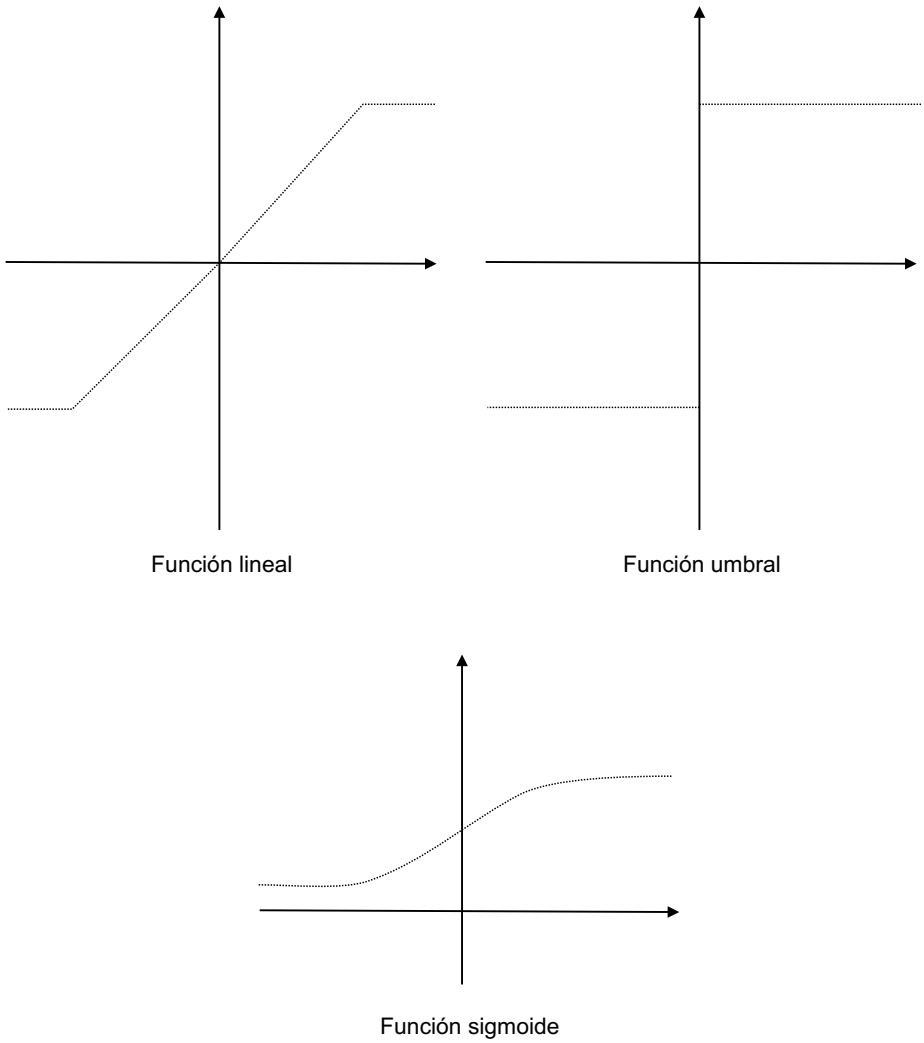

Figura 2.1: Funciones de activación utilizadas en las redes neuronales artificiales

Típicamente, las RNA han sido clasificadas en el estado del arte atendiendo a dos criterios: (i) su topología, relacionada con el número de capas disponibles en la red,

el número de neuronas por capa y el grado o tipo de conectividad entre las mismas, y (ii) el tipo de aprendizaje que la red es capaz de asumir, según ésta necesite o no un conjunto de entrenamiento inicial.

Primer criterio de clasificación: Topología de la RNA

Según su topología, podemos distinguir los siguientes tipos de RNA:

- **Redes monocapa**: Sólo cuentan con una capa de neuronas, que intercambian señales con el exterior y que funcionan a la vez como entrada y salida del sistema. En la literatura, las redes más representativas en esta categoría son el perceptrón simple [212] y la red Hopfield [98].
- **Redes multicapa**: Disponen de conjuntos de neuronas jerarquizadas en distintos niveles o capas, con al menos una capa de entrada y otra de salida y, eventualmente, una o varias capas ocultas. Dentro de éstas, podemos identiticar dos tipos básicos de RNA:
  - **Redes en cascada (*feedforward*)**: En estas redes, la información fluye unidireccionalmente de una capa a otra (desde la capa de entrada a las capas ocultas y desde éstas a la capa de salida), y además, no se admiten conexiones intracapa. Ejemplos representativos de este tipo de red son el perceptrón multicapa (MLP) [88], la red Madaline [215], las máquinas de Boltzmann [95] y Cauchy [239], la memoria lineal adaptativa [99], los modelos de Kohonen [145] y las redes *backpropagation* [214].
  - **Redes recurrentes (*feedback*)**: En este tipo de redes, la información puede volver a lugares por los que ya había pasado, formando bucles. Además, es posible el establecimiento de conexiones intracapa, e incluso la conexión de una neurona consigo misma. Estas últimas propiedades permiten que también puedan ser definidos modelos recurrentes para las redes monocapa descritas previamente (por ejemplo, la red Hopfield). Por su parte, entre las redes recurrentes multicapa propuestas en la literatura, destacan el Cognitrón [75], el Neocognitrón [251] y las redes de resonancia adaptativa (ART) [15].

Mientras las redes en cascada dan siempre soluciones estables, los modelos recurrentes proporcionan soluciones inestables o dinámicas, lo que no siempre es aconsejable. La principal aportación de John Hopfield, creador de la red que hereda su nombre, consistió precisamente en dotar de estabilidad a tales redes recurrentes.

La red neuronal Hopfield

La red Hopfield está formada por un conjunto de unidades de procesamiento —que están en estado activo o inactivo— conectadas tanto consigo mismas como entre sí a través de enlaces (bidireccionales) ponderados de forma simétrica. Su funcionamiento se basa en elegir una unidad de forma aleatoria y comprobar si alguno de sus nodos vecinos está activo. En caso afirmativo, se suman los pesos de todos los enlaces que conectan esta unidad con sus vecinos activos. Si esta suma es positiva, la unidad se activa; en caso contrario, se desactiva. Este proceso se repite, eligiendo otra unidad

al azar, hasta que la red alcanza un estado estable, en el cual no varían los estados de activación de las unidades. Este proceso de búsqueda recibe el nombre de *parallel relaxation* en la literatura [98].

Hopfield desarrolló tanto un modelo continuo como discreto para su red neuronal, definido en función de los valores que pudiesen tomar sus entradas y salidas. En el modelo discreto, estos valores son 0 y 1, y su función de activación es un escalón entre ambos valores; por el contrario, en el modelo continuo, dichas funciones son típicamente sigmoides cuyas salidas oscilan entre -1 y 1, o incluso, entre 0 y 1.

La principal contribución de Hopfield fue demostrar que, para cualquier combinación inicial de pesos y de estados de activación/desactivación de las unidades, el algoritmo propuesto convergía siempre a un estado estable. Con ello, se eliminaban las oscilaciones e inestabilidades asociadas tradicionalmente a los modelos recurrentes.

Segundo criterio de clasificación: Tipo de aprendizaje de la RNA

En lo que al tipo de aprendizaje se refiere, han sido varios los modelos de RNAs propuestos por diferentes autores:

- **Aprendizaje supervisado**: Las redes que adoptan este enfoque necesitan un conjunto de datos de entrada previamente clasificado o cuya respuesta objetivo sea conocida. Esta fase de entrenamiento consiste en aplicar las entradas a la RNA y observar la salida que produce. Si esta salida no coincide con la esperada, es necesario ajustar los pesos de cada neurona para, iterativamente, ir obteniendo las respuestas adecuadas del sistema. Los ejemplos presentados deben ser suficientemente representativos en el dominio de aplicación concreto, de forma que sea posible que la RNA *aprenda* a partir de ellos. Entre las redes que requieren este entrenamiento podemos destacar el perceptrón simple [212], la red Adaline [253], el perceptrón multicapa [88], la memoria asociativa bidireccional [119], las máquinas de Boltzmann [95] y Cauchy [239] y las redes *backpropagation* [214].

- **Aprendizaje no supervisado o autoorganizado**: En este caso es posible prescindir de la fase de entrenamiento inicial. Es el aprendizaje en el que se basan las memorias asociativas [119], las redes de Hopfield [98], los modelos de Kohonen [145], Cognitrón [75] y Neocognitrón [251].

- **Redes híbridas**: Se basan en un enfoque mixto en el que se utiliza una función de mejora para facilitar la convergencia. Un ejemplo de este último tipo de RNA son las redes de base radial (RBF) [164].

Gracias a sus capacidades de aprendizaje y clasificación automática, las RNA se han aplicado también en el campo de los sistemas recomendadores. Prueba de ello son los enfoques presentados en el sistema Re:Agent [30], encargado de filtrar los correos electrónicos de los usuarios en función de los mensajes leídos por éste recientemente, y en el recomendador de noticias personalizadas descrito en [106], donde sus autores entrenan una red neuronal para cada usuario, con el fin de aprender y representar sus preferencias. De forma análoga a la comentada en el resto de modelos basados en clasificadores, en la red usada en [106] los nodos identifican las palabras que aparecen en los artículos consultados por el usuario en el pasado, y las conexiones neuronales ponderadas representan la fuerza de las asociaciones existentes entre las palabras que

aparecen en estos documentos. De esta forma, a la hora de decidir si una noticia debe ser sugerida al usuario, el sistema extrae los términos más significativos de la misma y los coloca en las entradas de la red neuronal. Basándose en los ejemplos aprendidos durante una fase de entrenamiento supervisado, la red trata de descubrir patrones comunes entre dicha noticia y las preferencias del usuario. Obviamente, en caso de que éstos existan, esta noticia es finalmente sugerida al usuario.

Ya en el dominio de la TV personalizada, podemos destacar la herramienta presentada en [50] y el sistema TV Recommender Show [260, 40] en el que se emplea una red neuronal RBF [164] para combinar las sugerencias elaboradas por dos estrategias diferentes, construyendo así la lista final de programas mostrados al espectador. Ambas estrategias están basadas en clasificadores, concretamente en un árbol de decisión y una red Bayesiana.

2.3 Generación del perfil inicial

Una vez determinado el modelo de representación de las preferencias de los usuarios activos, es necesario recurrir a mecanismos que permitan inicializar sus perfiles personales, para, de esta forma, empezar a prestar los servicios de personalización.

En la literatura, se han propuesto mecanismos de muy diversa naturaleza para poder abordar esta tarea. Entre ellos, podemos destacar: (i) el reconocimiento automático de las preferencias del usuario, (ii) la inicialización manual de los perfiles, y (iii) la aplicación de procedimientos de clasificación semiautomáticos, basados en el uso de estereotipos y conjuntos de entrenamiento.

2.3.1 Reconocimiento automático de preferencias

Algunos recomendadores no necesitan crear un perfil inicial para cada usuario activo, sino que sus preferencias son reconocidas de forma automática a medida que éste interactúa con el sistema. Por su propia naturaleza, esta técnica se ha implantado con gran éxito en el dominio de las herramientas de personalización Web, en las que las recomendaciones son elaboradas a partir del historial de navegación de los usuarios. Algunos ejemplos de este tipo de recomendadores son WebSell [57], Fab [14] y Webmate [48].

2.3.2 Inicialización manual

Otros sistemas exigen a los usuarios una descripción explícita de sus intereses personales, para, de esta forma, poder construir su perfil inicial. Si bien esta técnica asegura una precisión máxima en lo que al modelado de preferencias se refiere, su principal limitación es el esfuerzo que deben asumir los usuarios a la hora de rellenar largos y tediosos formularios de registro.

2.3.3 Estereotipos

Esta técnica, propuesta por Elaine Rich en el año 1989 [207], consiste en definir diferentes categorías genéricas a las que pueden pertenecer los usuarios en función de

sus intereses y datos personales. De esta forma, cuando llega un nuevo usuario al sistema, éste debe proporcionar una serie de datos demográficos (por ejemplo, sexo, edad, ocupación, aficiones, intereses generales, etc.) que permitirán asignarle un estereotipo inicial utilizado como perfil. Por su naturaleza, este mecanismo de creación de perfiles iniciales suele ir acompañado del método de representación basado en características demográficas, descrito en la Sección 2.2.4.

Al igual que el método de inicialización manual, la principal debilidad de esta técnica está relacionada con la dificultad que supone conseguir este tipo de información por parte de los usuarios, no siempre dispuestos, por otro lado, a desvelar datos de carácter personal. A pesar de ello, el modelado de usuarios basado en estereotipos se ha aplicado con éxito en el sistema LifeStyle Finder [125], que sugiere diferentes productos comerciales a los usuarios en función de sus características personales y de los intereses de otros que comparten su estereotipo; y en la guía de programación de TV personalizada PPG [8], en la que a los espectadores se les asigna un estereotipo inicial en función de los intereses televisivos que hayan declarado (género, temática, canal, franja horaria preferente para ver televisión, etc.).

2.3.4 Conjuntos de entrenamiento

Algunos sistemas presentan un conjunto de productos a los usuarios para que identifiquen aquellos que les interesan y aquellos que no son relevantes para ellos. A partir de la información proporcionada, el sistema construye un conjunto de entrenamiento, procesado, generalmente, mediante alguno de los clasificadores descritos en la Sección 2.2.7.

La principal limitación de este mecanismo es la identificación de los ejemplos presentados a los usuarios, dado que éstos deben ser suficientemente representativos para así garantizar recomendaciones precisas.

Entre los sistemas recomendadores que utilizan conjuntos de entrenamiento para definir perfiles iniciales, podemos destacar News Weeder [130], ACR News [158], PSUN [232] y NewT [227], que selecciona noticias personalizadas para sus usuarios, la herramienta de filtrado de correos electrónicos Re:Agent [30], y los recomendadores Ringo [224] y Movielens [150], vinculados al dominio de la música y la TV personalizada, respectivamente.

2.4 Realimentación de relevancia

En el campo de la personalización, el término *realimentación de relevancia* identifica la información que los usuarios proporcionan al sistema recomendador, una vez evaluadas las sugerencias elaboradas por éste. Gracias a esta información, el sistema puede modificar las preferencias de los usuarios, con el propósito de agilizar la convergencia entre sus intereses reales y las sugerencias ofrecidas en el futuro.

Para poder obtener la realimentación de relevancia, los sistemas de personalización recurren a diversos métodos. El más sencillo y directo de todos ellos es aquel en el que el propio usuario declara explícitamente su (des)interés en relación a cada producto sugerido. Una solución mejorada con respecto a ésta consiste en inferir dicha información a partir de la interacción entre el usuario y el sistema. Por último, algunos sistemas optan por un enfoque híbrido en el que se combinan los dos modelos anteriores.

2.4.1 Sistemas recomendadores sin realimentación

En la literatura es posible identificar sistemas recomendadores que no actualizan automáticamente los perfiles a lo largo del tiempo, sino que delegan esta responsabilidad en los propios usuarios [256, 30]. Es obvio que este tipo de enfoques no necesitan la realimentación de relevancia; sin embargo, obligan al usuario a asumir un rol activo no siempre deseable.

2.4.2 Realimentación explícita

En los sistemas que adoptan este enfoque son los usuarios quienes declaran la realimentación de relevancia de forma explícita. Para acometer esta tarea, se han definido diferentes métodos:

- En algunos sistemas, los usuarios simplemente clasifican los productos como *interesantes* o *no interesantes*. Como ejemplos más representativos destacamos el recomendador Web Webmate [48], y las herramientas de filtrado de correo electrónico SIFT Netnews [256] y NewT [227].

- Otros recomendadores requieren que sus usuarios definan un nivel de interés concreto. Este enfoque ha sido adoptado en multitud de sistemas de personalización: Amazon [4], los recomendadores Web Amalthea [168], ifWeb [153] y Syskill & Webert [191], los sistemas de filtrado de noticias personalizadas INFOrmer [208] y PSUN [232], el recomendador de música Ringo [224] y los sistemas de TV personalizada Movielens [150] y Recommender [16].

- Por último, existen sistemas en los que los usuarios proporcionan realimentación mediante comentarios textuales. Este es el caso del recomendador de noticias personalizadas Grouplens [120] y el sistema de filtrado de correo electrónico Tapestry [79].

2.4.3 Realimentación implícita

La realimentación de relevancia puede ser inferida por el sistema recomendador a partir de las acciones que lleva a cabo el usuario [156].

Dependiendo del dominio de aplicación de cada sistema, los enfoques propuestos emplean diferentes indicativos para obtener dicha realimentación y, en consecuencia, modificar los perfiles de los usuarios activos: los enlaces por los que éstos han navegado [136, 157, 53, 158], su historial de consumo (por ejemplo, productos comprados mediante comercio electrónico) [4, 44, 125], el tiempo empleado en ver un programa de TV o en consultar una determinada página Web [167, 120, 118, 112], etc. Tal como decíamos antes, ciertas acciones de los usuarios también pueden ser útiles a la hora de aprender conocimiento adicional sobre sus preferencias: salvar e imprimir un documento [111], marcar una página Web, minimizar la ventana en la que ésta se visualiza [112], ordenar la grabación de un programa de TV, avanzar, retroceder o detener su reproducción [260, 8], etc.

2.4.4 Realimentación híbrida

La mayoría de los enfoques propuestos optan por un modelo híbrido, en el que la realimentación inferida por el propio sistema es completada con las clasificaciones explícitas de los usuarios activos. De esta forma, es posible aunar las ventajas de ambos métodos, aprovechando tanto la precisión de la técnica explícita, como el mínimo esfuerzo requerido a los usuarios en el modelo implícito. Entre los sistemas híbridos analizados destacamos LifeStyle Finder [125], Anatagonomy [112], Tapestry [79], Grouplens [120], y los recomendadores de comercio electrónico CDNow [44] y Amazon [4].

2.5 Técnicas para la adaptación del perfil de usuario

Las preferencias de los usuarios activos en un sistema recomendador, típicamente, varían a lo largo del tiempo, de ahí que la realimentación de relevancia sea un elemento clave a la hora de reflejar sus nuevos intereses o necesidades. Habida cuenta de esta evolución, en la literatura se han propuesto diferentes métodos para poder adaptar los perfiles de los usuarios a sus nuevas preferencias, eliminando aquellas que han quedado obsoletas. Entre estos enfoques podemos destacar: (i) la actualización manual por parte de los usuarios, (ii) la incorporación de nueva información al perfil, (iii) el uso de una función de olvido gradual, y (iv) la selección natural en ecosistemas de agentes.

2.5.1 Actualización manual por parte de los usuarios

Tal como adelantamos en las secciones previas, algunos sistemas exigen que los usuarios activos modifiquen sus propios perfiles, incorporando así los posibles cambios que hayan sufrido sus preferencias desde la última interacción con el recomendador [256, 30]. Este método plantea serias limitaciones: no sólo es una técnica costosa para los usuarios, obligados a participar activamente en el sistema, sino que además, la actualización manual es completamente inviable cuando las preferencias varían con mucha frecuencia.

2.5.2 Incorporación de nueva información al perfil

El enfoque más extendido entre los sistemas recomendadores propuestos en la literatura consiste en añadir nueva información a los perfiles de los usuarios activos, obtenida a partir de la realimentación de relevancia. Su principal ventaja es que permite que los perfiles se adapten, rápida y fácilmente, a la evolución en las preferencias de los usuarios; sin embargo, no contempla ningún mecanismo para que el sistema *olvide* los intereses que éstos tenían en el pasado.

Algunos de los sistemas que aplican este método son Amazon [4], Beehive [103], Anatagonomy [112], CDNow [44], infoFinder [126], Let's Browse [137], Syskill & Weber [191], INFOrmer [208], LifeStylefinder [125], WebWatcher [109] y Letizia [136], entre otros muchos. En el dominio de la TV personalizada, destacan Bellcore Video Recommender [94], Recommender [16], PTV [55], PTVPlus [229], TV Recommender Show [260], PPG [8], Movielens [150] y Moviefinder [169].

2.5.3 Función de olvido gradual

En 1996, Webb y Kuzmycz introdujeron el concepto de *olvido gradual*, basándose en la premisa de que, generalmente, las personas tienden a olvidar poco a poco el conocimiento que adquirieron en el pasado. Este comportamiento fue modelado por ambos investigadores mediante la llamada *función de olvido gradual* [247].

Dicha función asigna un peso a cada producto definido en el perfil del usuario activo, cuyo valor dependerá de su localización temporal. Así, cuando el sistema añade un nuevo producto, le asocia un peso de valor 1 y decrementa los índices de los productos que ya estaban almacenados en el perfil, de forma que las preferencias más recientes son las que tienen asociado un peso elevado. Éste es el enfoque utilizado en los recomendadores Web SiteIF [234] e ifWeb [153], y en el sistema de filtrado de documentos LaboUr [220].

Un caso particular de la función de olvido gradual ha sido adoptado en enfoques basados en ventanas temporales. Estos métodos aprenden las preferencias de los usuarios a partir de conjuntos de entrenamiento en los que sólo se consideran los datos más recientes. En realidad, dichas preferencias se aprenden desde los ejemplos de entrenamiento que están incluidos en una determinada ventana temporal [155], de forma que el sistema olvida todas aquellas preferencias que quedan fuera de la misma [143].

Una posible mejora de esta técnica se consigue introduciendo heurísticos para ajustar dinámicamente el tamaño de esta ventana, en función de la precisión de las estrategias de aprendizaje utilizadas (por ejemplo, clasificadores) [252]. Así, cuando se reduce la precisión, el sistema estrecha la ventana temporal, eliminando con ello el efecto negativo de aquellas preferencias del usuario activo que hayan quedado obsoletas.

2.5.4 Selección natural

El enfoque de la selección natural se ha asociado tradicionalmente a aquellos sistemas en los que un conjunto de agentes cooperan e interactúan en un determinado ecosistema. En un sistema de personalización, la evolución del ecosistema se basa en el siguiente principio: aquellos agentes que ofrecen las recomendaciones más precisas se reproducen, mientras que el resto son destruidos.

Este tipo de técnicas permiten que el recomendador se adapte a las nuevas preferencias de los usuarios activos, sin más que eliminar aquellos agentes que utilizan sus intereses pasados. Este enfoque ha sido adoptado en las herramientas de filtrado de noticias personalizadas PSUN [232] y NewT [227], y en los recomendadores Web Fab [14] y Amalthea [168].

2.6 Métodos de filtrado de información

Además de técnicas de modelado de usuarios, los sistemas recomendadores requieren mecanismos para seleccionar los productos que mejor se ajustan a sus preferencias o necesidades. En la literatura, son cuatro los métodos de filtrado de información más extendidos: (i) el filtrado demográfico, (ii) los métodos basados en contenido, (iii) el filtrado colaborativo, y (iv) los enfoques híbridos, que combinan estrategias de diferente naturaleza.

2.6.1 Filtrado demográfico

El filtrado demográfico emplea las características personales de los usuarios (edad, sexo, estado civil, ocupación profesional, historiales de compra, aficiones, etc.) —proporcionadas durante la fase de registro en el sistema— para descubrir las relaciones existentes entre un determinado producto y el tipo de usuarios a quien éste interesa [125]. Por su naturaleza, este método modela las preferencias de los usuarios como un vector de características demográficas, y recurre a la técnica de estereotipos para inicializar sus perfiles (ver Secciones 2.2.4 y 2.3.3). Así, todos aquellos usuarios que pertenezcan al mismo estereotipo recibirán las mismas sugerencias por parte del sistema.

El filtrado demográfico adolece de dos limitaciones principales: por una parte, puede conducir a recomendaciones demasiado generales e imprecisas para los usuarios, por considerar únicamente sus características demográficas; además, este método no permite que las sugerencias ofrecidas se adapten a posibles cambios en las preferencias de los usuarios, dado que sus datos personales suelen permanecer invariables a lo largo del tiempo.

A pesar de los inconvenientes comentados, el filtrado demográfico puede ser una estrategia útil si se combina con otros métodos.

2.6.2 Métodos basados en contenido

Esta técnica consiste en recomendar a un usuario activo productos que son *similares* a los que le han gustado en el pasado. Para poder comparar las preferencias del usuario activo frente a los productos objetivo, los métodos basados en contenido consideran sus principales características o atributos.

Si bien, por su propia naturaleza, es una estrategia muy precisa, son varias las limitaciones reconocidas en el filtrado basado en contenido:

- Por una parte, esta técnica suele recomendar productos excesivamente parecidos a los que el usuario activo ya conoce, e incluso, demasiado similares entre sí [231]. En este último caso, si alguno de los productos sugeridos no interesa al usuario su confianza en el sistema se verá comprometida (debido a que el resto de las sugerencias ofrecidas también se parecen a dicho producto).

  Esta limitación, acuñada en la literatura bajo el término de *sobreespecialización*, tiene su origen en las métricas de similitud adoptadas en los enfoques convencionales basados en contenido. Tales métricas únicamente detectan similitud entre dos productos cuando se advierte cierto solapamiento entre sus características o atributos, de ahí la falta de diversidad de las recomendaciones sugeridas.

  El problema de la *sobreespecialización* en los sistemas recomendadores basados en contenido ha sido combatido en algunos enfoques propuestos en la literatura [230, 202, 33]. Bajo una filosofía común, estas propuestas tratan de diversificar las recomendaciones ofrecidas a los usuarios sin necesidad de comprometer la personalización. Su objetivo es, por tanto, alcanzar un equilibrio justo entre dos parámetros: (i) la similitud existente entre las preferencias del usuario y las sugerencias ofrecidas, y (ii) la diversidad relativa entre cada producto sugerido y los restantes incluidos en la recomendación final. En consecuencia, tales enfoques seleccionan un conjunto de productos que no sólo son parecidos a los que inte-

resan al usuario activo, sino que además son diferentes entre sí, evitando así la posible pérdida de confianza del usuario en el sistema, mencionada al comienzo de la sección.

De nuestra explicación se desprende que este tipo de enfoques combaten la *sobreespecialización* desde un único frente: aseguran que los productos sugeridos son diferentes entre sí, pero no contemplan la posibilidad de que éstos aún pueden ser demasiado similares a las preferencias del usuario activo. En este escenario, el usuario podría llegar a cuestionar la utilidad de un sistema de personalización que únicamente sugiere productos muy parecidos a aquellos que él ya conoce.

- La segunda de las limitaciones de los métodos basados en contenido tiene su origen en la especificación de los atributos de los productos considerados, una tarea costosa que, en algunos casos, incluso requiere la participación de un experto de conocimiento, capaz de describir el dominio del sistema de forma precisa.
- Por último, destacar el llamado *new user ramp-up* [42], un problema relacionado con la llegada de nuevos usuarios al sistema. En este escenario, el recomendador dispone, típicamente, de muy poco conocimiento sobre sus preferencias personales, de ahí la baja precisión de las recomendaciones ofrecidas.

Una vez identificadas las limitaciones de las que adolecen los sistemas basados en contenido actuales, nos centramos en las técnicas que utilizan para comparar las preferencias de los usuarios activos frente a los productos objetivo.

Comparación sintáctica basada en palabras claves

Este método consiste en comparar sintácticamente las características definidas en el perfil del usuario activo, frente a los atributos de los productos objetivo considerados. Tiene las limitaciones propias de los enfoques que no razonan sobre el significado de los términos a través de su semántica (palabras con múltiples significados, sinónimos, etc.).

A pesar de ello, esta técnica se ha aplicado exitosamente en diferentes herramientas de personalización Web, como los buscadores de Internet convencionales, y los sistemas ifWeb [10] y SiteIf [234].

Categorización de texto (TC - *Text Categorization*)

El área de la categorización o clasificación textual surge con el propósito de salvar las limitaciones de los enfoques sintácticos convencionales. Sus técnicas se basan en aprender múltiples características a partir de descripciones textuales, y aplicarlas en la construcción de clasificadores eficientes. Entre las aplicaciones de estos últimos, podemos destacar el filtrado de información, el agrupamiento de documentos relacionados entre sí, la clasificación de los mismos en categorías predefinidas, etc.

Tales capacidades han sido aprovechadas en algunos de los sistemas recomendadores basados en contenido propuestos en la literatura. Estos sistemas procesan las descripciones textuales de las preferencias de los usuarios, identificando las categorías genéricas a las que pertenecen. Este modelo únicamente recomienda un producto objetivo al usuario, cuando el proceso de clasificación determina que está incluido en sus categorías de interés.

Como ejemplos de aplicación, destacamos el sistema recomendador de películas INTIMATE [142] y las herramientas presentadas en [165] y [23], que sugieren libros y noticias personalizadas, respectivamente.

Similitud basada en coseno

Esta técnica —tradicionalmente vinculada al campo de la recuperación de información— se utiliza en sistemas que representan las preferencias de los usuarios (y también los productos objetivo) mediante los vectores de características semánticas descritos en la Sección 2.2.2. El enfoque, propuesto por Salton en 1983, calcula la similitud entre dos vectores como el coseno del ángulo que forman, tal como se muestra en la Ec. 2.1. Así, si ambos vectores son idénticos, la similitud toma el valor 1, mientras que si son completamente diferentes y, por tanto ortogonales, la similitud entre ambos se anula [216].

$$sim_{Cos}(A, B) = \frac{\vec{A} \cdot \vec{B}}{\left|\vec{A}\right| \cdot \left|\vec{B}\right|} \tag{2.1}$$

La literatura revela la adopción de este enfoque en diferentes dominios de aplicación: filtrado de noticias (Anatagonomy [112] y NewT [227]), TV personalizada (Movielens [150] y Moviefinder [169]), y recomendadores Web (Amalthea [168], Let's Browse [137], Letizia [136], Webmate [48], Fab [14] y WebWatcher [9]), entre otros.

Vecinos más cercanos

Los algoritmos que utilizan este enfoque calculan la distancia entre los productos objetivo y las preferencias del usuario activo, de forma que sólo sugieren aquellos productos que son más parecidos a los que le interesaron en el pasado (i.e. que *están más cerca* de estos últimos).

Un caso particular de las técnicas basadas en la búsqueda de vecinos más cercanos es el razonamiento basado en casos (CBR - *Case Based Reasoning*). En este escenario, los perfiles de los usuarios se representan mediante una colección de experiencias o casos pasados. Dependiendo del dominio de aplicación del sistema, dichos casos pueden ser programas de TV vistos por el usuario, noticias y documentos consultados en la Web, sitios Web visitados o los últimos libros que compró. A la hora de recomendar un producto objetivo concreto, CBR aplica sobre este conjunto de entrenamiento diferentes métricas de similitud, para valorar si dicho producto se ajusta a las preferencias personales de cada usuario activo. Dependiendo del mecanismo elegido para representar los perfiles de los usuarios (ver Sección 2.2), dichas métricas pueden incluir:

- comparaciones sintácticas entre palabras clave,
- cálculo de similitud basada en coseno, o
- procesos en los que se comparan las preferencias del usuario con los productos objetivo, y se ponderan los valores de similitud resultantes mediante el nivel de interés del usuario en relación a cada uno de los casos definidos en su perfil.

Entre los sistemas que utilizan este enfoque podemos destacar LaboUR [220], WebSell [57], Re:Agent [30], Syskill & Webert [190] y Entree [42]. Este último sistema

merece una mención especial, ya que la estrategia de recomendación empleada lo diferencia, en algunos aspectos, del resto de enfoques descritos en este capítulo. Dicho sistema sugiere restaurantes a sus usuarios utilizando CBR como método para valorar la similitud entre las preferencias de los mismos y los casos (restaurantes) disponibles en la herramienta. La diferencia mencionada antes es debida a que el proceso de personalización no trata de minimizar la interacción entre el sistema y el usuario, sino que la herramienta dialoga activamente con el mismo para así precisar las características de los restaurantes recomendados. En cada interacción, el sistema presenta una selección de restaurantes al usuario, y éste tiene la posibilidad de establecer ciertos parámetros de búsqueda (como tipo de cocina, atmósfera del restaurante, precio, localización, etc.) para que el sistema Entree refine sus recomendaciones [42]. Si bien esta interacción entre sistema y usuario redunda en la precisión y calidad de las recomendaciones ofrecidas, no es menos cierto que, en algunos dominios de aplicación, puede convertirse en una tarea tediosa y molesta para el usuario, lo que limita considerablemente la utilidad de este tipo de enfoques.

Clasificación automática

En los sistemas basados en contenido, el proceso de elaboración de recomendaciones personalizadas se reduce a una simple tarea de clasificación. Para adoptar este enfoque, los sistemas deben representar los perfiles de los usuarios mediante los modelos basados en clasificadores automáticos, ya sean redes Bayesianas, árboles de decisión, reglas de asociación o redes neuronales.

Con arreglo a lo explicado en la Sección 2.2.7, en este tipo de propuestas, el sistema dispone de un conjunto de entrenamiento en el que se incluyen las características de los productos que interesaron al usuario en el pasado. Basándose en la frecuencia de aparición de dichas características en el conjunto, el recomendador predice un modelo de sus preferencias (su perfil), que permite clasificar los productos objetivo en categorías genéricas. Este proceso conduce a la recomendación de productos que comparten *exactamente* las características que interesan al usuario activo, de ahí el carácter *sobreespecializado* de este tipo de sugerencias (basadas en contenido).

Como ejemplos más representativos, destacamos el sistema de filtrado de correo electrónico Re:Agent [30], que emplea una red neuronal para clasificar los mensajes recibidos en dos carpetas etiquetadas como *trabajo* y *otras*; el recomendador Syskill & Webert [190], que emplea un árbol de decisión para decidir si las páginas Web son *interesantes* o *no interesantes* para los usuarios, y los sistemas de TV personalizada Recommender [16] y Movielens [150], que clasifican películas mediante reglas de aprendizaje inductivo.

2.6.3 Filtrado colaborativo

Es una de las técnicas de filtrado de información más empleadas en los sistemas recomendadores propuestos en la literatura. A diferencia de los métodos basados en contenido, a la hora de ofrecer una recomendación a un usuario activo, el filtrado colaborativo no considera únicamente sus preferencias personales, sino también las de otros usuarios con intereses similares a los suyos (llamados en adelante vecinos). Tal similitud se estima a partir de las clasificaciones (o índices de interés) definidas en sus perfiles personales; de ahí que los enfoques colaborativos no requieran las descripcio-

nes semánticas consideradas en los métodos basados en contenido.

En la literatura se han definido dos técnicas diferentes dentro de las estrategias colaborativas:

- *Filtrado colaborativo basado en usuario*: Sugiere a cada usuario activo aquellos productos que han interesado a sus vecinos. Para formar este vecindario, la estrategia considera que dos usuarios tienen preferencias similares si han clasificado los mismos productos en sus perfiles y les han asignado índices de interés parecidos.

- *Filtrado colaborativo basado en item*: Mediante este método, un producto es recomendado a un usuario activo si es similar a los definidos en su perfil personal. En este caso, se considera que dos productos son similares (o vecinos) si los usuarios que han clasificado uno de ellos tienden a clasificar el otro, asignándole índices de interés parecidos. Esta técnica da mejores resultados que la variante basada en usuario cuando el número de productos disponibles en el sistema recomendador es mucho menor que el número de usuarios [218].

En ambas técnicas, el objetivo es predecir el (nivel de) interés del usuario activo en relación a un producto objetivo dado y, en función de éste, sugerirlo (o no) a dicho usuario. En este proceso, se identifican tres fases:

1. En los enfoques colaborativos basados en usuario, la primera fase selecciona los usuarios cuyas preferencias son similares a las del usuario activo. Por el contrario, los enfoques basados en *item* extraen las preferencias del usuario que son más similares al producto objetivo.

2. A continuación, es necesario formar el vecindario —ya sea del usuario activo o de los productos definidos en su perfil—, a partir de la selección realizada en la etapa anterior.

3. Finalmente, tal como adelantábamos previamente, el sistema debe predecir el nivel de interés del usuario activo en relación al producto objetivo. Para ello, los enfoques basados en usuario consideran el nivel de interés de los vecinos del usuario activo en relación al producto objetivo. Por el contrario, la versión colaborativa basada en *item* considera el nivel de interés del usuario activo en relación a aquellos productos de su perfil que son más similares al objetivo.

Fase 1: Selección de preferencias similares

Los métodos más utilizados en la literatura a la hora de medir la similitud entre los perfiles de varios usuarios, son las técnicas de selección de los vecinos más cercanos, el *clustering* y los modelos basados en clasificadores. Además, la primera es también la estrategia más extendida entre los enfoques colaborativos basados en *item*.

- **Vecinos más cercanos**: De forma análoga a la explicada en los métodos basados en contenido, esta técnica también se utiliza para detectar similitud entre las preferencias de los usuarios en los enfoques colaborativos (tanto basados en usuario como basados en *item*).

En los enfoques de filtrado colaborativo basado en usuario, cada perfil se representa mediante un vector, cuyas componentes son las clasificaciones que el usuario ha asignado a cada uno de los productos incluidos en el mismo. Una vez modelados todos los usuarios, se aplican sobre sus respectivos vectores métricas como la similitud basada en coseno (ver Ec. 2.1) o la correlación Pearson-r (ver Ec. 2.2), detectando así parecidos entre sus preferencias.

$$corr_{Pea}(\vec{A}, \vec{B}) = \frac{\sum_r (\vec{A}[r] - \bar{A}) \cdot (\vec{B}[r] - \bar{B})}{\sqrt{\sum_r (\vec{A}[r] - \bar{A})^2 \cdot \sum_r (\vec{B}[r] - \bar{B})^2}} \tag{2.2}$$

donde $\bar{A}$ y $\bar{B}$ son los valores medios de las clasificaciones definidas por los dos usuarios cuyas preferencias son comparadas (y cuyos niveles de interés se registran en los vectores $\vec{A}$ y $\vec{B}$, respectivamente).

De acuerdo a la Ec. 2.2, es evidente que cuantos más productos hayan clasificado a la vez los usuarios comparados, y cuanto más parecidas sean las clasificaciones asignadas a los mismos, mayor será la correlación detectada entre sus preferencias.

Por su parte, tal como adelantábamos en la Sección 2.6.3, en lugar de computar la similitud entre dos usuarios concretos, los enfoques colaborativos basados en *item* calculan la similitud entre dos productos (*items*). Para ello, este tipo de propuestas seleccionan los usuarios que han clasificado a la vez los productos comparados, y construyen sendos vectores a partir de los niveles de interés que éstos han asignado a ambos productos. Finalmente, basta aplicar sobre estos vectores las métricas antes mencionadas (coseno, Pearson-r), para así obtener los valores de similitud concretos entre los productos considerados [218]. Razonando de forma análoga a la descrita en los sistemas colaborativos basados en usuario, es evidente que cuantos más usuarios hayan clasificado a la vez los dos productos comparados y cuanto más parecidos sean los niveles de interés asignados a los mismos, más significativo será el valor de similitud medido entre ambos.

Entre los sistemas colaborativos basados en la selección de los vecinos más cercanos destacamos los recomendadores de películas Bellcore Video Recommender [94] y Movielens [150], los sistemas de filtrado de noticias personalizadas GroupLens [120] y NewsWeeder [130], los recomendadores de listas musicales Ringo [224] y SmartRadio [91] y los sistemas Web Fab [14], WebWatcher [109] y WebSell [57].

- **Clustering**: Mediante el método de estereotipos introducido por Rich, descrito en la Sección 2.3.3, es posible identificar diferentes grupos o categorías a las que pueden pertenecer los usuarios en función de sus características y preferencias. Como ya señalamos anteriormente, la identificación de estos estereotipos es una tarea laboriosa, ya que supone gran aporte de información por parte de los usuarios.

  Como alternativa a este método, Orwant propuso en 1995 un método de *clustering*, basado en definir grupos de usuarios de forma dinámica a partir de los perfiles individuales disponibles en el sistema [183]. Las características que comparten varios usuarios son utilizadas precisamente como estereotipos, de forma que una vez identificados los grupos de usuarios comunes, el sistema predice el interés de cada uno de ellos promediando las clasificaciones del resto de usuarios pertenecientes a su mismo grupo.

De acuerdo a los resultados obtenidos por Breese [35], las recomendaciones elaboradas mediante los métodos basados en vecinos más cercanos son más precisas que las obtenidas mediante el *clustering*, hecho que limita la aplicación de esta última técnica en los enfoques colaborativos actuales.

- **Clasificadores**: El filtrado colaborativo está estrechamente relacionado con la clasificación automática (ver Sección 2.2.7). Típicamente, este tipo de enfoques representan los datos de entrada sobre los que trabaja el clasificador mediante las matrices descritas en la Sección 2.2.3, en las que se definen los niveles de interés que cada usuario (filas) asigna a los productos (columnas) registrados en su perfil. Dado que es poco probable que todos los usuarios clasifiquen exactamente los mismos productos, dicha matriz tendrá muchos elementos vacíos. Tal como se describió previamente, la tarea del filtrado colaborativo es precisamente eliminar estos huecos, prediciendo el interés de cada usuario en relación a todos aquellos productos a los que les corresponde una entrada vacía en la matriz de clasificaciones.

  En la literatura, existen enfoques que asignan un clasificador a cada usuario del sistema para poder predecir sus niveles de interés (en función de los gustos de los restantes), detectando con ello similitud entre sus respectivas preferencias. Algunos ejemplos del uso de clasificadores en este contexto pueden consultarse en [35, 22, 16, 80, 159, 139, 63]. El trabajo presentado en [35] adopta un modelo basado en redes Bayesianas, el enfoque descrito en [22] se basa en los árboles de decisión, mientras que los cinco enfoques restantes utilizan reglas de asociación para cuantificar la similitud entre los productos definidos en la matriz de clasificaciones del sistema.

Fase 2: Creación del vecindario de un usuario o de un producto

Son tres las técnicas utilizadas tradicionalmente a la hora de determinar el tamaño del vecindario en los enfoques colaborativos:

- **Combinación de correlación y umbral**: En los enfoques basados en usuario, este método establece que sólo aquellos usuarios cuya correlación con el activo supere un cierto umbral, podrán formar parte de su vecindario. De forma análoga, los enfoques basados en *item* consideran que únicamente aquellos productos del perfil del usuario activo cuya similitud con el objetivo es mayor que el umbral establecido, pueden estar incluidos en su vecindario.

- **Los n mejores vecinos**: Según la técnica colaborativa empleada (basada en usuario o basada en *item*), este método selecciona los n usuarios más parecidos al activo, o los n productos de su perfil más similares al objetivo.

  Es obvio que si n es muy elevado, se reduce la calidad de las recomendaciones ofrecidas, mientras que si se asumen valores reducidos, se limita mucho la capacidad predictiva del sistema. La principal ventaja de este método es que permite obtener recomendaciones más precisas que el anterior, aun cuando el solapamiento entre las preferencias de los usuarios es muy reducido (recordar la matriz de clasificaciones explicada en Sección 2.2.3) [80, 259].

- **Centroide**: En primer lugar, este enfoque selecciona el usuario más cercano al activo y calcula su centroide [219]. A continuación, incluye otros usuarios en el

vecindario del activo, utilizando como criterio de selección la mínima distancia entre éstos y el centroide, que es recalculado cada vez que se incorporan nuevos vecinos. Su comportamiento en escenarios con matrices de clasificaciones poco pobladas es peor que el de la estrategia de selección de los n vecinos más parecidos al usuario activo, razón por la que los enfoques basados en el cálculo de centroides no gozan de gran popularidad en la literatura.

De forma análoga a la comentada en los enfoques previos, este método también puede aplicarse en las técnicas colaborativas basadas en *item* para comparar los productos en lugar de los usuarios que los consumen.

Fase 3: Predicción basada en el vecindario creado

Una vez formado el vecindario del usuario activo, el sistema colaborativo debe predecir su (nivel de) interés en relación al producto objetivo considerado. También en este contexto, se han propuesto técnicas de diversa naturaleza:

- **Recomendación de los productos más frecuentes**: En los enfoques basados en usuario, este método selecciona, de entre todos los productos que interesan a los vecinos del usuario activo, aquellos que aparecen con mayor frecuencia en sus perfiles. Es decir, el recomendador sugiere al usuario activo un producto objetivo si éste ha sido muy relevante para la mayoría de sus vecinos.

  En los enfoques basados en *item* también se premian los productos más frecuentes en los perfiles de los usuarios, ya que éstos ayudan a incrementar los valores de similitud (entre las preferencias del usuario activo y el producto objetivo) utilizados en el proceso de predicción.

- **Recomendación basada en reglas de asociación**: Este método se utiliza en las técnicas colaborativas basadas en *item* y consiste en extraer reglas de asociación a partir de un conjunto de entrenamiento que contiene los perfiles de todos los usuarios del sistema. Estos enfoques [55, 184] cuantifican la similitud entre los productos disponibles en dichos perfiles a partir de la confianza de las reglas de asociación descubiertas entre ellos (ver Sección 2.2.7). Gracias a estos valores de similitud, el sistema colaborativo puede seleccionar los productos del perfil del usuario activo que más se parecen al producto objetivo y predecir su nivel de interés.

  Los enfoques basados en reglas de asociación comparten la filosofía apuntada en la primera técnica, dado que también sugieren productos que aparecen en los perfiles de los usuarios con mucha frecuencia. Sólo en este caso, es posible detectar las reglas que utilizan los enfoques colaborativos durante el proceso de predicción.

- **Media ponderada de clasificaciones**: En las propuestas colaborativas basadas en usuario, un método alternativo para predecir el interés del usuario activo en relación a un determinado producto objetivo, consiste en promediar las clasificaciones que sus vecinos han asignado a dicho producto, empleando como pesos los valores de correlación entre sus respectivas preferencias. Esta idea aparece reflejada en la Ec. 2.3, donde se muestra el nivel de interés $I_{U,i}$ que predice el enfoque para un usuario activo U y un producto objetivo i.

$$I_{U,i} = \frac{\sum_{k=1}^{M} sim(U, V_k) \cdot R_{V_k}(i)}{M} \tag{2.3}$$

donde:

- M es el tamaño del vecindario de U,
- V_k es el k-ésimo vecino de U,
- $sim(U, V_k)$ es la similitud (correlación) entre las preferencias de los usuarios U y V_k, y
- $R_{V_k}(i)$ es el nivel de interés del vecino V_k en relación al producto objetivo i.

De acuerdo a esta expresión, este modelo sugiere al usuario activo aquellos productos que hayan despertado mayor interés entre los vecinos con preferencias más similares a las suyas. Este enfoque asume que todos los usuarios emplean el mismo rango de valores a la hora de asignar clasificaciones a sus preferencias. Para poder contemplar el caso en que esta premisa no se cumpla, las propuestas existentes calculan la desviación de las clasificaciones de cada usuario respecto a su valor medio, tal como se mostró en la expresión de la correlación Pearson-r (ver Ec. 2.2). El recomendador de noticias Grouplens [120] es uno de los sistemas que adopta este enfoque. En el dominio de la TV personalizada, algunos ejemplos representativos son MovieLens [150], Moviefinder™ [169], TiVo™ [241], ReplaTV™ [205] y TV Scout [17].

La media ponderada de clasificaciones también puede ser utilizada en los sistemas colaborativos basados en *item*. Este enfoque predice el interés del usuario activo, calculando la suma de las clasificaciones que haya asignado a los productos vecinos del objetivo. Dichas clasificaciones son ponderadas mediante los valores de similitud medidos entre tales vecinos y el producto objetivo, tal como muestra la Ec. 2.4.

$$I_{U,i} = \frac{\sum_{j=1}^{N} sim(i, v_j) \cdot R_U(v_j)}{N} \tag{2.4}$$

donde:

- N es el tamaño del vecindario de i,
- v_j es el j-ésimo vecino de i,
- $sim(i, v_j)$ es la similitud entre los productos i y v_j, y
- $R_U(v_j)$ es la clasificación asignada por U al producto vecino v_j.

Ventajas e inconvenientes del filtrado colaborativo

Por su propia naturaleza, los enfoques colaborativos permiten superar la falta de diversidad asociada a los métodos basados en contenido, ya que las recomendaciones elaboradas no se basan únicamente en las preferencias personales del usuario activo, sino que consideran los intereses del resto de usuarios del sistema. Esta cualidad se aprecia fácilmente en el método colaborativo basado en *item*, donde dos productos pueden ser similares (y por tanto recomendados al usuario activo) aun cuando no compartan

ningún atributo semántico; simplemente es necesario que otros usuarios del sistema los hayan clasificado simultáneamente en sus perfiles.

Sin embargo, en el filtrado colaborativo también es posible identificar algunas limitaciones importantes:

- En primer lugar, el *new user ramp-up* sigue estando presente en este tipo de sistemas, lo que dificulta enormente la formación del vecindario asociado a un usuario que acaba de llegar al sistema, cuyo perfil registrará, típicamente, un número muy reducido de preferencias.

- La segunda de las limitaciones (conocida en la literatura como *sparsity problem*) está relacionada con la noción de vecindario adoptada en los enfoques colaborativos tradicionales[1]. Sus efectos se manifiestan a medida que aumenta el número de productos disponibles en el sistema recomendador. Ante tal diversidad, es menos probable que dos perfiles contengan exactamente los mismos productos (preferencias) y, por consiguiente, es más difícil encontrar vecinos tanto para el usuario activo como para el producto objetivo, fase crucial en los enfoques colaborativos basados en usuario y en *item*, respectivamente.

- El tercero de los problemas ligados a los sistemas colaborativos es el llamado *gray sheep*, especialmente perjudicial para aquellos usuarios que, al tener preferencias muy diferentes a las del resto de usuarios, no pueden recibir recomendaciones colaborativas (debido a la imposibilidad de crear su vecindario).

- Los problemas de escalabilidad también limitan los enfoques colaborativos tradicionales. Es evidente que a medida que aumenta el número de productos y usuarios en el sistema, también se incrementa el coste computacional del proceso de cálculo del vecindario (del usuario activo o de los productos definidos en su perfil), etapa crucial en este tipo de enfoques.

- Otra de las limitaciones reconocidas en los enfoques colaborativos tiene su origen en la incorporación de nuevos productos al recomendador (por esta razón, denominada en inglés *new item ramp-up*). De lo comentado en la Sección 2.6.3, se desprende que los enfoques colaborativos convencionales sólo sugieren productos incluidos en los perfiles de los usuarios del sistema. Como consecuencia de esto último, es posible identificar un tiempo de latencia desde que el sistema conoce nuevos productos hasta que éstos son recomendados a los usuarios activos (dado que deben ser clasificados por un número representativo de usuarios antes de ser incluidos en alguna recomendaciónr).

  Tal como apunta George Karypis en [89], esta limitación es crítica en algunos dominios de aplicación (véase la televisión digital), en los que no sólo hay un flujo constante de productos nuevos, sino que, en algunos casos, éstos sólo están disponibles durante un tiempo limitado. En este escenario, es especialmente importante reducir los tiempos de latencia mencionados antes, para así poder ofrecer a los usuarios los productos más novedosos del momento sin retardos innecesarios.

[1] Recordar que los enfoques tradicionales consideran que dos usuarios son similares si clasifican los *mismos* productos, mientras que dos productos son similares si los usuarios que clasifican uno de ellos tienden a clasificar también el otro (con niveles de interés similares).

- Por último, merece la pena destacar el llamado *cold start*, cuyos efectos se ponen de manifiesto durante las primeras etapas de funcionamiento de los sistemas colaborativos. Así, hasta que no se alcanza un número suficientemente elevado de usuarios registrados, el recomendador no dispone de información suficiente para crear de una forma precisa el vecindario (del usuario activo o del producto objetivo), minando de esta forma la calidad de las sugerencias ofrecidas.

2.6.4 Enfoques híbridos

Las limitaciones identificadas tanto en el filtrado demográfico, como en los métodos basados en contenido y el filtrado colaborativo, plantearon la necesidad de combinar varias de estas estrategias para así aunar sus ventajas y mitigar sus deficiencias. Bajo esta premisa, surgen los denominados sistemas híbridos.

Uno de los modelos que goza de mayor popularidad es el que combina el filtrado colaborativo y los métodos basados en contenido. La mayoría de las propuestas existentes adoptan un paradigma conocido en la literatura como *collaboration via content*, definido por Pazzani en 1999 [189]. Este autor propone calcular la similitud entre los usuarios del sistema utilizando tanto las descripciones semánticas de los productos definidos en sus perfiles (empleadas en los métodos basados en contenido), como los niveles de interés asignados a los mismos (considerados en el filtrado colaborativo).

Son muchos los sistemas de personalización que han adoptado este enfoque en dominios de aplicación muy diversos: Amazon [4], CDNow [44] y WebSell [57] (comercio electrónico); Fab [14], WebWatcher [9] y LaboUr [220] (Web personalizada); Anatagonomy [112] y NewsWeeder [130] (filtrado de noticias); y Recommender [16], PTV [55], PTVPlus [229] y Nakif [76] (TV personalizada).

Tal como apuntábamos al comienzo, los enfoques híbridos permiten neutralizar algunas de las limitaciones (individuales) asociadas a los métodos basados en contenido y al filtrado colaborativo:

- Por una parte, estos modelos permiten superar el excesivo parecido existente entre las recomendaciones basadas en contenido y las preferencias del usuario activo, gracias a que el enfoque colaborativo considera la experiencia del resto de usuarios del sistema, diversificando con ello las sugerencias ofrecidas.

- Por otro lado, el filtrado basado en contenido también contribuye a aliviar, en algunos casos, los tiempos de latencia que retrasan la recomendación de los productos que acaban de llegar a un sistema colaborativo (ver Sección 2.6.3). Esto es debido a que en un esquema híbrido, cualquier producto puede ser recomendado al usuario activo sin retardo alguno, siempre que los métodos basados en contenido consideren que se adapta a sus preferencias. Dadas las limitaciones de las métricas de similitud utilizadas en estos métodos, el problema de la latencia sólo será eliminado cuando los productos considerados sean muy parecidos a dichas preferencias; en caso contrario, serán sugeridos con retardos innecesarios debido a la inflexibilidad de las métricas adoptadas en los enfoques actuales.

- Por último, destacar que los enfoques híbridos también mitigan, en cierta medida, los efectos del *sparsity problem* ya que, gracias a las descripciones semánticas consideradas en el filtrado basado en contenido, pueden detectar que dos productos son similares aun cuando no hayan sido clasificados simultáneamente en los

perfiles de los usuarios. Sin embargo, esta solución aún no es efectiva, debido a las limitaciones de las métricas de similitud (sintácticas) adoptadas tradicionalmente.

Para una revisión exhaustiva de los enfoques híbridos se recomienda la consulta de [42], donde Robin Burke define hasta siete métodos diferentes para combinar estrategias de personalización de diferente naturaleza:

- **Ponderado** (*Weighted*): En un sistema híbrido ponderado se decide si un determinado producto objetivo es sugerido o no al usuario considerando las salidas de *todas* las estrategias de recomendación que conviven en el mismo. Como su nombre indica, algunas de las estrategias pueden tener más peso que otras durante el proceso de decisión, de forma que si las más determinantes deciden recomendar el producto, éste será ofrecido al usuario.

- **Conmutación** (*Switching*): En este caso, en lugar de ejecutar todas las estrategias simultáneamente, el sistema emplea algún criterio para conmutar entre ellas: cuando se cumplen ciertas condiciones el sistema emplea una estrategia y, en caso contrario, recurre a las restantes.

- **Mixto** (*Mixed*): Este esquema reúne en una misma recomendación productos que han sido sugeridos mediante las diferentes estrategias implementadas en el sistema híbrido.

- **Combinación de características** (*Feature combination*): En este modelo híbrido se funden en un único conjunto los datos que utilizan las diferentes estrategias de personalización, y con éste se ejecuta un solo algoritmo de recomendación. Por ejemplo, si se combinan los métodos basados en contenido y el filtrado colaborativo, este esquema podría utilizar la información colaborativa (clasificaciones de los usuarios) como una característica más de las preferencias de los usuarios, sobre las que finalmente se ejecutaría el filtrado basado en contenido.

- **Cascada** (*Cascade*): Un sistema híbrido en cascada funciona en dos etapas: primero se ejecuta una de las estrategias de recomendación sobre las preferencias del usuario, obteniendo un primer conjunto de productos candidatos a ser incluidos en la recomendación final; a continuación, una segunda estrategia refina la recomendación y selecciona sólo algunas de las sugerencias obtenidas en la primera etapa.

- **Incorporación de características** (*Feature augmentation*): En este esquema, una de las estrategias se emplea para calcular una clasificación para un producto concreto. A continuación, esa información se incorpora como dato de entrada para las siguientes técnicas de recomendación; en otras palabras, la salida de una de las estrategias de recomendación se utiliza como entrada en las siguientes.

  La diferencia fundamental entre este esquema híbrido y el modelo en cascada, es que, en este último, la segunda estrategia sólo trabaja sobre el conjunto de productos candidatos obtenidos por la primera de ellas, prescindiendo de cualquier tipo de información adicional calculada por ésta. Por el contrario, en el modelo basado en incorporación de características, toda la información que proporcione la primera de las técnicas, se utiliza en la(s) siguiente(s).

- **Metanivel** (*Meta-level*): En este caso, el modelo *completo* generado por una de las estrategias se utiliza como entrada en las restantes. La diferencia entre este sistema híbrido y el basado en incorporación de características, es que en este último, el modelo aprendido sólo se utiliza para generar características que se usan como entrada en las siguientes estrategias, mientras que en el híbrido de metanivel se utiliza todo el modelo como dato de entrada.

  Un esquema de este tipo estaría presente en un sistema híbrido que combine una primera etapa basada en contenido y una segunda colaborativa de la siguiente forma: la primera etapa construye un modelo de las preferencias del usuario (por ejemplo, un vector de preferencias y pesos asociados como el descrito en la Sección 2.2.2). A continuación, este modelo es empleado en la fase colaborativa para comparar las preferencias de los usuarios y formar los vecindarios oportunos, antes de elaborar las recomendaciones personalizadas.

2.7 Conclusiones

En este capítulo se han clasificado los sistemas recomendadores definidos en el estado del arte, con arreglo a los diferentes tipos de mecanismos que utilizan durante el proceso de personalización. Tales mecanismos incluyen:

- la representación de los perfiles de usuario,
- la generación del perfil inicial durante la fase de registro del usuario en el sistema,
- las técnicas que utilizan los usuarios a la hora de proporcionar información sobre la relevancia o precisión de las recomendaciones recibidas,
- los métodos empleados para lograr la adaptación del recomendador a medida que las preferencias de los usuarios evolucionan en el tiempo y, finalmente,
- las estrategias de filtrado de información utilizadas durante la elaboración de las recomendaciones.

Entre dichas estrategias cabe destacar el filtrado demográfico, los métodos basados en contenido, el filtrado colaborativo y los enfoques híbridos.

- Filtrado demográfico: Esta técnica basa sus recomendaciones en las características personales de los usuarios (edad, sexo, estado civil, profesión, etc.), de forma que todos aquellos que compartan un mismo estereotipo, recibirán las mismas sugerencias personalizadas.
- Filtrado basado en contenido: Esta estrategia recomienda a cada usuario productos similares a los que le interesaron en el pasado. Para poder comparar sus preferencias personales frente a los productos disponibles en el recomendador, el filtrado basado en contenido considera los principales atributos de los productos, así como su frecuencia de aparición en conjuntos de entrenamiento predefinidos.
- Filtrado colaborativo: En lugar de basarse únicamente en las preferencias personales de cada usuario activo, el filtrado colaborativo también considera los intereses de otros usuarios con gustos similares a los suyos. En los enfoques tradicionales, tal similitud se estima a partir de los niveles de interés definidos

en los perfiles de los usuarios del sistema, prescindiendo de las descripciones semánticas utilizadas en los métodos basados en contenido.

- Enfoques híbridos: Como su nombre indica, este tipo de enfoques combinan estrategias de diversa naturaleza (principalmente, filtrado basado en contenido y filtrado colaborativo).

A pesar de su implantación en diferentes sistemas reales, los enfoques enumerados comparten una misma limitación, ya que están basados en técnicas, más o menos sofisticadas, que omiten gran cantidad de información útil durante el proceso de personalización. Esta información está íntimamente relacionada con la semántica (es decir, el significado) de los atributos correspondientes a los productos disponibles en el sistema recomendador. La incorporación de la semántica en las estrategias de personalización adoptadas en los sistemas existentes, habilitaría ciertos procesos de razonamiento capaces de descubrir relaciones (semánticas) interesantes entre los productos disponibles y las preferencias personales de los usuarios. Mediante este tipo de razonamientos semánticos, el recomendador no sólo dispondría de más conocimiento para poder elaborar sus sugerencias, sino que, además, combatiría de forma efectiva las limitaciones identificadas en los enfoques actuales.

Para inferir tales relaciones semánticas, los sistemas recomendadores requieren metodologías que permitan consultar la base de conocimiento en la que se representen las descripciones semánticas de los productos, y recuperar información personalizada desde la misma. La revisión de este tipo de enfoques, vinculados al campo de la Web Semántica, es abordada en el siguiente capítulo.

Capítulo 3

Estado del arte en Web Semántica

Este capítulo describe un compendio de propuestas ligadas al dominio de la Web Semántica. Esta revisión incluye: (i) los enfoques utilizados en la misma para la representación y organización del conocimiento, (ii) la descripción de su arquitectura, y (iii) los mecanismos definidos en el ámbito del razonamiento semántico. En este sentido, el capítulo aborda tanto los enfoques inferenciales más sencillos como aquellos que, gracias a capacidades de razonamiento más sofisticadas, exploran grandes redes de conocimiento y descubren relaciones ocultas en las mismas.

3.1 Introducción

De acuerdo a la definición dada por Tim Berners-Lee [19], la Web Semántica es una extensión de la Web actual en la que se expresa el significado de la información mediante metadatos o anotaciones semánticas, que las máquinas pueden procesar de forma automática. Con ello, no sólo se amplía la interoperabilidad entre las máquinas, sino que además se reduce la necesidad de mediación de los agentes humanos en la realización de ciertas tareas que podrían ser tediosas para ellos. Uno de los aspectos claves en este dominio es la capacidad de descubrir nueva información a partir del conocimiento inicialmente aprendido, a través del establecimiento de relaciones semánticas entre los recursos Web descritos (anotados) mediante metainformación. En la literatura, este proceso de descubrimiento de conocimiento recibe el nombre de *inferencia* o *razonamiento*, y se lleva a cabo en los denominados *sistemas basados en conocimiento*.

En este capítulo se revisa el estado del arte en la Web Semántica, haciendo especial hincapié en los enfoques propuestos para organizar y representar formalmente la información manejada, almacenada en *bases de conocimiento*. Asimismo, analizamos la arquitectura que los creadores de la Web Semántica han propuesto para ésta, describiendo los elementos y funcionalidades asociadas a cada una de las capas establecidas en la misma.

Nuestra atención también se centra en los enfoques más relevantes propuestos en el ámbito del razonamiento semántico y, más concretamente, en los campos de la explo-

ración de bases de conocimiento y el establecimiento de relaciones semánticas entre las entidades formalizadas en las mismas. A la revisión de este tipo de propuestas dedicamos una parte importante del presente capítulo.

3.2 Representación del conocimiento

En esta sección describimos algunos conceptos de interés relativos a la *representación de conocimiento*, una disciplina vinculada al campo de la Inteligencia Artificial, estrechamente relacionada con los sistemas basados en conocimiento y el dominio de la Web Semántica.

3.2.1 ¿Qué es la representación de conocimiento?

La representación de conocimiento es una rama de la Inteligencia Artificial centrada en el diseño e implementación de lenguajes y sistemas orientados a expresar el conocimiento sobre el entorno. En realidad, esta disciplina constituye un medio para expresar conocimiento sobre un dominio de aplicación específico utilizando para ello descripciones semánticas. El objetivo principal es aplicar la información que proporcionan tales descripciones en procesos de razonamiento automáticos y computacionalmente eficientes.

En general, el lenguaje utilizado para representar el conocimiento determina en gran medida la clase de razonamiento que puede llevarse a cabo a partir del mismo, de ahí que la representación sea un paso previo al razonamiento sobre el conocimiento.

Por su propia naturaleza, los orígenes de la representación de conocimiento están asociados a las redes semánticas desarrolladas entre los años 1960 y 1970. En la literatura, se asume que la noción de red semántica fue propuesta por Quillian en una publicación desarrollada durante su estancia en el MIT (*Massachusetts Institute of Technology*) [199].

3.2.2 Redes semánticas

Tal como introdujimos brevemente en la Sección 2.2.5, las redes semánticas son grafos acíclicos dirigidos en los que los conceptos y las relaciones existentes entre ellos se representan mediante nodos y enlaces, respectivamente.

Originariamente, las redes semánticas estaban organizadas y estructuradas de forma heterogénea, mediante una semántica que dependía únicamente del grupo de investigadores que las habían desarrollado y del tipo de sistema en el que éstas se aplicaban. Dicho de otra forma, los métodos de razonamiento automático de este tipo de redes semánticas estaban basados en implementaciones particulares, propietarias, y no en lenguajes formales adoptados de forma generalizada entre la comunidad investigadora.

Conscientes de tal limitación, a finales de la década de los 70, los investigadores del campo se centraron en un objetivo común: la formalización coherente de una red semántica como un lenguaje basado en lógica y orientado a la representación de conocimiento. Uno de los frutos iniciales de este proceso fue KL-ONE, un lenguaje basado en marcos (*frames*) que surgió a principios de la década de los 80, y que está considerado en la literatura como la primera conceptualización formal de una red semántica [32].

3.2.3 Representación de conocimiento basado en marcos

Los marcos son simplemente estructuras de datos que permiten representar conceptos (llamados *clases*) y relaciones entre ellos (llamados *slots*). La noción de marco está inspirada en la terminología utilizada en el lenguaje de programación LISP, donde los atributos (a veces también llamados *propiedades*) permiten identificar las relaciones establecidas en un dominio de aplicación específico [154].

En los sistemas basados en marcos, el conocimiento de un dominio está formalmente representado mediante *entidades*, referidas tanto a las clases del dominio como a sus instancias concretas. Por su parte, en un lenguaje basado en marcos, una clase y todas sus instancias tienen asociadas un conjunto de características comúnmente aceptadas en el dominio concreto de una ontología. Entre dichas características destacan:

- la organización jerárquica de las clases,
- la posibilidad de que una clase tenga varias superclases,
- la capacidad de que las instancias hereden todas las propiedades y sus atributos desde la(s) clase(s) a la(s) que pertenecen, y
- la posibilidad de declarar valores por defecto para dichos atributos.

La perspectiva de marco difiere de la perspectiva de otros lenguajes de representación de conocimiento definidos en el dominio de las *Lógicas Descriptivas* (DL - *Description Logics*), en las que el conocimiento se formaliza mediante un conjunto de *axiomas*. Sin embargo, la inmensa mayoría de los lenguajes de representación de conocimiento modernos, así como los sistemas basados en éstos, tienen en cuenta ambas perspectivas, razón por la cual en la literatura se ha acabado asumiendo con naturalidad una relación de equivalencia entre las representaciones basadas en marcos y las basadas en lógica.

3.2.4 Lógicas Descriptivas

De la misma forma que las redes semánticas fueron evolucionando hacia los sistemas basados en marcos, estos últimos también condujeron al desarrollo de diferentes Lógicas Descriptivas. De hecho, el lenguaje basado en marcos KL-ONE [32] también se consolidó como primera DL, ya que además de soportar la representación basada en marcos, estaba formalizado lógicamente mediante estructuras formadas por *descripciones*.

En el dominio de las DL, una descripción es una expresión en un lenguaje formal que define un conjunto de instancias. Por tanto, una DL es un lenguaje para representar conocimiento que tiene una sintaxis y una semántica, de forma que la sintaxis se utiliza para construir las descripciones, mientras que la semántica define el significado de cada una de ellas.

Matemáticamente, en una DL pueden existir relaciones que no sean *decidibles*. Esto significa que cuando se formula una consulta a un sistema formalizado mediante dicha DL, es posible que éste no devuelva ningún resultado por llevarse a cabo la búsqueda sobre un espacio infinito. Por este motivo, las DL optan por limitar el tipo de conocimiento que puede ser formalizado (y con ello, la expresividad del lenguaje

de representación resultante), de forma que los problemas de *decidibilidad* puedan ser obviados en la implementación del sistema (ver Sección 3.2.1).

Precisamente en relación a esto último, podemos destacar que son dos las funciones encomendadas a las DL en el campo de la representación de conocimiento:

- En primer lugar, tal como dijimos a lo largo de la sección, deben proporcionar formalismos lógicos para poder representar y expresar conocimiento.
- Por otra parte, deben facilitar métodos de razonamiento correctos, resolubles y fundamentados en bases lógicas y teóricas sólidas.

Para cumplir ambos objetivos, los lenguajes de representación de conocimiento basados en DL deben combinar dos elementos fundamentales:

- Expresividad a la hora de formalizar el conocimiento: esta capacidad permite describir detalladamente la semántica del dominio de aplicación concreto representado mediante la DL.
- Métodos de razonamiento resolubles: Son imprescindibles para asegurar procesos de inferencia correctos que permitan a los sistemas basados en DL descubrir nuevo conocimiento de forma automática. En general, la obtención de este tipo de métodos de razonamiento supone el desarrollo de lenguajes lógicos con capacidades expresivas más limitadas.

Algunos de los lenguajes basados en DL que gozan de mayor popularidad en el estado del arte son LOOM [140], Classic [188], KRIS [11], FaCT [100] y RACER [85]. Los dos últimos, tal como veremos en este capítulo, se han incorporado en diferentes sistemas para poder razonar de forma automática en el dominio de la Web Semántica (precisamente en la literatura es posible encontrar referencias a los citados lenguajes bajo el término de *razonadores basados en DL*).

En esta sección, únicamente hemos revisado los principios fundamentales en los que se basan las DL. Para una descripción exhaustiva de las mismas, fuera del ámbito de esta tesis, se remite al lector a [12].

3.3 Organización del conocimiento

Son cuatro las estructuras más importantes propuestas en el estado del arte para organizar el conocimiento manejado en la Web Semántica. En orden creciente de acuerdo a su complejidad, tales estructuras son las taxonomías, los mapas temáticos, los tesauros y las ontologías.

3.3.1 Taxonomías

Las taxonomías son estructuras que permiten clasificar la información jerárquicamente. Esta información se representa mediante clases, más generales o más específicas, organizadas en el árbol de la taxonomía a partir de una única clase raíz. Dichas clases están dispuestas de forma que a medida que se desciende por la jerarquía desde el elemento raíz, las entidades son más específicas; mientras que, por el contrario,

al ascender hacia la raíz, se encuentran clases más generales (abstractas). Este tipo de clasificación se conoce con el nombre de *generalización/especialización*.

Habida cuenta de su organización jerárquica, las taxonomías modelan relaciones de herencia o inclusión (en terminología inglesa *subclassOf*) y, en algunos casos, relaciones de pertenencia (denominadas genéricamente *partOf*). En este sentido, en la literatura es común encontrar el término *subsumption principle* (en adelante, *principio de inclusión*), según el cual se cumple lo siguiente:

Siendo $\mathcal{T}$ una taxonomía, y a y b dos instancias pertenecientes a las clases C_a y C_b, respectivamente, se dice que a *subsume* a b si C_b es subclase de C_a (es decir, C_b *está incluido en* C_a).

Debido a su estructura en forma de árbol, en las taxonomías es posible identificar diferentes conceptos de interés, rescatados, en muchos casos, de la teoría de grafos [64]:

- **Profundidad**: La profundidad de a (representada como $depth(a)$) se define como el número de relaciones jerárquicas (de inclusión) existentes entre C_a y la clase raíz de la taxonomía $\mathcal{T}$ a la que pertenece.
- **Clases hojas**: Son clases de las que no dependen otras más específicas, es decir, aquellas que carecen de subclases en la estructura jerárquica.
- **Antecesor Común más Cercano** (LCA - *Lowest Common Ancestor*): Se define el LCA entre a y b (representado típicamente como $LCA_{a,b}$) como la clase más profunda que es antecesor tanto de C_a como de C_b en la taxonomía $\mathcal{T}$.

Las capacidades de clasificación disponibles en las taxonomías fueron aplicadas más tarde en unas estructuras llamadas mapas temáticos, que complementan esta funcionalidad con otras adicionales para así poder indexar el espacio conceptual de información disponible en la Web.

3.3.2 Mapas temáticos

Los mapas temáticos (*Topic Maps*) son una tecnología implantada con éxito en la Web que permite modelar y representar el conocimiento de forma intercambiable y procesable por las máquinas y, a la vez, ofrece un marco unificado para la gestión del conocimiento y de la información que también es comprensible para los humanos. Este tipo de tecnologías permiten, además, describir estructuras de conocimiento y asociarlas con otros recursos de información.

Como su nombre indica, un mapa temático es una colección de temas enlazados entre sí, que permiten indexar un conjunto de documentos basándose en su contenido. Los temas, asociaciones (entre temas) y apariciones (de los temas en los documentos) conforman un mapa temático, lo que comúnmente se conoce con el nombre de TAO (*Topics*, *Associations*, *Occurrences*). Los conjuntos de documentos así etiquetados ayudarán a los usuarios a encontrar información acerca de un tema a lo largo de una gran variedad de documentos.

Entre los enfoques que aplican este tipo de capacidades de los mapas temáticos durante procesos de búsqueda de información, destacamos [131] y [144]. El primero de ellos implementa un motor de búsqueda basado en construir perfiles para cada

usuario y valorar la relevancia de diferentes sitios Web para éstos en función de sus preferencias. Para ello, el enfoque emplea un mapa temático que permite identificar rápidamente aquellos temas que aparecen tanto en el perfil del usuario como en los sitios Web analizados y, con ello, entregar información de interés para éste. El segundo enfoque emplea los mapas temáticos para extender una base de datos relacional en la que se almacena información personal sobre un grupo de usuarios. Esta estrategia permite navegar de forma eficiente por la base de datos y descubrir nuevo conocimiento sobre las consultas (sirviéndose de los conceptos y relaciones representados en el mapa temático), mejorando así la calidad de los resultados devueltos.

Los mapas temáticos no sólo sirven para organizar la información y facilitar su recuperación, sino que también permiten representar el conocimiento de forma normalizada, de ahí su exitosa implantación en el dominio de la Web Semántica, tal como se describe en [97]. En concreto, esta iniciativa de normalización, que nace en 2001 bajo el nombre de XML Topic Maps (XTM) [193], fue impulsada por los miembros del grupo TopicMaps.Org. Esta norma provee un modelo abstracto y una gramática para representar la estructura de la información, definiendo los temas y las asociaciones establecidas entre éstos. Los detalles de la misma, fuera del ámbito de esta tesis, pueden consultarse en [193].

En la literatura también se han definido estructuras más complejas que las taxonomías y los mapas temáticos, que permitan completar el conocimiento que se puede representar en la Web Semántica y, con ello, mejorar las inferencias llevadas a cabo desde el mismo.

3.3.3 Tesauros

Un tesauro es un vocabulario controlado y estructurado formalmente, compuesto por términos entre los que se establecen relaciones semánticas genéricas.

Los términos contenidos en un tesauro responden al análisis del texto o materia tratada en la colección de documentos considerados. Así, un único tema se representa mediante una serie de ideas o conceptos que se pueden describir por medio de términos o descriptores. El tesauro incorpora todos esos términos en una base de datos y cada uno de ellos se convierte en un punto de acceso para la recuperación del documento. La potencia de un tesauro radica en la posibilidad de combinar todos los términos o descriptores, lo que le convierte en un lenguaje rico y potente. Un tesauro es pues, una herramienta de control terminológico muy útil para el análisis, descripción y recuperación automática de información.

Estructura de un tesauro

Son dos los elementos claves identificados en la estructura de un tesauro:

- *Unidades léxicas*: Dentro de éstas, podemos distinguir dos tipos:
    - *Descriptores*: Se corresponden con los términos permitidos en el tesauro. Un descriptor es una palabra o conjunto de palabras del lenguaje corriente que expresan un concepto. La relación entre ambos es unívoca, de forma que cada concepto se expresa mediante un único término y ese término responde a un único concepto. Dicho término, utilizado para el análisis y

la recuperación de los documentos descritos en el tesauro, puede ser de diferentes tipos:

- *Términos simples o unitérminos*: se emplean cuando el concepto es claro en sí mismo, sin necesidad de añadir ninguna otra palabra, o cuando se usa como genérico de términos más concretos o específicos.
- *Términos compuestos*: la mayoría de los términos de un tesauro adquieren su verdadero sentido cuando se convierten en términos compuestos, ya que así se logra una mayor especificidad. La estructura sintáctica de los términos compuestos suele ser la siguiente:
  - sustantivo + adjetivo (por ejemplo, "documentos digitales"),
  - sustantivo + sintagma preposicional ("documentos de archivo"),
  - sustantivo + sintagma aposicional ("documentos HTML").

- *No descriptores*: Para poder eliminar la sinonimia del lenguaje natural durante la recuperación de información desde un tesauro, se elige un único término (de entre todos aquellos que son sinónimos) y se rechaza el resto. Los no descriptores son precisamente este conjunto de términos desechados.

- *Relaciones semánticas*: Son todas aquellas relaciones que se establecen entre los términos de un tesauro. Los tesauros incorporan hasta tres tipos diferentes de relaciones semánticas:
  - *Relaciones de equivalencia*: Se derivan de la gran proximidad fonética o sémica entre los descriptores y no descriptores (sinónimos, cuasi-sinónimos, etc.).
  - *Relaciones jerárquicas*: Son relaciones asimétricas establecidas entre los descriptores, en la que un término es superior o genérico de otro. Esta estructura clasificatoria distingue al tesauro de las listas alfabéticas de materias o palabras clave, así como de diccionarios y léxicos. De esta forma, se crean campos conceptuales que clasifican los términos de los documentos y los agrupan en disciplinas organizadas jerárquicamente.
  - *Relaciones asociativas*: Permiten asociar descriptores que contienen conceptos que comparten un lexema común (por ejemplo, "universidad" y "universitario").

Entre los tesauros propuestos en la literatura merecen especial mención *Roget's Thesaurus* [211] y *WordNet* [254]. El primero puede ser considerado la herramienta pionera en este campo de investigación. Los creadores de los tesauros que surgieron más tarde optaron por sentar sus bases sobre él y completar la información que contenía, a fin de mejorar los procesos de recuperación de información a partir de este tipo de herramientas. En la actualidad, el tesauro que ha alcanzado un mayor grado de madurez es WordNet, una base de datos léxica para el lenguaje Inglés incorporada con gran éxito en numerosas aplicaciones de búsqueda en la Web [161, 162, 166, 172].

3.3.4 Ontologías

El origen del término ontología se sitúa en el campo de Filosofía, en el que se define como "la rama de la metafísica que estudia la naturaleza de la existencia". Este concepto fue rescatado desde el área de la Inteligencia Artificial, en el que Gruber definió

una ontología como una "especificación explícita y formal de una conceptualización compartida" [82]. En esta definición, convertida ya en estándar, *conceptualización* se refiere a un modelo abstracto del dominio de aplicación específico que se representa en la ontología; *explícita* hace referencia a la necesidad de especificar los diferentes conceptos incluidos en la misma; *formal* indica que la especificación debe llevarse a cabo mediante un lenguaje formalizado, y *compartida* alude al hecho de que una ontología debe representar conocimiento consensuado, ampliamente aceptado, al menos en la comunidad de usuarios en la que ésta vaya a ser utilizada.

Una definición más concreta de ontología fue dada por Weigand en 1997, en la que se refería a este tipo de estructuras como "bases de datos que describen los conceptos de un determinado dominio de aplicación, así como sus propiedades y las relaciones existentes entre ellos" [250]. Ésta es precisamente la noción de ontología utilizada en el dominio de la Web Semántica [233].

Además de ofrecer la capacidad de formalizar un dominio específico, las ontologías permiten que el conocimiento que en ellas se representa pueda ser reutilizado fácilmente, tanto entre usuarios como entre agentes software. Esta característica es la que ha convertido a las ontologías en la piedra angular de la Web Semántica en lo que a conceptualización de conocimiento compartido se refiere.

Los elementos claves de una ontología, cuya descripción más detallada será abordada a lo largo del capítulo, son las clases y las propiedades. Las primeras identifican los conceptos representativos del dominio de la ontología, mientras que las segundas se refieren a las relaciones existentes entre dichos conceptos. Una vez definidas las clases y propiedades, es necesario representar en la base de conocimiento instancias específicas de las mismas[1]. Junto a éstas, las ontologías también definen reglas, declaraciones lógicas utilizadas habitualmente para modelar conocimiento que no puede ser representado mediante los tres componentes enumerados anteriormente (clases, propiedades e instancias). Estas reglas son fácilmente interpretadas por los usuarios humanos y además, pueden ser utilizadas de forma eficiente en cálculos computacionales llevados a cabo por agentes software. En función de su propósito, dichas reglas se pueden clasificar en tres grupos:

- Reglas para la creación de conocimiento: Permiten obtener nuevas sentencias lógicas a partir de la información almacenada en la base de conocimiento.

- Restricciones: Indican las propiedades que debe cumplir el modelo formalizado en la ontología, y permiten además detectar posibles inconsistencias en el mismo.

- Reglas reactivas: Determinan las acciones que debe tomar un sistema basado en conocimiento como consecuencia del cumplimiento de ciertas condiciones (de ahí que cada regla conste de un antecedente y un consecuente).

[1] Algunos enfoques hacen una distinción entre ontologías y bases de conocimiento, argumentando que las primeras representan las clases y las propiedades, mientras que las segundas incluyen sus instancias específicas. En esta tesis, asumimos que ambos términos son equivalentes, de ahí que los utilicemos indistintamente a la hora de referirnos al conjunto de clases, propiedades e instancias formalmente representadas en la conceptualización.

3.4 Arquitectura de la Web Semántica

A lo largo de este capítulo, hemos presentado la Web Semántica como una extensión de la Web tradicional en la que se anotan los recursos disponibles mediante descripciones semánticas. Por ello, no debe extrañar que su arquitectura consista en un conjunto de capas –de forma que cada nueva capa se construye a partir de las anteriores–, cuyas bases se asientan sobre los lenguajes comúnmente utilizados en el dominio de la Web (como XML), tal como se muestra en la Fig. 3.1.

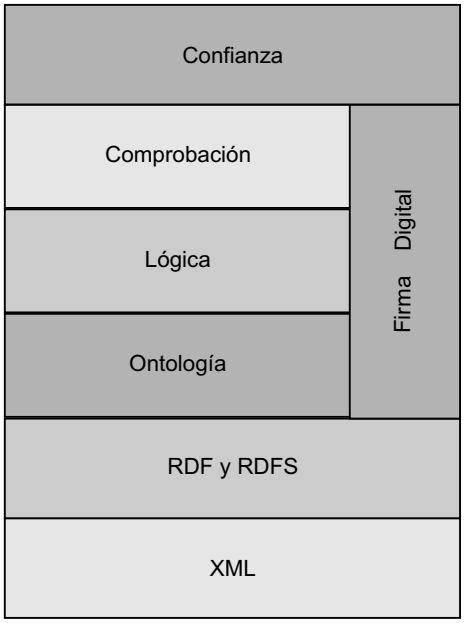

Figura 3.1: Arquitectura propuesta para la Web Semántica

Dichos lenguajes son especialmente importantes en la Web Semántica, ya que permiten el intercambio y la reutilización de conocimiento –a través de las ontologías– de una forma normalizada y consensuada. Por otra parte, dado que cualquiera puede acceder a la Web e incorporar cualquier tipo de información, es necesario definir mecanismos que aseguren la confianza de los datos utilizados en el dominio de la Web Semántica. Precisamente para este propósito se incluyó en la arquitectura la firma digital representada en la Fig. 3.1.

3.4.1 Capa XML

En [34] se define XML (*eXtensible Markup Language*) como un conjunto de reglas para definir etiquetas semánticas que permiten organizar los documentos de texto de una forma estructurada. Sin embargo, la adopción generalizada de este lenguaje fue debida a su capacidad para intercambiar datos (estructurados) arbitrarios, lo que convirtió a XML en uno de los elementos claves para la compartición de información en el dominio de la Web.

La especificación XML proporciona los medios necesarios para declarar estructu-

ras de datos muy simples que son almacenadas en documentos que pueden procesar las máquinas. Sin embargo, este lenguaje adolece de una importante carencia, ya que está definido únicamente a nivel sintáctico, de forma que las máquinas no pueden determinar a partir de sus declaraciones el significado correcto de cada etiqueta XML utilizada en un documento conforme a este formato.

Conscientes de esta deficiencia, los integrantes del consorcio W3C optaron por diseñar nuevos lenguajes, en los que se añade una semántica formal a las especificaciones sintácticas definidas en XML. Como resultado de este proceso surgen RDF [18] y RDF Schema (RDFS) [36].

3.4.2 Capa RDF y RDF Schema

RDF (*Resource Description Framework*) es el lenguaje básico en el dominio de la Web Semántica. Un modelo RDF consta de un conjunto de triplas formadas por un sujeto, un predicado y un objeto. El sujeto identifica el recurso que se describe mediante RDF, el predicado se refiere a una característica concreta de dicho recurso y el objeto indica el valor que toma el predicado para el sujeto especificado. Los recursos en RDF están unívocamente identificados mediante URIs (*Universal Resource Identifiers*), de forma que es posible designar nuevos recursos que deben ser descritos mediante RDF sin más que asignar una URI a cada uno de ellos.

RDF Schema (RDFS) añade ciertas funcionalidades al lenguaje RDF, entre las que podemos destacar:

- Definición de clases y sus instancias: En RDFS aquellos objetos que comparten características similares se organizan en clases, y los individuos pertenecientes a estas últimas se identifican mediante instancias.
- Declaración de propiedades binarias entre objetos: RDFS permite establecer relaciones tanto entre clases como entre instancias concretas de estas últimas.
- Organización jerárquica de clases y propiedades: El *principio de inclusión* puede aplicarse tanto a las clases como a las propiedades definidas en una ontología RDFS.
- Tipos de propiedades y restricciones de rango y dominio: Este lenguaje permite restringir las entidades (clases e instancias) que pueden actuar como rango o como dominio en una propiedad concreta. El dominio de una propiedad identifica a aquellos objetos a las que se puede asociar la propiedad, mientras que el rango identifica las entidades que pueden tomar como valor dicha propiedad. En función del rango, es posible distinguir dos tipos de propiedades: (i) *Object*, donde el rango es una instancia de una clase de la ontología, y (ii) *Datatype*, cuyo rango corresponde a tipos de datos concretos (enteros, cadenas, etc.).

Si bien mediante las funcionalidades de RDFS que acabamos de describir ya es posible formalizar ontologías en la Web Semántica, el conocimiento representado en las mismas aún podría ser completado de forma considerable. Entre las características que se echan en falta en este lenguaje destacan:

- Definición del ámbito local de las propiedades: En RDFS no es posible declarar restricciones de rango y dominio aplicables sólo sobre algunas de las clases de la

ontología. Por ejemplo, en la propiedad *padreDe* no es posible distinguir que el padre de un persona pertenece siempre a la clase *Personas*, mientras que el padre de un animal es siempre una instancia de la clase *Animales*.

- Disyunción entre clases: En RDFS sólo es posible relacionar dos clases mediante herencia (o mediante propiedades), siendo imposible declarar que éstas son disjuntas.

- Combinación *booleana* de clases: RDFS no contempla la posibilidad de combinar varias clases de la ontología mediante los operadores *booleanos* típicos, tales como unión, intersección y complemento.

- Restricciones de cardinalidad: Este tipo de funcionalidad permite limitar los distintos valores que puede tomar una propiedad. Tales restricciones permiten establecer también requisitos de cardinalidad mínima, máxima o fijar ésta a un valor concreto (por ejemplo, es posible declarar que la propiedad *tienePadre* de la clase *Persona* tiene una cardinalidad exactamente igual a uno).

- Definición de características especiales de las propiedades: RDFS no soporta la declaración de tipos especiales de propiedades, tales como propiedades transitivas (por ejemplo, *mayorQue*), únicas (*tienePadre*) o inversas (*hijoDe* y *padreDe*).

Este tipo de funcionalidades se han incorporado en lenguajes con una capacidad expresiva más rica que la ofrecida por RDFS, entre los que podemos citar DAML [58], OIL [71], DAML+OIL [59] y OWL [147].

3.4.3 Capa de Ontologías

De todas las capas incluidas en la torre mostrada en la Fig. 3.1, es la correspondiente a las ontologías la que ha alcanzado un mayor grado de madurez. Esto último es debido, en gran parte, a la aparición de OWL (*Web Ontology Language*), el lenguaje más expresivo actualmente definido en el dominio de la Web Semántica y estandarizado por el consorcio W3C [147].

Su sintaxis y semántica están basadas en gran medida en DAML+OIL, primer candidato sobre el que trabajaron los desarrolladores del modelo OWL, centrándose en tareas relacionadas con las capacidades de expresividad y razonamiento ofrecidas. En OWL podemos diferenciar tres niveles; en orden creciente de acuerdo a estas capacidades, tales niveles son: OWL Lite, OWL DL (de *Description Logic*) y OWL Full. A costa de introducir costes computacionales mayores y posibles problemas de *decidibilidad*, los niveles más altos del lenguaje extienden los inferiores, de forma que en cada uno de ellos se van incorporando nuevas funcionalidades. Así, una conclusión válida en el dominio de OWL Lite, es también válida en OWL DL y OWL Full, mientras que una conclusión válida en OWL DL lo es también en OWL Full, pero no necesariamente en OWL Lite. De forma análoga, la validez de una conclusión en OWL Full no tiene por qué verificarse ni en OWL DL ni en OWL Lite.

OWL Lite

Entre las principales características de OWL Lite podemos destacar:

- Permite definir clases y propiedades, así como instancias concretas de ambas (mediante los constructores *owl:Class*, *rdf:Property* y *rdfs:type*, respectivamente).

- Es posible declarar relaciones de herencia entre las clases y también entre las propiedades (utilizando el constructor *rdfs:subClassOf* y *owl:subproperty*, respectivamente).

- Permite declarar que dos clases o propiedades son equivalentes mediante los constructores *owl:equivalentClass* y *owl:equivalentProperty*, respectivamente. Se dice que dos clases (o propiedades) son equivalentes si tienen exactamente las mismas instancias (por ejemplo, las clases *Coche* y *Automóvil*).

- OWL Lite también permite definir una relación de igualdad entre dos instancias concretas, mediante el constructor *owl:sameAs* (por ejemplo, *Clinton* y *Bill Clinton*).

- Contempla diferentes tipos de propiedades, de forma que éstas pueden establecerse entre instancias específicas de una clase, o bien entre una instancia y un tipo de datos genérico XML (mediante los constructores *owl:objectProperty* y *owl:datatypeProperty*, respectivamente).

- En OWL Lite, las propiedades pueden ser: simétricas (*owl:SymmetricProperty*), transitivas (*owl:TransitiveProperty*), inversas (*owl:inverseOf*), funcionales y funcionales inversas (*owl:FunctionalProperty* y *owl:InverseFunctionalProperty*, respectivamente).

- Permite definir un dominio y un rango concreto para cualquier propiedad de la ontología OWL Lite (*owl:domain* y *owl:range*, respectivamente).

- Es posible restringir el rango de las propiedades mediante los cuantificadores existenciales $\forall$ y $\exists$ (constructores *owl:allValuesFrom* y *owl:someValuesFrom*, respectivamente). Ambas restricciones se declaran sobre una propiedad con respecto a una clase.

  - *owl:allValuesFrom* establece que si una instancia de una determinada clase está relacionada mediante una propiedad restringida con una segunda instancia, esta última necesariamente debe pertenecer a la clase indicada como rango de dicha propiedad (por ejemplo, si existe en una ontología una instancia de la clase *Persona* relacionada con una segunda instancia mediante la propiedad *tieneHija*, ésta necesariamente debe pertenecer a la clase *Mujeres*).

  - *owl:someValuesFrom* establece que necesariamente debe existir al menos una relación entre dos instancias mediante una propiedad restringida, de forma que la segunda de las instancias pertenezca a la clase señalada como rango en dicha propiedad (por ejemplo, la clase *ArtículosSobreWS* debería tener una propiedad restringida *palabraClave* que declare que alguno de sus valores debe pertenecer necesariamente a la clase *WebSemántica*).

- Por último, destacar que OWL Lite permite declarar restricciones de cardinalidad de valores 0 y 1.

OWL DL

OWL DL enriquece las capacidades expresivas de OWL Lite, incorporando funcionalidades como las apuntadas a continuación:

- Permite restricciones de cardinalidad no limitadas a 0 y 1.
- Es posible definir nuevas clases utilizando el constructor *owl:hasValue*, de forma que se limitan las clases (o instancias) que pueden figurar como rango en las propiedades asociadas a la clase definida (por ejemplo, la clase *Españoles* se puede definir restringiendo el rango de la propiedad *tieneNacionalidad* a la instancia *Española*, perteneciente a la clase *Nacionalidades*).
- Contempla la creación de nuevas clases a partir de la combinación *booleana* de las ya existentes en la ontología (mediante los constructores *owl:unionOf*, *owl:intersectionOf* y *owl:complementOf*).
- Es posible declarar clases disjuntas (*owl:disjointWith*).
- Permite describir nuevas clases enumerando los individuos que forman parte de la misma (por ejemplo, la clase *Estaciones* incluye únicamente los individuos *Primavera*, *Verano*, *Otoño* e *Invierno*).

Todas las primitivas de modelado mediante el lenguaje OWL que hemos descrito hasta este punto, pueden ser utilizadas en el último nivel del lenguaje, en el que se incorporan, además, nuevas funcionalidades.

OWL Full

Además de ser totalmente compatible con RDF, tanto en su sintaxis como en su semántica, OWL Full ofrece una capacidad expresiva máxima. Ésta permite, por ejemplo, que las clases puedan ser interpretadas como colecciones o como individuos, indistintamente, lo que convierte a OWL Full en un lenguaje difícil de tratar en términos computacionales.

En definitiva, el lenguaje OWL es un formalismo lógico, potente y expresivo, que no sólo permite representar aspectos básicos sobre los conceptos y las relaciones existentes entre éstos en un dominio de aplicación específico, sino que también habilita ciertos mecanismos inferenciales automáticos que lo convierten en la piedra angular de la capa ontológica de la Web Semántica.

Herramientas de soporte

Con el ánimo de asistir a los desarrolladores de ontologías, en la literatura se han propuesto diferentes herramientas que facilitan considerablemente su proceso de implementación. A continuación, enumeramos algunas de las herramientas existentes, señalando las referencias oportunas en las que el autor puede consultar una descripción detallada. En general, la mayoría de estas herramientas incorporan un editor gráfico que permite visualizar la jerarquía de clases de la ontología, así como insertar, modificar o eliminar de la misma clases, propiedades, instancias y/o reglas.

- CORESE: herramienta RDF basada en grafos conceptuales [54].

- Jena [105]: es una herramienta implementada en Java para el desarrollo de aplicaciones para la Web Semántica, cuyas capacidades permiten representar modelos RDF, RDFS y OWL, así como consultar y visualizar datos conformes a estos formatos.
- KAON: gestor de código abierto que incluye un conjunto de herramientas para crear y gestionar ontologías, y otras herramientas para construir aplicaciones basadas en las mismas [113].
- KIM: plataforma que permite la anotación semántica mediante los lenguajes RDF, RDFS y OWL Lite [115].
- Kowari: herramienta desarrollada en Java y de código abierto que soporta RDF y OWL [123].
- KPOntology: biblioteca que permite gestionar y acceder a ontologías RDF y OWL, sin necesidad de trabajar directamente con estos lenguajes [124].
- Mindswap: editor de ontologías hipermedia basado en OWL [152].
- OilEd: editor de ontologías basado inicialmente en el lenguaje OIL, y extendido posteriormente a DAML+OIL [174].
- OntoEdit: herramienta que permite representar gráficamente las ontologías y manipularlas en una base de datos relacional [176, 236].
- OntoLingua: es un entorno colaborativo distribuido, desarrollado en la Universidad de Stanford, que incorpora facilidades para la creación, edición, navegación y consulta de ontologías [178].
- Ontomat: herramienta de código abierto que soporta anotación OWL. Incluye un navegador de ontologías para poder explorar sus clases, propiedades e instancias, así como un navegador HTML que muestra las partes anotadas de los documentos [180].
- OntoShare: una herramienta para compartir conocimiento en un entorno distribuido a través de ontologías RDF(S) [60].
- OntoWeaver: permite el diseño y desarrollo de sitios Web basándose en ontologías [181].
- OntoWebber: sistema de gestión de sitios Web basado en ontologías [182].
- PROTON (PROTo ONtology): plataforma que sirve para construir ontologías, facilitando las tareas de anotación semántica, indexación y recuperación de información desde las mismas [198].
- SESAME [222, 39]: esta herramienta ofrece un repositorio basado en RDFS junto a algunas facilidades para su acceso y consulta, así como la traducción de datos a modelos conformes a este lenguaje. El acceso a la información representada en RDF está basado en el lenguaje RQL (*RDF Query Language*) [114], cuyas consultas básicas permiten extraer: (i) todas las clases de la ontología, (ii) las instancias de una clase dada, (iii) las subclases de una clase concreta, (iv) todas propiedades asociadas a una clase, (v) todas las instancias que tengan como rango de una propiedad dada un determinado valor, etc. Consultas similares a

éstas están incluidas en otros lenguajes propuestos en la literatura para el acceso a modelos RDF, como SquishQL (*Simple RDF Query Language*) [151] y RDQL (*RDF Data Query Language*) [221]. La descripción detallada de los mismos, fuera del ámbito de esta tesis, está disponible en [13], donde se revisan los diferentes paradigmas de consulta propuestos en el dominio de la Web Semántica.

- TM-Builder: es un constructor de ontologías basado en mapas temáticos (ver Sección 3.3.2) [135].
- WebOnto: consiste en un applet Java que permite navegar y editar modelos de conocimiento sobre la Web [249].
- WSMO Studio: editor de ontologías para modelado de servicios de la Web Semántica [255], desarrollado por el Web Service Modeling Ontology Group [246].

Una de las herramientas que merece especial mención, por su implantación generalizada en la comunidad de la Web Semántica, es Protégé [197]. Este editor, desarrollado en la Universidad de Stanford, ofrece una interfaz gráfica muy intuitiva que permite especificar las clases, propiedades e instancias de la ontología, así como las relaciones existentes entre ellas e, incluso, las restricciones que pueden tener asociadas. Además de su esquema de representación de conocimiento basado en marcos y del soporte de los principales lenguajes de implementación de ontologías utilizados actualmente en la Web Semántica (RDF, RDFS y OWL), una de las características más destacables de Protégé es su arquitectura abierta y modular. De hecho, con el objetivo de obtener un entorno extensible, los desarrolladores de Protégé optaron por una arquitectura basada en componentes que permiten añadir nuevas funcionalidades mediante *plugins*. Entre las funcionalidades ofrecidas por estos últimos, destacan el almacenamiento de bases de conocimiento en diversos formatos (RDFS, XML, OIL y DAML+OIL), la visualización avanzada de las mismas (*plugins* OntoViz y Jambalaya), la realización de inferencias lógicas o la fusión de varias ontologías en un único modelo. Para abordar esta última tarea, Protégé utiliza un mecanismo muy simple (basado en la declaración *import*) para poder incluir en una ontología dada otras disponibles públicamente, rescatando así la ya mencionada capacidad de reutilización de conocimiento asociada a este tipo de conceptualizaciones.

En definitiva, mediante Protégé y su filosofía de *plugins* es posible configurar un entorno que cubre prácticamente todas las fases de creación y utilización de ontologías, de ahí que sea la herramienta que goza de mayor aceptación entre los desarrolladores de este tipo de bases de conocimiento.

Por otra parte, con el objetivo de potenciar el uso de ontologías en diferentes dominios de aplicación y aprovechar así las capacidades inferenciales vinculadas a la Web Semántica, se han desarrollado herramientas específicas relacionadas con la utilización y búsqueda de ontologías en la Web, entre las que podemos destacar:

- OntoAgent [175], que indica a los usuarios las ontologías ya existentes y sus características para poder utilizarlas en su sistema,
- OntoJava [177], un compilador que traduce ontologías desarrolladas mediante el editor Protégé a bases de datos Java, y
- RDFCrawler [204] orientada a la búsqueda de datos RDF en el dominio de la Web.

Por último, mencionar OWL-S, que no es exactamente una herramienta, sino una ontología conforme al lenguaje OWL facilitada a los desarrolladores de servicios Web, para que describan a través de sus constructores las funcionalidades principales ofrecidas en los mismos.

La mayoría de las herramientas actualmente disponibles aprovechan las capacidades de los razonadores DL descritos en la Sección 3.2.4, lo que permite llevar a cabo procesos inferenciales muy útiles en el dominio de la Web Semántica, ubicados en la capa lógica de la arquitectura representada en la Fig. 3.1.

3.4.4 Capa Lógica

En esta capa se aplican diversas inferencias lógicas sobre los conceptos y relaciones formalmente representados en una ontología de conocimiento, a fin de deducir nuevos hechos a partir de los ya conocidos. Para ello, actualmente se recurre a los razonadores de la DL, capaces de llevar a cabo inferencias automáticas como las descritas a continuación:

- *Pertenencia a una clase*: Permite deducir si una instancia pertenece o no a una clase dada de la ontología.
- *Clasificación automática*: Mediante esta deducción, un razonador DL obtiene todas las relaciones jerárquicas establecidas en la ontología, a partir de las declaradas en la misma.
- *Equivalencia de clases*: Esta capacidad no sólo es útil para identificar clases equivalentes (es decir, aquellas que tienen exactamente las mismas instancias), sino que también detectan redundacias (innecesarias) durante la fase de desarrollo de la ontología.
- *Consistencia de una clase*: Los razonadores existentes son capaces de comprobar las propiedades definidas para cada clase (o instancia) de la ontología, detectando posibles inconsistencias en las mismas. Por ejemplo, si las clases *Profesor* y *Estudiante* se declaran disjuntas en una ontología, la clase *Estudiantes-Profesores*, definida como la intersección de las dos anteriores, es claramente inconsistente.
- *Consistencia de la ontología*: La capacidad de razonamiento que acabamos de describir también puede ser aprovechada para comprobar la consistencia de la ontología completa. Siguiendo con el ejemplo anterior, si la conceptualización declara que la instancia *Juan* pertenece a las clases disjuntas *Profesor* y *Estudiante*, la ontología es inconsistente, dado que las dos restricciones definidas (pertenencia de la instancia a una clase y disyunción entre clases) no se verifican a la vez en el modelo de la ontología.

Estas capacidades de razonamiento se han convertido en un elemento clave en el dominio de la Web Semántica, ya que permiten que los desarrolladores pueden diseñar sus ontologías y, a continuación, verificar la corrección y consistencia de las mismas mediante este tipo de herramientas lógicas.

Tal como hemos mencionado a lo largo de este capítulo, la capacidad de razonamiento lógica está estrechamente relacionada con la expresividad del lenguaje empleado en la implementación de la base de conocimiento. Así, por ejemplo, hoy en día

no existe ningún motor de inferencia válido para OWL Full, debido precisamente al gran poder expresivo que incorpora este lenguaje. Sin embargo, sí se han desarrollado razonadores válidos tanto para OWL Lite como para OWL DL, capaces de realizar inferencias lógicas automáticas como las apuntadas anteriormente. Junto a RACER [85] y FaCT [100], también Pellet [192] goza de gran popularidad como razonador lógico externo en aplicaciones de la Web Semántica. Todos ellos son motores de inferencia de código abierto que proporcionan facilidades para consultar modelos OWL, comprobar la consistencia de ontologías OWL DL y clasificar automáticamente sus instancias en las taxonomías apropiadas.

Tal como hemos mencionado en la Sección 3.3.4, uno de los elementos que componen las ontologías son las reglas, incorporadas a los lenguajes RDFS y OWL a fin de mejorar sus capacidades de razonamiento automático. Dichas reglas (también incluidas en la capa lógica en la Fig. 3.1) permiten definir declaraciones lógicas de la forma "*si P es cierto, entonces Q también es cierto*" y, con ello, representan conocimiento no formalizable mediante el resto de componentes incluidos en el modelo ontológico. Los lenguajes basados en reglas más extendidos en el estado del arte son:

- TRIPLE [228]: Es un lenguaje de reglas basado en cláusulas de Horn, orientado a la consulta y transformación de modelos RDF, cuya sintaxis es muy parecida a la de SQL o Prolog. Los programas TRIPLE consisten en declaraciones de hechos y reglas desde las que es posible realizar inferencias y responder a las consultas planteadas sobre la base de conocimiento RDF. Además de este tipo de mecanismos de razonamiento, TRIPLE proporciona también herramientas que permiten traducir modelos RDF a TRIPLE y viceversa.

  También en el dominio de la personalización es posible encontrar enfoques que utilizan el lenguaje TRIPLE. Este es el caso de SIMPLE [67], un sistema educacional (basado en hipermedia) que se adapta a las necesidades y al estado de aprendizaje de cada usuario mediante un proceso de razonamiento lógico. En esta propuesta, el conocimiento adquirido por el usuario se utiliza para modelar su perfil personal. A partir de este último, el sistema selecciona aquellos contenidos que son relevantes para su proceso de aprendizaje utilizando un conjunto de reglas lógicas de primer orden.

- RuleML (*Rule Markup Language*): Este lenguaje nació con el claro objetivo de permitir la reutilización y compartición de conocimiento en el dominio de la Web, empleando para ello reglas basadas en sintaxis XML. RuleML combina la programación lógica con la orientada a objetos, y construye una jerarquía de diferentes sublenguajes de reglas basados en XML, RDF, XSLT y OWL [213].

  Entre los sistemas basados en reglas de inferencia RuleML, destacamos la herramienta RACOFI en el dominio del *e-learning* [134]. Este sistema combina el filtrado colaborativo (ver Sección 2.6.3) con un conjunto de reglas RuleML, a fin de presentar a los usuarios objetos de aprendizaje personalizados de acuerdo a sus preferencias. Así, para predecir el nivel de interés del usuario en un objeto dado, RACOFI utiliza, en primer lugar, las clasificaciones que el resto de usuarios del sistema han asignado a dicho objeto y, en segundo lugar, aplica las reglas RuleML para mejorar este proceso de predicción y personalizar así los objetos seleccionados.

- SWRL (*Semantic Web Rule Language*): Esta propuesta integra los lenguajes de reglas con las tecnologías de la Web Semántica. En concreto, SWRL es una

combinación de los lenguajes semánticos OWL Lite y OWL DL, y un subconjunto de RuleML [238]. La ventaja de SWRL, además del nivel de estandarizacion y abstracción que proporciona, es que actualmente dispone de numerosas herramientas de apoyo (como SWRL-Protégé Plugin [195]), que permiten aprovechar la funcionalidad de las reglas en tareas de recuperación de información.

3.4.5 Capas de Comprobación y Confianza

Estas capas son, sin duda, las menos maduras de la arquitectura propuesta para la Web Semántica, de ahí la necesidad de especificar y desarrollar sus principios teóricos y prácticos. En realidad, el objetivo en este nivel es ejecutar las reglas definidas en la capa lógica y, utilizando los mecanismos proporcionados por la capa superior de la torre, comprobar si las aplicaciones pueden confiar o no en el conocimiento deducido mediante las mismas.

Si bien los lenguajes de reglas existentes permiten mejorar las capacidades de razonamiento incorporadas en la capa lógica de la Web Semántica, los motores de inferencia propuestos en el estado del arte no consideran el descubrimiento de relaciones semánticas de naturaleza *compleja*, detectadas a partir de las propiedades explícitamente declaradas entre las instancias de una ontología.

De este tipo de propuestas nos ocupamos en las siguientes secciones, en las que analizamos dos enfoques propuestos en la literatura para la exploración de conocimiento en el dominio de la Web Semántica: las *asociaciones semánticas complejas* y las *técnicas de propagación de activación*.

3.5 Asociaciones semánticas complejas

La aparición de la Web Semántica ha provocado el desarrollo de diversos enfoques que incorporan capacidades de consulta semántica, tal como se ha indicado en la sección previa. A pesar de su expresividad, los lenguajes existentes no permiten descubrir relaciones semánticas complejas, ocultas en una base de conocimiento. En su lugar, estos lenguajes ofrecen paradigmas de consulta de la forma "*Extrae todas los recursos que estén relacionadas con el recurso A mediante la relación R*". Este tipo de propuestas exigen a los usuarios tener un conocimiento detallado de la estructura de la base de conocimiento, para, de esta forma, poder identificar entidades y relaciones específicas a la hora de formular sus consultas. Es obvio que, en muchos dominios de aplicación, dichas exigencias resultan excesivas y, por ello, limitan notablemente la utilidad de este tipo de enfoques.

Conscientes de tal limitación, los integrantes del proyecto de investigación SemDis[2] (*Semantic Discovery*) se centraron en el diseño, implementación y evaluación de un sistema homónimo que ofrece capacidades para consultar y descubrir relaciones semánticas ocultas en una base de conocimiento RDFS, denominadas **asociaciones semánticas complejas** [5].

El paradigma de consulta manejado en su trabajo es de la forma "*Extrae todas las posibles relaciones existentes entre los recursos A y B*". En este caso, el conocimiento exigido al usuario sobre la conceptualización se ha relajado notablemente (se ha eliminado la componente de las relaciones), incrementando la flexibilidad y utilidad de

[2]Información adicional disponible en *http://lsdis.cs.uga.edu/projects/semdis/*

la propuesta. Si bien este enfoque es genérico, uno de los dominios en los que ha demostrado ser especialmente útil es el de las aplicaciones relacionadas con la seguridad nacional. Por ejemplo, en el contexto de la seguridad aérea, el hecho de que dos pasajeros de un mismo vuelo sean ciudadanos de un país vinculado a diversos grupos terroristas de fama internacional, y hayan recibido recientemente entrenamiento aéreo, podría ser indicativo de una asociación entre ellos. El descubrimiento de este tipo de relaciones, de innegable valor desde el punto de vista de la seguridad, son las que avalan la utilidad del prototipo SemDis.

Además de definir las asociaciones semánticas, los investigadores de SemDis proponen dos mecanismos para poder clasificarlas en función de lo relevantes que sean para la consulta formulada por el usuario. Empecemos revisando algunos conceptos de interés, antes de presentar las asociaciones semánticas y revisar este tipo de procesos de clasificación.

3.5.1 Secuencias de propiedades

Tal como se describe en [5], para descubrir las asociaciones semánticas complejas es necesario analizar las propiedades formalizadas en la ontología RDFS considerada. Mediante la exploración de tales propiedades, es posible identificar dos estructuras genéricas:

- Secuencia de propiedades (PS): Consta de un conjunto de clases unidas entre sí mediante propiedades, de forma que el rango de una de ellas coincide con el dominio de la siguiente.
- Instancia de una secuencia de propiedades (ps): Como su nombre indica, consiste en un conjunto de instancias de clases unidas entre sí mediante propiedades, de la misma forma descrita antes[3].

En ambos casos, la longitud de la (instancia de la) secuencia coincide con el número de propiedades contenidas en la misma. Además, en [5] también se definen dos funciones para poder acceder a los nodos incluidos en las secuencias y en sus instancias; éstas son $PS.NodesOfPS()$ y $ps.PSNodesSequence()$, respectivamente. Además, la primera instancia entregada por $ps.PSNodesSequence()$ recibe el nombre de *origen*, mientras que la última se denomina *término*. Pongamos un sencillo ejemplo para entenderlo mejor.

Siendo $PS = [P_1, P_2, \ldots, P_n]$ una secuencia de propiedades, se tiene que la función $PS.NodesOfPS()$ devuelve el conjunto $\{C_1, C_2, \ldots, C_{n+1}\}$, donde C_i son las clases que figuran como rangos y dominios de las propiedades P_j mencionadas antes. De forma análoga, la función $ps.PSNodesSequence()$ devuelve el conjunto $\{v_1, v_2, \ldots, v_{n+1}\}$, donde v_i identifica cada una de las instancias de las clases C_i, siendo v_1 y v_{n+1} el origen y el término de ps, respectivamente.

3.5.2 Relaciones entre secuencias de propiedades

En [5], también se identifican las posibles relaciones que se pueden establecer entre dos secuencias de propiedades. Para su definición, consideremos dos secuencias de lon-

[3] La notación utilizada en [5] establece que las instancias de las secuencias se representan siempre en letra minúscula, y las secuencias en mayúscula.

gitudes n y m, respectivamente: $PS_1 = [P_1, P_2, \ldots, P_n]$ y $PS_2 = [Q_1, Q_2, \ldots, Q_m]$.

- **Secuencias de propiedades unidas** (*Joined Property Sequences*): Se dice que PS_1 y PS_2 son secuencias unidas (representado como $PS_1 \bowtie_\rho PS_2$) si tienen algún nodo en común (por esta razón, llamado *nodo de unión*). Formalmente, podemos expresarlo así:

$$PS_1 \bowtie_\rho PS_2 \Leftrightarrow PS_1.NodesOfPS() \cap PS_2.NodesOfPS() \neq \oslash$$

- **Secuencias de propiedades ρ-isomórficas** (*ρ-Isomorphic Property Sequences*): Se dice que PS_1 y PS_2 con $n = m$ (es decir, de la misma longitud) son secuencias ρ-isomórficas (representado como $PS_1 \cong_\rho PS_2$) si sus respectivas propiedades están relacionadas, una a una, mediante igualdad o *inclusión* (representado mediante el símbolo $\sqsubseteq$). Formalmente, se expresa como sigue:

$$PS_1 \cong_\rho PS_2 \Leftrightarrow \forall i, 1 \leq i \leq n : P_i = Q_i \ \text{ o } \ P_i \sqsubseteq Q_i \ \text{ o } \ Q_i \sqsubseteq P_i$$

3.5.3 Tipos de asociaciones semánticas complejas

Precisamente tomando como punto de partida las relaciones existentes entre las secuencias de propiedades, en SemDis identifican hasta cuatro tipos diferentes de asociaciones semánticas.

Sean ps_1 y ps_2 dos instancias de las secuencias de propiedades PS_1 y PS_2, y x e y dos instancias definidas en la ontología RDFS.

- **Asociación *ρ-path***: Se dice que existe una asociación *ρ-path* entre x e y (representado como *ρ-path* (x,y)) si es posible encontrar una instancia ps_1, cuyo origen es x y cuyo término es y, o viceversa.

- **Asociación *ρ-join***: Se dice que existe una asociación *ρ-join* entre x e y (representado como *ρ-join* (x,y)) si es posibe encontrar dos secuencias PS_1 y PS_2 unidas (es decir, $PS_1 \bowtie_\rho PS_2$) y dos instancias de las mismas ps_1 y ps_2 en las que se cumple:

  - x es el origen de ps_1 e y es el origen de ps_2, o bien
  - x es el término de ps_1 e y es el término de ps_2.

- **Asociación *ρ-cp***: *ρ-cp* (x,y) se cumple si existen dos secuencias de propiedades unidas (es decir, $PS_1 \bowtie_\rho PS_2$) y dos instancias de las mismas ps_1 y ps_2 para las que se satisface la asociación *ρ-join* (x,y). Ademas, las propiedades de PS_1 y PS_2 deben ser de la forma *rdf.typeOf.(rdfs:subClassOf)**. Es decir, ambas secuencias tienen una primera propiedad (*typeOf*) que permite identificar la clase a la que pertenece x (ó y para la otra secuencia), y el resto de propiedades de la secuencia (de tipo *rdfs:subClassOf*) identifican relaciones de herencia entre dicha clase y otras definidas en la jerarquía de la ontología, hasta alcanzar el nodo de unión en el que terminan ambas secuencias y que provoca la asociación de tipo *ρ-join* mencionada antes.

  Dicho de otra forma, para poder definir la asociación *ρ-cp* (x,y), ambas instancias deben pertenecer a clases que tengan un padre común (*common parent*) que actúa como nodo de unión.

A fin de evitar la inferencia de asociaciones semánticas poco significativas, [5] establece algunas restricciones:

- En primer lugar, es posible especificar la clase más general de la ontología RDFS que puede actuar como nodo de unión, denominada por esta razón *techo* de la asociación *ρ-cp*. En el contexto de SemDis, la especificación del techo es una tarea encomendada al usuario del sistema.
- Por otra parte, si consideramos que las secuencias unidas PS_1 y PS_2 originan una nueva secuencia, cuya longitud se corresponde con la suma de las longitudes individuales de cada una de ellas, el enfoque de SemDis también permite poner una cota superior a este parámetro, asegurando con ello que las instancias x e y no estén demasiado distantes en la ontología. Esto último es importante, ya que una longitud elevada se traduce en el descubrimiento de asociaciones poco relevantes para la consulta del usuario.

- **Asociación *ρ-iso***: *ρ-iso* (x, y) se cumple si existen dos secuencias isomórficas (esto es, $PS_1 \cong_\rho PS_2$) y dos instancias de las mismas ps_1 y ps_2, tales que x es el origen de ps_1 e y es el origen de ps_2.

3.6 Valoración de la relevancia de las asociaciones semánticas

Una vez identificadas las asociaciones semánticas en la base de conocimiento RDFS, los integrantes de SemDis vieron la necesidad de recurrir a algún mecanismo que permitiese valorar la relevancia de las mismas con respecto a la consulta formulada al sistema. En dichas consultas, el usuario especifica dos entidades RDFS y el enfoque propuesto recupera únicamente aquellas relaciones establecidas entre ambas que realmente aportan conocimiento útil en el dominio de aplicación concreto del sistema.

La propuesta de SemDis [87] consiste en calcular una función de clasificación para cada una de las asociaciones inferidas. El valor de dicha función, que será más significativo para las asociaciones más relevantes, depende de varios criterios intermedios, convenientemente ponderados en la expresión final. Estos criterios se organizan en torno a dos grupos genéricos de métricas: métricas semánticas y métricas estadísticas.

3.6.1 Métricas semánticas

Las métricas semánticas definidas en esta aproximación están estrechamente relacionadas con la ontología RDFS utilizada. En concreto, consideran aspectos relativos a sus instancias y a las asociaciones establecidas entre éstas. En este grupo, son tres las métricas propuestas, denominadas de forma genérica *contexto*, *inclusión* y *confianza*.

Contexto

Antes de formular una consulta en el sistema SemDis, es preciso que el usuario identifique el contexto de la misma. Para ello, basta con identificar una o varias regiones (subgrafos) dentro de la ontología, correspondientes a sus dominios de interés preferentes. Además, para cada una de dichas regiones, es necesario indicar un peso,

a modo de nivel de interés, de forma que los índices de todas las regiones definidas deben sumar uno. Así, el usuario puede especificar las clases y/o propiedades que son más relevantes para él, de forma que las asociaciones que involucren estas entidades serán más significativas que aquéllas en las que no figuren dichos elementos.

Veamos cómo se define formalmente esta métrica. Sean:

- A, una asociación semántica en la que dos nodos (clases y/o instancias) están conectados entre sí mediante un conjunto de propiedades;
- R_i, una región formada por las clases y propiedades que capturan el dominio de interés relativo a una determinada consulta;
- c, un componente de A (i.e. una clase, instancia o propiedad incluida en tal asociación), y
- n, el número de regiones de interés declaradas por el usuario.

Además, para formalizar la métrica, los autores de [87] definen sendos conjuntos:

$$X_i = \{c \mid (c \in R_i) \wedge (c \in A)\}$$
$$Z = \{c \mid (\forall i \mid 1 \leq i \leq n)\ (c \notin R_i) \wedge (c \in A)\}$$

donde X_i identifica el conjunto de componentes de A que también están definidos en la región de interés R_i, y Z contiene todos los componentes de A que no pertenecen a ninguna de las regiones especificadas por el usuario.

A partir de ambos conjuntos, [87] define la métrica *contexto* (C_A) como sigue:

$$C_A = \frac{1}{length(A)} \cdot \sum_{i=1}^{n} (w_{R_i} \cdot |X_i|) \cdot (1 - \frac{|Z|}{length(A)}) \tag{3.1}$$

donde w_{R_i} es el índice de interés declarado por el usuario en relación a la región R_i y $length(A)$ es el número de componentes incluidos en la asociación A.

A la vista de la Ec. 3.1, el valor de la métrica *contexto* C_A será más significativo cuanto mayor sea el número de componentes de la asociación A que también pertenecen a las regiones de mayor interés definidas por el usuario.

Inclusión

Esta métrica considera las relaciones de herencia (jerárquicas) establecidas en la ontología RDFS utilizada en SemDis. El valor que toma esta métrica para una asociación específica depende exclusivamente de la localización de sus clases en la jerarquía considerada. La contribución de cada una de estas clases (representada como csw_i, de *component subsumption weight*) se calcula mediante la Ec. 3.2.

$$csw_i = \frac{H_{c_i}}{H_{depth}} \tag{3.2}$$

donde H_{c_i} es la posición que ocupa la componente c_i de la asociación A en la jerarquía H, y H_{depth} es la profundidad de esta taxonomía, es decir, el número de relaciones

jerárquicas existentes entre las clases que ocupan el nivel más específico de la jerarquía y el nodo raíz de la misma.

Una vez aplicada la Ec. 3.2 sobre cada componente de A, se reúnen todas las contribuciones para computar la métrica *inclusión* de dicha asociación, representada como S_A en la Ec. 3.3.

$$S_A = \prod_{i=1}^{length(A)} csw_i \tag{3.3}$$

Esta expresión conduce a valores más significativos para aquellas asociaciones semánticas cuyas componentes pertenezcan a los niveles más alejados de la raíz de la jerarquía, dado que son éstos los más específicos y, por tanto, los que aportan más conocimiento al usuario.

Confianza

Las clases e instancias incluidas en la base de conocimiento RDFS usada en SemDis, han sido extraídas de distintas fuentes (importadas desde otras ontologías, declaradas por un experto en el dominio de aplicación formalizado en las mismas, etc.). Algunas de estas fuentes son más fiables que otras, de ahí la necesidad de incluir una métrica que permita identificar aquellas asociaciones que contienen componentes de confianza.

De esta forma, a cada componente le corresponde un valor concreto de confianza, que debería ser asignado a los metadatos formalizados en la ontología durante la fase de descripción o anotación semántica. Este valor, obviamente, dependerá de la fiabilidad que proporcione cada una de las fuentes de las que se extraen dichos metadatos, de ahí la necesidad de recurrir a expertos en la materia.

El enfoque adoptado en SemDis asume que la *confianza* de una asociación semántica A (representada como T_A) coincide con el valor correspondiente a su componente menos fiable (t_{c_i}), tal como se muestra en la Ec. 3.4.

$$T_A = min(t_{c_i}) \tag{3.4}$$

3.6.2 Métricas estadísticas

Este tipo de métricas consideran aspectos estadísticos de la ontología RDFS empleada en el enfoque, basados fundamentalmente en el número de instancias pertenecientes a las clases definidas en la misma. Son tres las métricas clasificadas en este grupo: *rareza*, *popularidad* y *longitud de la asociación*.

Rareza

Dado el gran volumen de información manejada en la Web Semántica, es obvio que son muchas las entidades (clases e instancias) y relaciones (propiedades) que se pueden formalizar en una base de conocimiento concreta.

Desde el punto de vista de las consultas de asociaciones semánticas, es posible que aquellas relaciones y entidades que ocurran con poca frecuencia en la ontología sean

más relevantes o interesantes para el usuario, por la dificultad que puede entrañar su descubrimiento. Sin embargo, en ciertos dominios de aplicación, son precisamente las relaciones más frecuentes las más significativas (por ejemplo, en el contexto de las actividades delictivas).

Para considerar ambas aproximaciones, el enfoque introduce la métrica *rareza*, cuyo valor se calcula a partir de los pesos correspondientes a cada una de las componentes c_i incluidas en una asociación semántica A. Para obtener dichos pesos (representados como rar_i) en una base de conocimiento K, es necesario aplicar la Ec. 3.5.

$$\mathbf{rar_i} = \frac{|\mathbf{M}| - |\mathbf{N}|}{|\mathbf{M}|} \tag{3.5}$$

En esta expresión, M identifica el conjunto de todas las instancias y relaciones definidas en K (Ec. 3.6), y N representa el subconjunto de estas últimas que, además, son del mismo tipo[4] que la componente c_i (Ec. 3.7).

$$M = \{res \mid res \in K\} \tag{3.6}$$

$$N = \{res_j \mid (res_j \in K) \wedge (type(res_j) = type(c_i))\} \tag{3.7}$$

A la vista de las expresiones que acabamos de introducir, el peso rar_i representa la frecuencia de ocurrencia de la componente c_i en la base de conocimiento K. El valor de la métrica *rareza* correspondiente a la asociación A se calcula a partir de las contribuciones de cada uno de sus componentes. Cabe destacar, además, que el enfoque propone dos expresiones para esta métrica, en función del tipo de asociaciones que resulten más interesantes para el usuario: aquellas que involucran componentes poco frecuentes en K (Ec. 3.8), o aquellas otras cuyas relaciones y entidades son comunes en la misma (Ec. 3.9).

$$R_A = \frac{1}{length(A)} \cdot \sum_{i=1}^{length(A)} rar_i \tag{3.8}$$

$$R_A = 1 - \frac{1}{length(A)} \cdot \sum_{i=1}^{length(A)} rar_i \tag{3.9}$$

Popularidad

En una base de conocimiento algunas entidades tienen más propiedades asociadas que otras. Para reflejar este desequilibrio, el enfoque de SemDis introduce la noción de *popularidad* de una entidad, definida precisamente como el número de propiedades en las que dicha entidad figura como rango o como dominio.

En algunos dominios de aplicación, las asociaciones más relevantes podrían ser aquellas que involucran entidades muy populares en la base de conocimiento (por ejemplo, si el usuario formula una consulta en el dominio de la investigación, aquellas publicaciones que han sido objeto de abundantes citas serían las entidades preferentes). Sin embargo, de forma análoga a la comentada en la métrica *rareza*, también es posible que las asociaciones más significativas sean las que contienen entidades con pocas relaciones en la conceptualización.

[4]Recordar que los componentes de una asociación A pueden ser instancias, clases o propiedades.

En [87], la *popularidad* de una asociación A se calcula a partir de las contribuciones que aportan cada uno de sus componentes. En concreto, para calcular el peso p_i de la componente c_i, este enfoque considera la entidad más popular de todas aquellas que pertenecen al mismo tipo que c_i, tal como muestra la Ec. 3.10.

$$p_i = \frac{|pop_{e_i}|}{max(|pop_{e_j}|)} \quad \text{con} \quad typeOf(e_i) = typeOf(e_j), 1 \leq j \leq n \tag{3.10}$$

donde n es el número total de entidades formalizadas en la base de conocimiento K, pop_{e_i} es la popularidad de la entidad e_i, y e_j es la entidad más popular de K que pertenece al mismo tipo que e_i.

De la misma forma que en la métrica *rareza*, este enfoque permite que el usuario elija el tipo de asociaciones que le interesan, en función de la popularidad de sus entidades; de ahí que existan dos expresiones matemáticas diferentes para formalizar la *popularidad*: la Ec. 3.11 si se desean premiar las asociaciones populares y la Ec. 3.12, en caso contrario.

$$P_A = \frac{1}{n} \cdot \sum_{i=1}^{n} p_1 \tag{3.11}$$

$$P_A = 1 - \frac{1}{n} \cdot \sum_{i=1}^{n} p_1 \tag{3.12}$$

Longitud de la asociación

La longitud de una asociación es una parámetro clave a la hora de determinar el tipo de relación establecida entre sus entidades. Éstas serán más directas si las asociaciones son cortas, y más indirectas en el caso de que dichas asociaciones involucren un número elevado de entidades. Estas últimas son especialmente útiles en dominios en los que hay una gran cantidad de conocimiento útil oculto en la conceptualización (por ejemplo, actividades delictivas).

Para poder reflejar ambos tipos de relaciones en su enfoque, el enfoque propuesto en [87] introduce la métrica *longitud* (representada como L_A para la asociación A), formalizada mediante dos expresiones diferentes: la Ec. 3.13 si el usuario está interesado en asociaciones directas, y la Ec. 3.14 para aquellas más indirectas.

$$L_A = \frac{1}{length(A)} \tag{3.13}$$

$$L_A = 1 - \frac{1}{length(A)} \tag{3.14}$$

3.6.3 Criterio de clasificación global

Una vez cuantificadas las métricas semánticas y estadísticas en relación a la asociación A, el enfoque reúne todas estas contribuciones bajo un criterio de clasificación global, representado como W_A en la Ec. 3.15.

$$W_A = k_1 \cdot C_A + k_2 \cdot S_A + k_3 \cdot T_A + k_4 \cdot R_A + k_5 \cdot P_A + k_6 \cdot L_A \tag{3.15}$$

Obviamente, las asociaciones más relevantes para cada consulta son aquellas para las que la Ec. 3.15 devuelve valores más elevados. Los pesos k_j permiten que el usuario ajuste el tipo de consulta que desea formular al sistema, tal como hemos descrito para cada una de las métricas introducidas en las secciones previas.

Si bien esta aproximación es suficientemente flexible, su implantación tropieza precisamente con la exigencia de que tenga que ser el usuario quien establezca todos los parámetros de la consulta. Conscientes de tal limitación, los integrantes del proyecto SemDis propusieron un enfoque alternativo (acuñado bajo el nombre *SemRank*), que permite al usuario formular consultas interaccionando mínimamente con el sistema, y obtener aquellas asociaciones semánticas que son más relevantes en su contexto.

3.7 SemRank: Clasificación de asociaciones semánticas

El objetivo fundamental de este enfoque es presentar al usuario una lista de asociaciones semánticas, establecidas entre las dos entidades que éste especifica, ordenadas según lo relevantes que sean para la consulta formulada. Para llevar a cabo tal proceso de clasificación, la aproximación descrita en [6] asigna un valor numérico (llamado *SemRank*, de *Semantic Ranking*) a cada asociación semántica, en función del cual es posible establecer el orden que ocupa cada una de ellas en la lista final mostrada al usuario. Obviamente, las primeras posiciones de la misma corresponderán a aquellas asociaciones semánticas cuyo índice *SemRank* sea más significativo.

La funcionalidad descrita en este enfoque es similar a la ofrecida por los buscadores definidos en el dominio de la Web: el usuario especifica unos términos de consulta y la herramienta debe recuperar y ordenar la información solicitada. Por ello, podemos interpretar la herramienta descrita en [6] como un buscador de relaciones semánticas, cuyos términos de consulta son entidades específicas formalizadas en una ontología, en lugar de palabras clave en formato textual. Esto último supone una diferencia fundamental entre ambos tipos de herramientas en lo que a la utilización de técnicas de recuperación y clasificación de información se refiere. Los buscadores convencionales establecen comparaciones sintácticas entre los términos de consulta y la información considerada durante el proceso de búsqueda y, en función del grado de coincidencia existente, determinan la posición de los resultados devueltos. Por el contrario, los buscadores de asociaciones parten de dos entidades formalizadas en una ontología, de ahí que no tenga sentido emplear métodos sintácticos que ignoren la semántica representada en la base de conocimiento. Como consecuencia de ello, no es posible reutilizar los mecanismos de ordenación empleados en el dominio de la Web [37, 51, 101, 116, 185, 90] para poder clasificar las asociaciones semánticas complejas, de ahí la necesidad de explorar enfoques alternativos en esta dirección.

Una de las primeras aproximaciones fue desarrollada en el campo de las redes semánticas [258]. Este enfoque propone un esquema para ordenar las relaciones establecidas en una red de este tipo. Su principal limitación es que dicho esquema impone un único tipo de ordenación en el conjunto de la información recuperada. En su lugar, parece más adecuado permitir cierta flexibilidad a la hora de decidir la relevancia de la información, de forma que un mismo conjunto de datos pueda ser ordenado según diferentes criterios, dependiendo de las necesidades de búsqueda de cada usuario.

Con esta idea en mente, los autores de *SemRank* introducen en su herramienta la posibilidad de formular dos clases de consultas: *consultas convencionales* y *consultas basadas en descubrimiento de conocimiento*. Para resolver ambos tipos, el enfoque recurre a un criterio basado en lo predecibles que son los resultados recuperados en cada consulta. De esta forma, en las consultas convencionales, el enfoque clasifica en primer lugar asociaciones directas, fáciles de predecir por parte de los usuarios. Por el contrario, en las consultas basadas en descubrimiento de conocimiento, son las asociaciones más indirectas y, por ello difícilmente descubiertas por el usuario, las que figuran en los primeros lugares de la clasificación.

Para poder calcular el valor *SemRank* de una asociación concreta, el enfoque propuesto considera las contribuciones de las secuencias de propiedades a partir de las cuales ésta fue descubierta. De forma análoga, la contribución de cada una de dichas secuencias se obtiene a partir de las propiedades contenidas en la misma. Para computar dichos pesos, la aproximación define cuatro métricas, cuyos valores dependen del tipo de consulta seleccionada por el usuario en la herramienta de clasificación.

3.7.1 Métricas empleadas

Las métricas empleadas en [6], denominadas de forma genérica *especificidad*, *entropía*, *refracción* y *S-Match*, intentan reflejar lo fácil o difícil que puede ser para el usuario predecir cada una de las asociaciones semánticas consideradas, para, de esta forma, poder diferenciar los dos tipos de consultas soportados en el sistema *SemRank*.

Especificidad

En [6], una propiedad será más específica cuantas menos instancias de la misma se hayan definido en la base de conocimiento. En concreto, la especificidad de una propiedad P se define como la probabilidad de encontrar alguna de sus instancias en el conjunto de todas las propiedades incluidas en la ontología, tal como indica la Ec. 3.16.

$$SP(P) = \frac{|Inst(P)|}{|Inst(Props)|} \tag{3.16}$$

donde $Inst(P)$ representa el conjunto de todas las instancias de P, e $Inst(Props)$ contiene todas las instancias de propiedades de la conceptualización.

Partiendo de la expresión mostrada en la Ec. 3.16, el enfoque *SemRank* cuantifica la especificidad de una secuencia de propiedades completa. Así, siendo $PS = [P_1, P_2, \ldots, P_n]$ una secuencia de propiedades y ps una instancia de la misma, esta aproximación estima la especificidad de PS como el valor correspondiente a la propiedad menos específica contenida en ella, tal como muestra la Ec. 3.17.

$$SP(ps) = min\{SP(P_i)\}, \forall i \tag{3.17}$$

En base a las posibles relaciones establecidas entre dos secuencias de propiedades concretas (ver Sección 3.5.2), el enfoque aproxima el valor de especificidad correspondiente a las secuencias fruto de dichas relaciones. Éstas contendrían simplemente todas las clases e instancias incluidas en las dos secuencias relacionadas. Así, por ejemplo, si PS_1 y PS_2 son secuencias unidas ($PS_1 \bowtie_\rho PS_2$) o isomórficas ($PS_1 \cong_\rho PS_2$), y ps_1

y ps_2, sus respectivas instancias, el valor de especificidad $SP(ps_1, ps_2)$ se aproxima empleando un criterio análogo al utilizado en la Ec. 3.17.

$$SP(ps_1, ps_2) = min\{SP(ps_1), SP(ps_2)\} \tag{3.18}$$

Entropía

El concepto de entropía fue propuesto por Claude Shannon en 1948 [223]. En el campo de la Teoría de la Información, la entropía es interpretada como la información que aporta la ocurrencia de un determinado evento, esto es, la incertidumbre que es posible eliminar al conocer que dicho evento ha ocurrido.

Así, siendo $\mathcal{X}$ una variable aleatoria discreta (un evento) cuyos posibles valores son $x_1, x_2, \ldots, x_n$, ocurriendo con probabilidades $pr_1, pr_2, \ldots, pr_n$, Shannon aproxima la entropía de $\mathcal{X}$ como:

$$H(\mathcal{X}) = -\sum_{i=1}^{n} pr_i \cdot log(pr_i) \tag{3.19}$$

De acuerdo a la expresión 3.19, la entropía de un evento será máxima cuando todos sus posibles valores ocurran con la misma probabilidad. En este caso, la Ec. 3.19 conduce a un valor de entropía $H(\mathcal{X}) = log(n)$.

En el enfoque *SemRank*, sus autores aplican la noción de entropía para medir la cantidad de información que aportan las asociaciones semánticas consideradas durante el proceso de clasificación. Para ello, en primer lugar se propone una medida de la entropía de cada propiedad que conecta las dos entidades especificadas por el usuario en su consulta y, a continuación, dicha medida es aplicada en el cálculo de la entropía de una secuencia de propiedades completa.

Siendo θ el conjunto de todas las propiedades establecidas en la base de conocimiento entre dos entidades a y b, y p una propiedad incluida en dicho conjunto, [6] define la entropía de P (es decir, $H(P)$) como sigue:

$$H(P) = -\sum_{P \in \theta} Pr(P \in \theta) \cdot log[Pr(P \in \theta)] \tag{3.20}$$

Una vez cuantificada la entropía de una propiedad, el enfoque extiende este modelo con funcionalidades adicionales que permitan calcular la entropía de una secuencia de propiedades. Para este propósito, [6] recurre a la noción de entropía conjunta de Shannon [223], formalizada a continuación.

Sean $\mathcal{X}$ e $\mathcal{Y}$ dos variables aleatorias cuya función de distribución de probabilidades conjunta es Pr ($\mathcal{X}$, $\mathcal{Y}$). La entropía conjunta $H(\mathcal{X}, \mathcal{Y})$ es:

$$H(\mathcal{X}, \mathcal{Y}) = -\sum Pr(\mathcal{X}, \mathcal{Y}) \cdot log[Pr(\mathcal{X}, \mathcal{Y})] \tag{3.21}$$

donde $Pr(\mathcal{X}, \mathcal{Y}) = Pr(\mathcal{X}) \cdot Pr(\mathcal{Y}|\mathcal{X})$. Además, si ambas variables son independientes, la Ec. 3.21 conduce a un valor de entropía conjunta $H(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y})$.

Dado que las secuencias son conjuntos de propiedades, su entropía puede aproximarse a partir de la entropía conjunta de cada una de éstas. En concreto, asumiendo la secuencia de propiedades $PS = [P_1, P_2, \ldots, P_n]$ y su instancia ps, se cumple que:

$$H(ps) = -\sum Pr(P_1, P_2, \ldots, P_n) \cdot log[Pr(P_1, P_2, \ldots, P_n)] \tag{3.22}$$

Las probabilidades conjuntas incluidas en la Ec. 3.22 se definen como sigue:

$$Pr(P_1, \ldots, P_n) = \prod_{j=1}^{n-1} Pr(P_j | \bigcup_{k=1}^{j-1} P_k) \tag{3.23}$$

donde, siendo $Inst(Props)$ el conjunto de todas las instancias de propiedades incluidas en la base de conocimiento, e $Inst(P_1, \ldots, P_{n-1}, P_n)$ el conjunto de instancias de las propiedades $P_1, \ldots, P_n$, se cumple que:

$$Pr(P_n | P_1, P_2, \ldots, P_{n-1}) = \frac{Inst(P_1, \ldots, P_{n-1}, P_n)}{Inst(Props)} \tag{3.24}$$

A partir de la expresión de la entropía de una secuencia de propiedades, *SemRank* computa también el valor correspondiente a secuencias unidas o isomórficas, de forma análoga a la descrita para la *especificidad*. En este caso, considerando de nuevo dos secuencias unidas o isomórficas PS_1 y PS_2 y dos instancias de las mismas ps_1 y ps_2, se estima que $H(ps_1, ps_2)$ coincide con el valor correspondiente a la secuencia de mayor entropía, tal como muestra la Ec. 3.25.

$$H(ps_1, ps_2) = max\{H(ps_1), H(ps_2)\} \tag{3.25}$$

Refracción

Esta métrica está estrechamente relacionada con una de las capacidades ofrecidas por los lenguajes de la Web Semántica como RDF, RDFS y OWL. Se trata de la posibilidad de que una misma instancia pertenezca a varias clases de la conceptualización, lo que en la literatura se conoce como clasificación múltiple. En el dominio de las asociaciones semánticas, dicha capacidad provoca que ciertos caminos definidos a través de propiedades entre las instancias de la base de conocimiento, no se correspondan con ninguna secuencia de propiedades establecida entre las clases de la ontología. Aclaremos este concepto con ayuda de la Fig. 3.2, en la que C_i y c_j representan clases e instancias, respectivamente.

En esta figura, es posible identificar dos secuencias de propiedades $PS_1 = [P_1, P_2]$ y $PS_2 = [P_3, P_4]$, y una instancia $ps = [p_1, p_2, p_3, p_4]$, para la cual no existe ninguna secuencia de la forma $PS = [P_1, P_2, P_3, P_4]$. En este escenario, los autores de *SemRank* establecen que la instancia c_3 se refracta: inicialmente, ps sigue fielmente el modelo establecido en la secuencia PS_1 (de ahí las dos primeras propiedades p_1 y p_2 incluidas en la misma); sin embargo, debido a que c_3 también pertenece a la clase C_4, es posible que ps se desvíe de PS_1 y añade las propiedades de PS_2.

La refracción es una métrica adecuada para poder valorar lo fácil o difícil que sería para un usuario predecir una determinada asociación semántica, de forma que aquellas asociaciones que contengan instancias que se refractan serían menos predecibles. Es precisamente por esta razón por la que este concepto ha sido incluido en el enfoque *SemRank*. Para cuantificar dicha métrica, este trabajo asigna un índice RI a cada secuencia de propiedades, definido como la proporción de refracciones que ocurren en

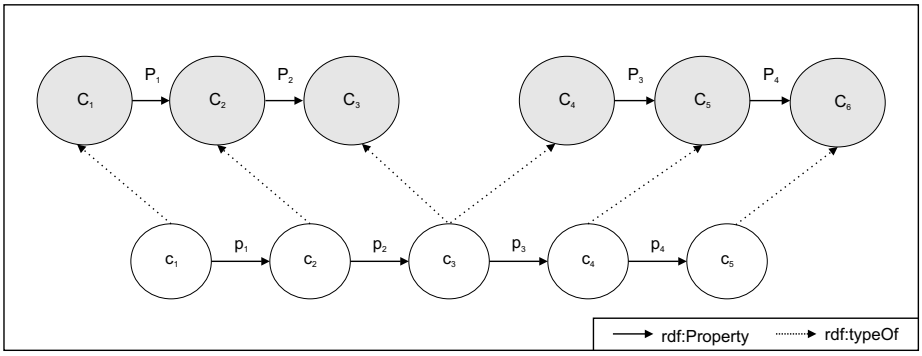

Figura 3.2: El concepto de la refracción

dicha secuencia, en relación al número de refracciones posibles que podrían producirse en la misma, tal como indica la Ec. 3.26.

$$RI(ps) = \begin{cases} \frac{1}{n-1} \cdot \sum_{i=1}^{n-1} Refrac(P_i, P_{i+1}) & \text{si } n \geq 2 \\ 0 & \text{en otro caso} \end{cases} \tag{3.26}$$

donde n es la longitud de la secuencia ps y $Refrac(P_i, P_{i+1}) = 1$, si hay refracción y 0, en caso contrario.

También en este caso, el enfoque aproxima la refracción de las secuencias unidas e isomórficas, proponiendo para ello la Ec. 3.27.

$$RI(ps_1, ps_2) = max\{RI(ps_1), RI(ps_2)\} \tag{3.27}$$

S-Match

Con el propósito de ofrecer una herramienta de búsqueda de asociaciones semánticas al más puro estilo de los buscadores convencionales, los autores de *SemRank* permiten que los usuarios completen sus consultas mediante palabras clave. Si bien estos términos permiten precisar la consulta y, con ello, incrementar las posibilidades de obtener resultados adecuados, también suponen cierta limitación para el sistema. Ésta es debida a que dichas palabras claves deben identificar propiedades de la base de conocimiento empleada; es decir, el usuario define dos entidades para conocer las asociaciones existentes entre ellas y, además, especifica propiedades concretas para destacar aquellas relaciones que más le interesan. Es obvio que este mecanismo obliga a los usuarios a conocer con detalle la ontología empleada, lo que en algunos dominios se torna una exigencia excesiva.

Dada la naturaleza de las palabras clave, el enfoque *SemRank* cuantifica la similitud entre cada una de ellas y las propiedades contenidas en las asociaciones semánticas descubiertas. A la hora de comparar dos propiedades, esta aproximación considera únicamente las relaciones de herencia establecidas entre éstas, de forma que cuanto más cerca estén en la jerarquía de propiedades, más similares son. Precisamente esta noción de distancia (número de enlaces jerárquicos entre las dos propiedades) se incluye en la métrica de similitud expresada en la Ec. 3.28 mediante el parámetro d. En esta expresión, además de la secuencia $PS = [P_1, \ldots, P_n]$, se incluye el conjunto de palabras

clave especificadas por el usuario $K = \{k_i, \ldots, k_m\}$, y el valor de similitud entre dos propiedades concretas $SemMatch(k_i, p_j)$.

$$SemMatch(P_i, k_j) = 0 < 2^d \leq 1 \tag{3.28}$$

La Ec. 3.28 conduce a un valor de similitud máximo cuando las palabras claves coinciden exactamente con las propiedades consideradas, ya que en este caso el parámetro d toma un valor 0.

De forma análoga a la comentada en las métricas anteriores, el enfoque propuesto parte de la Ec. 3.28 aplicada sobre propiedades individuales, y calcula el valor correspondiente a cada secuencia completa —denominado de forma genérica S-Match—, tal como se muestra en la Ec. 3.29.

$$\text{S-Match}(ps) = \sum_{i=1}^{n} max_{j=1}^{k}\{SemMatch(P_i, k_j)\} \tag{3.29}$$

La expresión aplicada en el caso de dos secuencias ps_1 y ps_2 unidas y/o isomórficas, es la indicada a continuación.

$$\text{S-Match}(ps_1, ps_2) = max\{\text{S-Match}(ps_1), \text{S-Match}(ps_2)\} \tag{3.30}$$

3.7.2 Criterio de clasificación global

Una vez calculadas las expresiones correspondientes a las cuatro métricas consideradas en el enfoque, [6] reúne todas las contribuciones en el índice *SemRank* relativo a cada asociación semántica. La forma en que dichas metricas se combinan se modula mediante un parámetro μ, que permite identificar el modo de consulta seleccionado por el usuario: convencional ($\mu = 0$) o basada en descubrimiento de conocimiento ($\mu = 1$).

Revisemos, a continuación, cuál es la influencia de dicho parámetro en cada una de las métricas propuestas en [6]. Para ello, el enfoque combina el parámetro μ citado antes, con las expresiones que define cada métrica para las secuencias de propiedades.

- **Especificidad**: Las secuencias de propiedades más específicas resultan más difíciles de predecir por parte de los usuarios. Es por ello que dichas secuencias son las recuperadas en el modo de búsqueda basada en descubrimiento de conocimiento. Por el contrario, las consultas convencionales se centran únicamente en aquellas secuencias que son más comunes en la conceptualización (y, por tanto, menos específicas). Ambas ideas aparecen reflejadas en la Ec. 3.31:

  $$SP(ps_\mu) = \mu \cdot SP(ps) + (1 - \mu) \cdot (1 - SP(ps)) \tag{3.31}$$

  Es evidente que en las consultas convencionales, las secuencias más específicas tendrán una componente $SP(ps_\mu)$ poco significativa, mientras que para las más comunes dicho valor será muy elevado. El razonamiento contrario corresponde al modo de consulta basado en descubrimiento de conocimiento.

- **Entropía**: Tras comprobar que la expresión de la entropía de una secuencia de propiedades (ver Ec. 3.22) favorecía a aquellas cuya longitud era mayor, el enfoque introdujo un proceso de normalización, dando como resultado el valor $NH(ps)$ mostrado en la Ec. 3.32.

$$NH(ps) = \frac{H(ps)}{length(ps)} \tag{3.32}$$

La forma en que la entropía normalizada de una asociación semántica contribuye a la clasificación de la misma, se modula de nuevo mediante el parámetro μ, tal como se muestra en la Ec. 3.33:

$$NH(ps_\mu) = \mu \cdot NH(ps) + (1 - \mu) \cdot (1 - NH(ps)) \tag{3.33}$$

En el modo de consulta basado en descubrimiento de conocimiento, los valores $NH(ps_\mu)$ más elevados corresponden a las secuencias de entropía alta, ya que son éstas las más difíciles de predecir para el usuario. Por el contrario, aquellas secuencias de entropía poco significativa resultan más fáciles de adivinar, de ahí que ocupen los primeros lugares en las consultas convencionales.

- **Refracción**: Teniendo en cuenta la definición de refracción dada en la sección previa, las secuencias que contienen instancias que se refractan serán las más difíciles de predecir para los usuarios. Por esta razón, en las consultas basadas en descubrimiento de conocimiento, el factor $RI(ps_\mu)$ incluido en la Ec. 3.34 toma valores más significativos para aquellas secuencias en las que ocurre el fenómeno de la refracción. Desde esta expresión, también es posible inferir que las consultas convencionales recuperan únicamente secuencias sin refracción.

$$RI(ps_\mu) = \mu \cdot RI(ps) \tag{3.34}$$

- **S-Match**: En este caso, no es necesario hacer diferencias entre ambos modos de consulta, ya que en los dos casos se clasifican en los primeros puestos aquellas secuencias cuyas propiedades son más similares a las palabras clave especificadas por el usuario, tal como indica la Ec. 3.35.

$$\text{S-Match}(ps_\mu) = \sum_{i=1}^{n} \text{S-Match}(p_i) \tag{3.35}$$

Finalmente, la Ec. 3.36 muestra la expresión correspondiente al índice *SemRank* de una asociación semántica A, cuyo valor depende de las métricas presentadas a lo largo de la Sección 3.7.1.

$$SemRank(A) = SP(ps_\mu) \cdot NH(ps_\mu) \cdot (1 + RI(ps_\mu)) \cdot (1 + \text{S-Match}(ps_\mu)) \tag{3.36}$$

3.8 Aplicación de las asociaciones semánticas complejas

Son varios los enfoques que han aprovechado el potencial de las asociaciones semánticas como mecanismo de inferencia de conocimiento. Entre tales propuestas destacamos el sistema BioPatentMiner [170], una herramienta que permite recuperar información sobre patentes biomédicas. Este sistema combina la información sobre las patentes con el conocimiento extraído desde diversas ontologías del dominio de la medicina. Sus capacidades de consulta soportan tanto la búsqueda basada en palabras

claves como el descubrimiento de asociaciones semánticas entre dos instancias especificadas por el usuario (referidas a los recursos identificados en las patentes). Una vez inferidas las asociaciones, BioPatentMiner clasifica los resultados de la búsqueda filtrando la información irrelevante para la consulta. Para ello, los autores de [170] definen una métrica que cuantifica la *importancia* de cada entidad definida en la base de conocimiento del sistema, basándose tanto en las relaciones jerárquicas establecidas entre ellas como en su especificidad (ver Sección 3.7.1).

Un segundo ejemplo de aplicación de las asociaciones semánticas complejas se describe en [194]. Este enfoque infiere dichas asociaciones entre las entidades RDF disponibles en una red *Peer-to-Peer* (P2P). En concreto, cada usuario tiene la posibilidad de especificar dos instancias concretas para que el enfoque recupere las asociaciones semánticas establecidas entre ellas. Si bien cada *peer* dispone de una o varias bases de conocimiento, es posible que las entidades especificadas no estén representadas en éstas. En este caso, es necesario acceder a los repositorios del resto de *peers* de la red para obtener la información requerida en la consulta del usuario. Para ello, los autores de [194] definen una topología jerárquica en la red P2P, en la que los *peers* se organizan en grupos, llamados *super-peers*, que cooperan activamente durante el proceso de consulta: los *super-peers* se comunican con sus homólogos en la red P2P a fin de descubrir nuevo conocimiento, y los *peers* lo emplean durante la inferencia de las asociaciones semánticas solicitadas por el usuario.

Un enfoque similar a éste se describe en [203], en el que el objetivo es explorar una grafo RDF para descubrir relaciones significativas entre dos instancias también especificadas por el usuario. Para ello, este trabajo cuantifica la relevancia de los nodos y propiedades relacionados con ambas instancias en el grafo, recurriendo a métricas inspiradas en la especificidad y la refracción definidas en el proyecto SemDis. Como resultado de este proceso, el enfoque selecciona un subgrafo en el que se identifican las relaciones establecidas entre las dos instancias identificadas inicialmente, a través de otros nodos conectados con ellas mediante secuencias de propiedades.

A pesar de sus capacidades inferenciales, los enfoques descritos en esta sección también presentan ciertas limitaciones. La más relevante es la falta de flexibilidad del paradigma de consulta que implementan, que sólo permite descubrir asociaciones semánticas entre dos instancias previamente especificadas.

Además de la inferencia de las asociaciones semánticas, en la literatura se han definido mecanismos alternativos que también permiten explorar bases de conocimiento y descubrir en las mismas información relevante en el contexto de cada sistema. Entre estos enfoques, de gran utilidad en el dominio de la Web Semántica, destacan las técnicas de propagación de activación.

3.9 Técnicas de Propagación de Activación

El modelo *Spreading Activation* (en lo que sigue SA) fue usado por primera vez en el campo de la Inteligencia Artificial, y más tarde se extendió a áreas de muy diversa naturaleza, entre las que podemos destacar las bases de datos, la psicología, la biología o incluso la recuperación de información.

Si bien este tipo de técnicas se han aplicado normalmente en las redes semánticas, algunos trabajos han adoptado un enfoque diferente, empleándolas como estructura de procesado en cualquier red genérica en la que los conceptos y las relaciones estable-

cidas entre éstos se representan mediante nodos y enlaces, respectivamente. Debido a esto último, en la literatura es posible encontrar propuestas en las que estas técnicas se utilizan en redes asociativas e, incluso, en redes neuronales, tal como veremos en las próximas secciones.

El funcionamiento de este tipo de técnicas se basa en lo siguiente: los nodos de la red tienen una relevancia implícita en la misma, denominada (nivel de) activación. Inicialmente, se seleccionan un conjunto de nodos y se propagan sus niveles hasta alcanzar a sus vecinos, es decir, los nodos de la red con los que están unidos mediante enlaces ponderados. Durante este proceso de propagación, los niveles de los nodos alcanzados se actualizan a partir de los valores correspondientes a sus vecinos y de los pesos de los enlaces que los unen. Tras aplicar este proceso en toda la red, los niveles más elevados identifican nodos relacionados de forma significativa con los inicialmente activados. Esta funcionalidad ha permitido la aplicación de las técnicas SA como mecanismos eficientes para la exploración e inferencia de conocimiento, a partir de las relaciones conceptuales representadas en redes genéricas, de ahí su indudable utilidad en el dominio de la Web Semántica.

Este modelo básico, acuñado bajo el término de **técnicas SA puras**, ha sido mejorado incluyendo un conjunto de restricciones que permiten controlar el proceso de propagación de la activación, de ahí que dichos métodos reciban el nombre de **técnicas SA restringidas**. Además de revisar ambos enfoques, en esta sección nos centramos también en la descripción de diferentes algoritmos basados en técnicas SA propuestos en el estado del arte.

3.9.1 Técnicas puras de propagación de activación

Tal como se describe en [56], el modelo de las llamadas técnicas SA puras adopta la estructura genérica mencionada al comienzo de la sección: una red de nodos (conceptos) interconectados mediante enlaces (relaciones) sobre la que se aplican procedimientos de procesado sencillos. Dicho procesado incluye tres etapas: ajuste inicial (*preadjustment*), propagación (*spread*) y ajuste final (*postadjustment*).

La primera y la tercera de las etapas son opcionales, y permiten controlar en qué condiciones se lleva a cabo la propagación en la segunda etapa. Por ejemplo, durante ambas fases es posible determinar qué nodos de la red van a ser más relevantes durante el proceso de propagación (activados inicialmente) o, incluso, establecer alguna condición específica para que éste se vea interrumpido. Como su nombre indica, la etapa intermedia consiste en propagar valores numéricos (llamados *salidas de activación*) desde un nodo a todos aquellos que están directamente conectados con él, es decir, sus nodos vecinos. Estos valores son usados para calcular las llamadas *entradas de activación* de cada nodo alcanzado en el proceso de activación, tal como muestra la Ec. 3.37 en la que I_j es la entrada de activación del nodo j, O_i es la salida del nodo i conectado a j, y w_{ij} el peso asociado al enlace que conecta ambos nodos.

$$I_j = \sum_i O_i \cdot w_{ij} \tag{3.37}$$

Los valores asignados a las entradas y salidas de activación y a los pesos de los enlaces dependen de la aplicación que se modele en la red. Así, por ejemplo, en redes semánticas los pesos suelen ser binarios (0 ó 1), o incluso pueden reflejar valores de activación e inhibición (1 y -1, respectivamente). Sin embargo, en las redes neuronales

estos pesos suelen tomar valores reales que indican la fuerza de la relación establecida entre los nodos conectados.

Una vez calculada la entrada de activación de un nodo j, es necesario calcular su *nivel de activación* a partir de ésta, tal como se muestra en la Ec. 3.38.

$$A_j = f(I_j) \tag{3.38}$$

donde A_j y f son el nivel de j y la función de activación, respectivamente.

Las funciones de activación empleadas en las técnicas de propagación coinciden con las utilizadas en las redes neuronales artificiales descritas en la Sección 2.2.7. Los modelos más empleados, correspondientes a funciones lineales, umbrales y sigmoides, pueden ser consultados en la Fig. 2.1 representada en dicha sección.

En el caso concreto de la función umbral, los niveles de activación resultantes serían los mostrados en la Ec. 3.39.

$$A_j = \begin{cases} 0 & \text{si } I_j < k_j \\ 1 & \text{en otro caso} \end{cases} \tag{3.39}$$

donde k_j es el umbral establecido para el nodo j de la red. Al igual que el resto de valores, dicho umbral también depende del dominio de aplicación de cada sistema, y puede variar incluso de un nodo a otro de la red.

Típicamente, la salida de cada nodo suele coincidir con su nivel de activación. Así, de acuerdo a la Ec. 3.37, si un nodo tiene nivel de activación nulo, propaga valores nulos a todos los nodos conectados con él en la red.

3.9.2 Técnicas de propagación de activación restringida

La implementación de las técnicas SA puras en sistemas reales puso de manifiesto algunas limitaciones importantes:

- Por un lado, diversos experimentos demostraron que, a menos que las etapas de ajuste inicial y final estuvieran muy controladas, la activación tendía a propagarse rápidamente por toda la red.

- Por otra parte, las técnicas de propagación de activación puras omiten la semántica de las relaciones identificadas en la red, lo que dificulta la aplicación de procesos inferenciales sobre ésta.

Estas limitaciones podrían superarse estableciendo restricciones que controlen el proceso de propagación llevado a cabo en la red. Este enfoque dio lugar a un modelo más elaborado que las técnicas SA puras, conocido como técnicas de propagación de activación restringida (*Constrained Spreading Activation* o CSA). Algunas de las restricciones en las que se basa la CSA son las enumeradas a continuación:

- *Restricción de distancia*: Está basada en limitar la máxima distancia a la que puede propagarse la activación en la red. En esta restricción se asume que el nivel de activación de un nodo decrece con la distancia entre éste y los nodos inicialmente activados.

- *Restricción de fan-out*: El proceso de propagación debería detenerse al alcanzar un nodo con un *fan-out* elevado, esto es, un nodo conectado con otros muchos en la red. Esto es debido a que dicho nodo podría provocar un proceso de propagación incontrolado en el que se activase la red en su totalidad.

- *Restricción de camino favorito*: La activación se propagaría a lo largo de un conjunto de caminos preferidos que deberían ser identificados en función del dominio de aplicación. Esta restricción se puede modelar empleando los pesos asignados a los enlaces de la red.

- *Restricción de activación*: Establece un valor umbral sobre el nivel de activación de un único nodo. Así, si dicho nivel es inferior al umbral de activación, el proceso de propagación no debería continuar a partir de este nodo (en la técnica pura el nivel no se propaga únicamente cuando su valor es nulo).

Una vez definidos los modelos de las técnicas SA, los investigadores se centraron en desarrollar algoritmos que aprovechasen sus capacidades para explorar grandes redes de conocimiento —tanto redes semánticas como neuronales—, y extraer desde las mismas conceptos relevantes en cada dominio de aplicación. En el estado del arte, son dos los algoritmos que gozan de mayor popularidad, denominados de forma genérica *Branch-and-Bound* y *Hopfield Net*. Ambas propuestas trabajan sobre un conjunto de conceptos especificados por el usuario, a partir de los cuales se extraen aquellos nodos de la red que están relacionados con los primeros de forma significativa.

3.9.3 El algoritmo Branch-and-Bound

Este algoritmo adopta como modelo de representación de conocimiento una red semántica, de ahí que los enlaces tengan asociadas etiquetas semánticas que identifican las relaciones existentes entre los conceptos definidos en la misma [46]. Basándose en el principio de búsqueda *branch-and-bound*, este algoritmo (en adelante BNB) explora los caminos de longitud mínima establecidos entre los conceptos que el usuario ha especificado y aquellos que son finalmente recuperados. De esta forma, el algoritmo garantiza que dichos conceptos son los más relevantes para la consulta del usuario, ya que son los que están relacionados de forma más significativa (a través de los caminos más cortos) con los indicados por el mismo.

Inicialmente, el algoritmo BNB activa el conjunto de nodos correspondientes a la consulta (que en adelante suponemos serán $\{S_1, S_2, \ldots, S_m\}$), a los que asigna un nivel de activación de valor 1, y construye dos colas:

- La cola de prioridad ($Q_{priority}$): En la primera iteración del algoritmo, esta cola registra el conjunto de nodos inicialmente activados, ordenados de acuerdo a su nivel de activación, de forma que la cabeza de la misma la ocupa siempre el nodo de mayor nivel.

$$Q_{priority} = \{S_1, S_2, \ldots, S_m\}$$

- La cola de salida (Q_{out}): Esta cola, inicialmente vacía, almacena todos los nodos que se vayan activando durante el proceso de propagación, junto a sus respectivos niveles.

$$Q_{out} = \{\ \}$$

En cada iteración, el algoritmo:

1. Extrae de la cola de prioridad el nodo de mayor nivel de activación. En caso de que sean varios los nodos que registran este valor máximo, todos ellos son extraídos en la misma iteración.

2. Activa únicamente los vecinos de este nodo, unidos directamente a él mediante enlaces en la red de propagación.

3. Calcula los nuevos niveles de activación de dichos vecinos. Para ello, en cada iteración el nivel del vecino se calcula multiplicando el nivel que tenía en la iteración anterior el nodo que lo activó (seleccionado en 1) por el peso del enlace que une ambos nodos. Así, siendo j el nodo vecino e i el nodo extraído de la cola de prioridad, en $t+1$ se cumple que $A_j(t+1) = A_i(t) \cdot w_{ij}$.

4. Añade dichos vecinos a la cola de prioridad, colocándolos en el lugar que les corresponda según su nuevo nivel de activación.

5. Dado que los nodos vecinos han sido activados, deben ser añadidos a la cola de salida. Destacar que, en caso de que dichos nodos ya se hubiesen almacenado en iteraciones previas en esta cola, sus niveles de activación son actualizados sumando al valor que ya tenían, el calculado mediante el procedimiento descrito en el paso 3.

El algoritmo itera hasta que se alcanza una determinada condición de parada, por ejemplo, número fijo de elementos en la cola de salida, cola de prioridad vacía, etc.

El enfoque adoptado en BNB puede tener una utilidad limitada en algunos dominios de aplicación, en concreto, en aquellos en los que las relaciones más significativas para el usuario no sean descubiertas a partir de caminos de longitud mínima. Precisamente para descubrir dichos caminos, el algoritmo activa en cada iteración únicamente los vecinos del nodo más relevante de la red, es decir, aquel que tiene mayor nivel de activación, y que, por tanto, ocupa la cabeza de la cola de prioridad. En caso de que el tipo de relaciones de interés no sean tan directas, parece más adecuado computar y evaluar los niveles de todos los nodos de la red en una misma iteración. Este enfoque es el adoptado en el algoritmo Hopfield Net.

3.9.4 El algoritmo Hopfield Net

El algoritmo Hopfield Net [47] (en adelante, Hopfield) está basado en la red neuronal que lleva el mismo nombre [98], cuyas principales características han sido revisadas en la Sección 2.2.7 del capítulo anterior. Por esta razón, los nodos y enlaces de la red de propagación coinciden con las neuronas y sinapsis, respectivamente, en la red neuronal.

Hopfield propone una alternativa interesante al proceso de activación aplicado en el algoritmo BNB. Para ello, incorpora dos de las características más destacables de la red neuronal Hopfield: su principio de búsqueda en paralelo y sus propiedades de convergencia.

- Búsqueda en paralelo: Por un lado, el algoritmo activa, en cada iteración, todos los nodos de la red en paralelo, calculando los niveles de cada uno de ellos en función de los niveles de los restantes.
- Convergencia: Por otra parte, el enfoque recorre los diferentes nodos de la red, iteración tras iteración, hasta que sus niveles de activación convergen a un valor estable. En este sentido, destacar que una de las principales contribuciones de Hopfield fue demostrar que, dado cualquier estado inicial de la red neuronal y cualquier combinación de pesos en los enlaces (sinapsis), el algoritmo conducía siempre a un estado estable [117].

El funcionamiento detallado del algoritmo se describe a continuación:

1. Inicialmente, y al igual que en BNB, el algoritmo asigna un peso de valor uno a todos los nodos seleccionados por el usuario, y de valor cero a los restantes. Así, asumiendo una red de propagación con n nodos, tenemos que el nivel de activación de un nodo genérico x en t=0 será:

$$A_x(0) = \begin{cases} 1 & \text{si } x \in \{S_1, \ldots, S_m\} \\ 0 & \text{en otro caso} \end{cases} \tag{3.40}$$

2. En la iteración correspondiente a t, se calculan los niveles de todos los nodos de la red, esto es:

$$A_j(t+1) = f_S \left[\sum_{i=0}^{n-1} A_i(t) \cdot w_{ij} \right], \quad 0 \leq j \leq n-1 \tag{3.41}$$

donde f_S es la función sigmoide [117] representada en la Fig. 2.1, cuya expresión analítica se muestra a continuación:

$$f_S(net_j) = \frac{1}{1 + exp\left[\frac{\theta_1 - net_j}{\theta_2}\right]} \tag{3.42}$$

donde

- $net_j = \sum_{i=0}^{n-1} A_i(t) \cdot w_{ij}$,
- θ_1 es un umbral configurable y
- θ_2 es un parámetro usado para modificar la forma de la función sigmoide.

La Ec. 3.41 muestra la propiedad de búsqueda en paralelo de la red Hopfield, ya que en cada iteración del algoritmo todos los nodos de la red se activan a la vez. Basándose en ese principio de activación en paralelo, el nivel de activación de cada nodo individual se calcula sumando las contribuciones de *todos* sus nodos vecinos (tal como demuestra la expresión $net_j = \sum_{i=0}^{n-1} A_i(t) \cdot w_{ij}$ en la Ec. 3.41). Tal como adelantábamos en la sección previa, este proceso de activación es la principal diferencia entre los dos algoritmos revisados. Así, en BNB si un nodo tiene varios vecinos, en cada iteración su nivel se calcula considerando sólo la contribución del vecino que lo activó, en lugar de considerar el vecindario completo, tal como propone Hopfield.

3. El proceso descrito se repite hasta que los niveles de activación de todos los nodos de la red converjan. Para comprobar esta condición, se consideran las dos últimas iteraciones del algoritmo y se aplica la Ec. 3.43, donde es obvio que ξ debe ser un parámetro configurable que toma valores poco significativos.

$$\sum_{j=0}^{n-1} |A_j(t+1) - A_j(t)| \leq \xi \tag{3.43}$$

A modo de conclusión, en la Tabla 3.1 se resumen las principales características de los dos algoritmos basados en técnicas SA descritos en esta sección.

Algoritmo BNB	**Algoritmo Hopfield**
• Basado en redes semánticas. • Activación en serie: en cada iteración se activan sólo los vecinos del nodo extraído de $Q_{priority}$. • El nivel de un nodo activado se calcula multiplicando el nivel del nodo que lo activó por el peso del enlace que los une. • Condiciones de parada: - Número determinado de nodos en Q_{out}. - Cola $Q_{priority}$ vacía.	• Basado en redes neuronales. • Activación en paralelo: en cada iteración se activan todos los nodos de la red. • El nivel de activación de un nodo se calcula considerando los niveles de todos sus vecinos y los pesos de los enlaces que los unen. • Condición de parada: - Convergencia de los niveles de activación de los nodos de la red.

Tabla 3.1: Comparativa entre los algoritmos Branch-and-Bound y Hopfield Net

Dado que las técnicas SA ofrecen facilidades para explorar bases de conocimiento y realizar inferencias muy útiles, no es de extrañar su aplicación en el dominio de la Web Semántica, donde se han definido diferentes enfoques para asistir a los usuarios en las tareas de búsqueda y recuperación de información.

3.10 Búsqueda de información en la Web Semántica

La revisión del estado del arte nos conduce a numerosos motores de búsqueda que se alimentan de las capacidades de razonamiento automático provistas en la Web Semántica.

3.10.1 Motores de búsqueda semánticos

Herramientas basadas en técnicas SA

Entre los enfoques propuestos empezamos revisando el trabajo presentado en [209], donde se describe una herramienta semántica que combina los métodos de búsqueda clásicos definidos en la Web con las técnicas SA, con el objetivo de recuperar aquellos documentos que son más relevantes para la consulta formulada por el usuario. Dado que la mayoría de los usuarios están acostumbrados a expresar dichas consultas mediante una secuencia de palabras claves, esta aproximación propone un motor de búsqueda que, a la vez que ofrece este tipo de facilidades, incorpora capacidades de procesado semántico. Tales capacidades utilizan las técnicas SA a la hora de encontrar conceptos relacionados con aquellos que los usuarios especifican en sus consultas.

En concreto, [209] implementa las técnicas SA restringidas (ver Sección 3.9.2) mediante el algoritmo BNB. Tal como explicamos en la Sección 3.9.3, este algoritmo requiere tanto la activación inicial de un conjunto de conceptos de la red de propagación, como la asignación de un peso numérico a cada uno de los enlaces definidos en la misma.

La herramienta propuesta ofrece una interfaz a través de la cual los usuarios formulan sus consultas especificando un conjunto de palabras claves. El enfoque propuesto recupera, basándose en los métodos sintácticos clásicos, una relación de documentos Web relevantes para dicha consulta, y los procesa como sigue: en primer lugar, la herramienta extrae de los documentos seleccionados aquellos conceptos que están formalmente representados en la ontología de conocimiento considerada en el sistema; son precisamente éstos los conceptos inicialmente activados en las técnicas CSA aplicadas en [209]. A continuación, el enfoque calcula el peso correspondiente a cada enlace de la red de propagación. Para ello, esta aproximación extrae conocimiento desde las instancias representadas en la ontología, y cuantifica así la fuerza de las relaciones establecidas en el modelo. Este proceso de cuantificación se basa en dos métricas diferentes, finalmente combinadas mediante su producto.

- *Similitud*: Esta métrica establece el grado de similitud existente entre las dos instancias relacionadas mediante un enlace explícito en la red de propagación. En concreto, siendo C_j y C_k dos instancias en la ontología considerada, su similitud se calcula mediante la Ec. 3.44:

$$Sim(C_j, C_k) = \frac{\sum_{i=1}^{N} n_{ijk}}{\sum_{i=1}^{N} n_{ij}} \tag{3.44}$$

  donde n_{ij} toma el valor 1 si existe una propiedad explícita entre los conceptos C_i y C_j, y 0 en caso contrario. De forma análoga, n_{ijk} vale 1 si tanto C_j como C_k están relacionados con C_i mediante sendas propiedades. Dicho de otra forma, la Ec. 3.44 conduce a valores de similitud elevados entre aquellos conceptos que comparten muchas relaciones con otros nodos de la red.

- *Especificidad*: Si bien la idea en la que se basa esta métrica es la misma que la propuesta en el proyecto SemDis (ver Sección 3.7.1), la expresión analítica empleada es diferente:

$$Sp(C_j, C_k) = \frac{1}{\sqrt{n_k}} \tag{3.45}$$

donde n_k identifica el número de propiedades establecidas entre C_k y las instancias de la clase a la que pertenece C_j. En caso de existir pocas instancias de la misma clase que C_j relacionadas con C_k, el enlace que conecta ambos conceptos en la red de propagación sería muy específico y, por tanto, la métrica $Sp(C_j, C_k)$ tomaría un valor elevado.

Además de combinar ambas métricas (mediante su producto), el enfoque descrito en [209] asigna a cada enlace de la red SA un índice de relevancia, que permite dar más peso a aquellas propiedades de la ontología que se consideren más significativas en cada dominio de aplicación[5]. Reuniendo estas ideas, resulta la Ec. 3.46, en la que w_{jk} y f_{jk} son el peso correspondiente al enlace que une los conceptos C_j y C_k en la red de propagación, y el índice de relevancia de dicho enlace, respectivamente.

$$w_{jk} = Sim(C_j, C_k) \cdot Sp(C_j, C_k) \cdot f_{jk} \tag{3.46}$$

Habida cuenta de la aplicación de las técnicas CSA en [209], el enfoque también requiere el establecimiento de los parámetros de configuración relativos a cada una de las restricciones consideradas, tal como se describió en la Sección 3.9.2: distancia de propagación, *fan-out*, nivel de activación mínimo, etc.

Una vez definidos los valores de todos los parámetros de configuración, el enfoque propuesto aplica el algoritmo BNB tal como se explicó en la Sección 3.9.3. De esta forma, se inicializa la cola de prioridad con los conceptos inicialmente activados, y cada vez que uno de ellos es extraído de la misma, se comprueba si cumple las restricciones consideradas. En caso afirmativo, se activan sus vecinos, se calculan sus nuevos niveles de activación y se reorganizan tanto la cola de prioridad como la de salida. Tras finalizar el proceso de propagación, la cola de salida almacena los conceptos relevantes para la consulta del usuario, ordenados con arreglo a su nivel de activación, y a partir de los cuales es posible indexar rápidamente los documentos que deben ser recuperados para el usuario.

Un trabajo de características similares a las del propuesto en [209] se describe en [235], donde sus autores definen un enfoque que permite clasificar, según su relevancia, los resultados recuperados para las consultas formuladas por los usuarios. Este proceso de clasificación no sólo considera la estructura subyacente de la ontología de conocimiento, sino que se basa en métricas como la especificidad —análoga a la propuesta en [209] (ver Ec. 3.45)— que permite valorar la importancia de cada propiedad (o conjunto de propiedades) con respecto a dichas consultas.

Junto a la ausencia de las técnicas SA, la principal diferencia entre ambas propuestas es que [235] exige a los usuarios que formulen sus peticiones especificando instancias concretas de la ontología utilizada, en lugar de palabras clave. A consecuencia de esto último, los usuarios deben tener cierto grado de conocimiento sobre la estructura de la conceptualización, para así poder formular sus consultas correctamente.

Al igual que en el enfoque propuesto en [209], el sistema ONTOCOPI [173] también aplica técnicas SA, con el objetivo de formar las llamadas *comunidades de prácticas* dentro de una organización, es decir, aquellos conceptos que están relacionados con uno dado de forma significativa. Para ello, este trabajo representa el conocimiento sobre el dominio de la organización considerada mediante una ontología. Su aproximación consiste en partir de aquellas instancias cuya comunidad se desea formar, y recorrer

[5]En los enfoques tradicionales, se considera que todas las propiedades son igual de importantes, de ahí que dicha función tome el valor 1.

las propiedades que están asociadas a ellas en la base de conocimiento, aplicando las técnicas SA tal como se ha descrito en la Sección 3.9.

En ONTOCOPI, los pesos numéricos correspondientes a cada una de estas propiedades pueden ser asignados de forma manual o automática. En el primer escenario, son los usuarios quienes asumen un rol activo en el sistema, y establecen los pesos más elevados para aquellas propiedades que son más relevantes en su consulta. Por ejemplo, imaginemos que un usuario está interesado en conocer la comunidad de prácticas de una instancia referida a un profesor de Universidad en una ontología sobre el dominio académico. En este escenario, dicho usuario podría asignar los pesos más elevados a las propiedades que permitan identificar a otros colegas que trabajan en las mismas áreas de investigación que dicho profesor (*comparteInvestigacionCon*), que han publicado artículos con él (*CoautorCon*), alumnos a los que ha dirigido la tesis doctoral (*directorTesisDe*), etc.

Por el contrario, en el caso de que la asignación de pesos sea automática, el sistema pondera las relaciones (enlaces en la red de propagación) basándose en su frecuencia de aparición en la ontología. Así, el enfoque considera que aquellas propiedades que son más frecuentes en la base de conocimiento, son también más relevantes, de ahí que les corresponda un peso numérico elevado.

Una vez establecidos los pesos e identificada la instancia cuya comunidad debe ser formada, ONTOCOPI aprovecha el procesado semántico que ofrecen las técnicas SA para recorrer la red semántica, extrayendo los conceptos relacionados con dicha instancia y clasificándolos en función de los pesos de las propiedades atravesadas en la misma. En realidad, el enfoque emplea una de las restricciones de las técnicas CSA, de forma que únicamente es posible recorrer un número fijo de enlaces desde la instancia original. Este umbral, configurable en el sistema, evita que instancias poco significativas estén incluidas en la comunidad de prácticas resultante.

Otro de los enfoques que aprovecha las capacidades de inferencia de conocimiento de las técnicas SA ha sido propuesto en [122]. La aproximación allí descrita se basa en crear un modelo de propagación de activación modificado, en el que los pesos de los enlaces se calculan considerando el contexto individual de cada usuario. Como consecuencia, el peso asociado a un mismo enlace en dicho modelo semántico puede variar de un usuario a otro, en función de la información contextual relativa a su consulta. Para este propósito, el enfoque asocia a cada enlace de la red de propagación, un nuevo nodo que representa dicho contexto. La contribución de dicho nodo debe ser considerada durante la ponderación de los enlaces a los que está asociado. Al variar el peso de estos enlaces, los autores de [122] también deben modificar la expresión del cálculo del nivel de activación de un nodo (dependiente tanto de sus vecinos como de los enlaces establecidos entre ellos) para poder llevar a cabo el proceso de propagación descrito en la Sección 3.9. Combinando el conocimiento inferido mediante este proceso con la información contextual de la consulta, el enfoque mejora la calidad de los resultados mostrados a los usuarios.

Otros ejemplos de aplicación de las técnicas SA como métodos de procesado para recuperar e indexar documentos relevantes para las consultas de los usuarios, pueden ser consultados en [45, 102, 110]. Estos enfoques reutilizan las ideas descritas en los trabajos anteriores, adaptándolas al dominio de aplicación concreto de cada tipo de sistema. Por su estrecha relación con el ámbito de esta tesis, describimos el enfoque propuesto en [102]. Este trabajo aprovecha la capacidad de exploración eficiente de conocimiento de estas técnicas, para mitigar los efectos del *sparsity problem* asociado

a los enfoques colaborativos tradicionales (ver Sección 2.6.3). Su aproximación utiliza el proceso de propagación de este mecanismo para descubrir asociaciones *transitivas* entre los usuarios del sistema recomendador, que permiten detectar similitud entre sus preferencias aun cuando no hayan clasificado *exactamente* los mismos productos en sus perfiles.

El sistema QuizRDF

Otro enfoque alternativo en el campo de la búsqueda de información en la Web Semántica fue propuesto en [62]. Al igual que el trabajo descrito en [209], el sistema QuizRDF combina las búsquedas tradicionales basadas en palabras claves con la posibilidad de consultar y explorar las anotaciones RDF de los recursos Web. Así, a diferencia de las herramientas tradicionales, QuizRDF no sólo busca dichas palabras claves en la información textual definida en los documentos Web, sino también en las descripciones RDF presentes en los mismos. Esta característica, sin embargo, no supone una mejora sustancial con respecto a los motores de búsqueda existentes en el Web, ya que, si bien el enfoque incorpora las anotaciones RDF a los documentos, su mecanismo de recuperación de información sigue siendo sintáctico, carente de cualquier proceso de razonamiento automático.

Tal limitación se refleja en el tipo de consultas que ofrece QuizRDF, tal como apuntan sus autores en [62]. Éstas giran únicamente en torno a las clases de la ontología RDF, de forma que es imposible formular consultas encadenadas en las que se incluyan clases e instancias de éstas junto con propiedades que permiten relacionarlas con otras entidades de la ontología. Por ejemplo, el sistema permite resolver consultas como "Extraer todos los pintores cuyo nombre es Pablo", pero no soporta otras de la forma "Extraer todos los cuadros cuyo autor sea un pintor apellidado Picasso". Gracias a las capacidades de procesamiento semántico incluidas en las técnicas SA, este tipo de consultas encadenadas serían fácilmente resueltas en enfoques como [209].

El sistema Semantic Search

Como su nombre indica, el sistema Semantic Search es un motor de búsqueda construido sobre la infraestructura de la Web Semántica, diseñado para superar las limitaciones de las que adolecen los enfoques convencionales [83]. Para este propósito, la herramienta completa las consultas formuladas por los usuarios, recurriendo a metadatos extraídos de diferentes fuentes. Al igual que en las herramientas revisadas previamente, dichas consultas son especificadas mediante un conjunto de palabras claves que deben ser traducidas a conceptos formalizados en la ontología subyacente, para, de esta manera, poder recuperar la información solicitada.

La idea de recorrer las instancias de la base de conocimiento a través de las propiedades de la misma está también contemplada en esta aproximación, pero el objetivo es completamente diferente al planteado en enfoques como [209] u ONTOCOPI [173]. En Semantic Search, el proceso de exploración de la ontología persigue únicamente encontrar los conceptos que se han añadido a la consulta del usuario, ya que éstos se consideran más relevantes para la misma. Por el contrario, en los dos enfoques citados anteriormente el propósito es realizar inferencias gracias a la aplicación de las técnicas SA, en lugar de identificar únicamente un conjunto de entidades previamente seleccionadas.

El sistema SCORE

Otro de los buscadores semánticos propuestos en la literatura es SCORE [226] (*Semantic Content Organization and Retrieval Engine*), un sistema cuya característica distintiva es la capacidad de resolver consultas en diferentes dominios de aplicación. Para identificar el contexto concreto de cada consulta, esta herramienta emplea tecnologías que ofrecen facilidades de clasificación automática y extracción de información. Tal como describimos en la Sección 2.6.2 del capítulo anterior, las primeras son útiles a la hora de clasificar los documentos Web considerados durante el proceso de búsqueda en diferentes categorías genéricas, relevantes en el dominio de aplicación concreto vinculado a la consulta. Una vez identificadas estas categorías, SCORE extrae de forma automática los metadatos semánticos que describen sus principales características, formalmente representados en una ontología de conocimiento.

SCORE incorpora un motor semántico que procesa las consultas (basadas en palabras claves) especificadas por los usuarios. Para ello, el sistema no incorpora mecanismos de inferencia avanzados, sino que simplemente recorre las instancias y propiedades definidas en la ontología de cada dominio de aplicación, identificando (sintácticamente) los metadatos más relevantes en el contexto de la consulta, de una forma similar a la descrita en la herramienta Semantic Search [83]. Aquellos documentos que contengan dichos metadatos son los finalmente recuperados para el usuario.

El sistema Froogle

Froogle es el motor de búsqueda especializado en productos comerciales desarrollado por Google [74]. Los usuarios del sistema especifican una relación de palabras claves sobre los productos que desean comprar (por ejemplo, nombre, modelo, funcionalidad, etc.), y la herramienta los recupera de forma automática. Para ello, este enfoque establece correspondencias (sintácticas) entre las palabras claves indicadas por el usuario y los metadatos que describen los productos conceptualizados en una base de conocimiento. Es por tanto un enfoque de capacidades inferenciales limitadas, en la misma línea que SCORE [226] y Semantic Search [83].

El sistema Swoogle

Swoogle es otro de los sistemas de recuperación de información propuestos en el dominio de la Web Semántica [66]. Esta herramienta permite descubrir, de forma automática, tanto documentos anotados mediante RDF u OWL, como ontologías completas, a partir de un conjunto de palabras claves especificadas por el usuario. Así, por una parte, Swoogle permite encontrar ontologías concretas, en las que incluso es posible especificar determinadas restricciones que deben cumplir sus clases y/o propiedades. Por otro lado, en el caso de los documentos OWL y RDF, el sistema indexa los metadatos que describen la semántica de los mismos y recupera aquellos que son más relevantes para cada consulta, empleando en este proceso tanto la información textual contenida en dichos documentos como sus anotaciones semánticas, de forma similar a la comentada en la herramienta QuizRDF [62].

A la hora de presentar los resultados a los usuarios, Swoogle emplea un mecanismo para clasificar los documentos recuperados, inspirado en el algoritmo PageRank [185, 90] propuesto en Google [81]. PageRank se basa en medir la probabilidad de visitar un documento elegido de forma aleatoria, alcanzado directamente o bien a través de

los enlaces incluidos en otros documentos que lo apuntan. Así, un documento será más relevante cuanto mayor sea la probabilidad de ser visitado, es decir, cuanto mayor sea el número de documentos en la Web que lo apuntan.

Utilizando esta misma aproximación, Swoogle infiere varios tipos de enlaces entre los documentos de la Web Semántica, y los emplea a la hora de clasificar los mismos y mostrarlos al usuario en un orden adecuado. Dados dos documentos A y B, el sistema es capaz de identificar hasta cuatro tipos diferentes de relaciones entre ellos:

- *imports* (A,B): En este caso, el documento A importa todo el contenido de B.
- *uses-term* (A,B): A emplea términos definidos en B, sin necesidad de importar su contenido.
- *extends* (A,B): En esta categoría, A extiende las definiciones de los términos incluidos en B.
- *asserts* (A,B): El documento A contiene declaraciones sobre las instancias definidas en B.

El mecanismo de clasificación propuesto en Swoogle asigna pesos numéricos diferentes a cada uno de los enlaces que acabamos de describir. El peso de un enlace entre dos documentos será más elevado cuanto más probable sea alcanzar uno de estos documentos, apuntado desde el otro. Por ejemplo, si un usuario visita un documento A en el que existe la relación *uses-term* (A,B), es lógico que siga este enlace y visite también el documento B, ya que ambos incluyen términos pertenecientes a un mismo dominio; por esta razón, Swoogle asigna un peso elevado a dicho enlace. En el caso concreto de la relación *uses-term* (A,B), un documento B será más relevante para el sistema cuantos más términos del mismo sean refenciados desde A, ya que con ello se incrementa la probabilidad de que un usuario visite el documento B una vez visitado A. Un razonamiento análogo puede ser aplicado en los restantes tipos de relaciones mencionadas antes.

A modo de conclusión, destacar que un documento es más relevante para Swoogle cuanto más significativos sean los pesos correspondientes a cada una de las relaciones descubiertas entre este documento y los restantes disponibles en la Web, y cuanto mayor sea el número de dichas relaciones.

Enfoques basados en Indexación Semántica Latente (LSI)

La indexación semántica latente (en lo que sigue LSI, de *Latent Semantic Indexing*) es un método estadístico diseñado con el mismo objetivo que las aproximaciones que hemos revisado en esta sección: superar las limitaciones de los sistemas de recuperación de información (sintácticos) existentes en el dominio de la Web.

A la hora de recuperar la información relevante para la consulta de cada usuario, la LSI no representa los documentos Web mediante vectores de los términos que éstos contienen (como ocurre en los enfoques tradicionales), sino que se basa en los conceptos semánticos definidos en dichos documentos. Esta aproximación parece bastante lógica si tenemos en cuenta que un mismo concepto puede ser descrito en varios documentos empleando muchos términos diferentes.

De esta forma, es posible que las herramientas de búsqueda basadas en LSI consideren en sus consultas un espacio conceptual de dimensión reducida (en comparación

con las herramientas sintácticas convencionales), debido precisamente a que el número de conceptos es mucho menor que el número de términos relacionados con los mismos. Además de un incremento en la eficiencia y de las ventajas computacionales que conlleva esta reducción dimensional, la LSI también permite recuperar documentos que contienen términos no necesariamente definidos en las consultas especificadas por los usuarios.

Para poder crear el mencionado espacio conceptual de dimensión reducida, la LSI emplea una técnica matemática denominada descomposición singular de valores (en adelante SVD, de *Singular Value Descomposition*). El proceso de aplicación de esta técnica en LSI se describe como sigue:

- En primer lugar, es necesario generar una matriz A, cuyas columnas representan los documentos considerados durante el proceso de búsqueda de información, y cuyas filas se refieren a los términos que figuran en éstos, de forma que un elemento a_{ij} de A identifica la frecuencia de ocurrencia del término i en el documento j.
- A continuación, se utiliza SVD para descomponer la matriz A en tres matrices diferentes, tal como se representa en la Fig. 3.3. La primera de ellas representa los m conceptos (columnas) a los que se refieren cada uno de los términos (filas) definidos en A. La segunda identifica las relaciones existentes entre los conceptos considerados (filas y columnas), y la tercera de las matrices representa los documentos (columnas) en los que dichos conceptos (filas) están definidos.

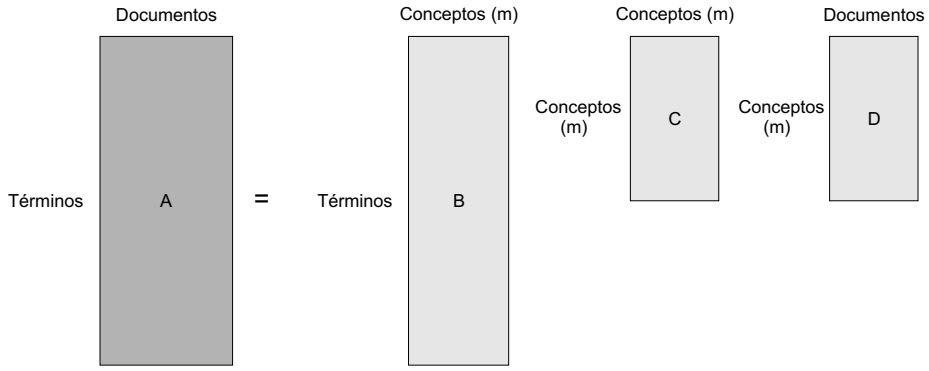

Figura 3.3: Descomposición de matrices mediante la técnica SVD

- Una vez completado el procesado de la matriz A, las tres matrices resultantes pueden ser utilizadas de una forma inmediata para recuperar los documentos que interesan al usuario. Así, dados los términos que definen su consulta, es posible descubrir los conceptos a los que éstos se refieren, empleando la matriz B representada en la Fig. 3.3. Por su parte, C permite descubrir con qué otros conceptos están relacionados los obtenidos en el paso previo y, finalmente, D facilita la recuperación de los documentos en los que esa relación de conceptos están presentes. Este proceso pone de manifiesto las capacidades inferenciales de este tipo de herramientas, en las que a partir de los términos especificados por los usuarios, es posible descubrir documentos y conceptos no necesariamente incluidos en sus consultas.

Los numerosos estudios abordados por diferentes investigadores [21, 187, 129, 20, 72] revelaron algunas de las limitaciones más críticas de la técnica LSI, asociadas a la dificultad de establecer un número óptimo de conceptos (parámetro m) en cada aplicación concreta, y a la degradación de prestaciones obtenida a medida que se incorporan nuevos términos y/o conceptos al proceso de búsqueda y recuperación de información.

A pesar de tales inconvenientes, la LSI ha conducido a resultados muy prometedores en el campo de la búsqueda de información en el dominio de la Web [70, 73, 257]. Además de estos enfoques, basados en el procedimiento descrito antes, se han definido otras aplicaciones también relacionadas en cierto modo con la recuperación de información. Una de las más interesantes incluye la utilización de LSI en sistemas orientados a la recuperación de documentos a partir de consultas especificadas en diferentes idiomas [128]. El objetivo es distinguir los múltiples significados que puede tener una misma palabra (en distintos idiomas) en función del contexto en el que se utilice. Para ello, en primer lugar, se evalúan mediante LSI párrafos textuales enteros, en los que a cada palabra se le asigna un vector dependiente del resto de palabras que aparecen en otros párrafos del documento cerca del analizado. A continuación, se asigna a cada consulta un vector que es la media de los vectores correspondientes a cada una de las palabras presentes en la misma. Los resultados obtenidos en su evaluación muestran que, para una misma consulta en diferentes idiomas (inglés y francés), el enfoque recupera los mismos documentos el 30 % de las veces.

Las capacidades de la LSI también han sido aprovechadas en el dominio de la personalización para aliviar la complejidad computacional de los sistemas recomendadores. Así, por ejemplo, en [217] se presenta una herramienta de personalización para comercio electrónico basada en filtrado colaborativo, en la que se aplican las técnicas SVD para reducir la dimensión de las bases de datos utilizadas durante los procesos de predicción del interés de cada usuario, formación de su vecindario y elaboración de las sugerencias finales mostradas al mismo. En el campo de la búsqueda personalizada y la recuperación de información desde la Web, también puede ser destacado el enfoque propuesto en [43], donde se presenta un motor de búsqueda inteligente que, dado un contexto específico al que se asocia la consulta del usuario, filtra los documentos más relevantes para la misma, utilizando la indexación semántica latente para encontrar similitudes entre los términos presentes en dichos documentos y las características contextuales de la consulta.

En el estado del arte, también es posible encontrar enfoques de recuperación de información basados en la noción de *similitud semántica*. Este proceso consiste en comparar los conceptos formalizados en una estructura de conocimiento y cuantificar su parecido (ver Sección 3.3). Su aplicación en el campo de la búsqueda y recuperación de información es directa y se basa en seleccionar aquellos conceptos más similares a los especificados en las consultas de los usuarios.

3.10.2 Recuperación de información basada en similitud semántica

En la literatura se han propuesto enfoques basados en similitud semántica de naturaleza y complejidad muy diversas: desde aproximaciones que comparan los conceptos organizados en una misma taxonomía, hasta propuestas más elaboradas cuyo objetivo es evaluar la similitud entre entidades formalizadas en diferentes bases de conocimiento. Los primeros enfoques empezaron a utilizarse en aplicaciones de recuperación de documentos Web a partir de las consultas de los usuarios; los segundos, más recientes,

se han vinculado al dominio de la Web Semántica.

Coeficientes Dice

Los coeficientes Dice permiten cuantificar la similitud entre dos términos concretos, en dominios representados únicamente mediante vectores de características (ver Sección 2.2.2). En concreto, siendo c_1 y c_2 dos términos incluidos en dichos vectores, su similitud se calcula mediante la Ec. 3.47.

$$sim_{Dice}(c_1, c_2) = \frac{2 \cdot P(c_1, c_2)}{P(c_1) + P(c_2)} \tag{3.47}$$

donde $P(c_1)$ y $P(c_2)$ son las probabilidades de que los vectores de características considerados contengan el término c_1 o c_2, respectivamente; mientras que $P(c_1, c_2)$ identifica la probabilidad de que ambos términos aparezcan juntos en dichos vectores.

Con el propósito de extender el concepto de similitud semántica a modelos de representación de información más complejos que los citados vectores de características, surgen los enfoques de Rada [201] y Lee [132].

Enfoques de Rada y Lee

En concreto, sus métricas de similitud asumen que el dominio de aplicación se modela mediante una red semántica (ver Sección 2.2.5) en la que únicamente se identifican relaciones jerárquicas entre sus nodos. En estos enfoques, la similitud entre dos conceptos dados depende de la distancia existente entre los nodos que los identifican en la jerarquía, de forma que cuanto más corto es el camino que separa ambos conceptos, más similares son éstos.

Enfoque de Ganesan *et al.*

Un enfoque basado en jerarquías similar al anteriormente descrito fue propuesto por Ganesan *et al.* [77]. En éste, se cuantifica la similitud entre dos colecciones de conceptos (nodos) considerando únicamente la estructura y organización de la jerarquía a la que pertenecen. En concreto, los valores de similitud medidos dependen tanto de la profundidad de los conceptos comparados como de la existencia y posición en la taxonomía de un posible antecesor común (LCA) entre ellos. En su trabajo, los autores de [77] concluyen que la similitud entre dos nodos es mayor cuanto mayor es su profundidad (es decir, cuanto más específicos sean dichos nodos), y cuanto más profundo es su LCA, ya que en este caso, ambos nodos están relacionados de una forma muy significativa.

Una forma alternativa de medir la similitud semántica en una estructura jerárquica fue propuesta por Resnik [206].

Enfoque de Resnik

Este autor describió una métrica inspirada en la teoría de la información de Shannon [223]. En su aproximación, Resnik calcula la probabilidad de encontrar en una jerarquía una instancia del concepto c, y la representa como $p(c)$. Además, dados dos

conceptos c_1 y c_2, donde c_1 está *incluido en* c_2, este enfoque establece que $p(c_1) < p(c_2)$. Rescatando la noción de contenido de la información, Resnik considera que la información de un concepto c es $-log[p(c)]$, de forma que cuantas menos instancias de c haya en la jerarquía, más información proporciona este concepto. Reuniendo todas estas ideas resulta la Ec. 3.48, en la que la similitud entre c_1 y c_2 se representa como la entropía del concepto que aporta más información, de entre todos los antecesores de c_1 y c_2 en la taxonomía considerada (representados mediante el conjunto $S(c_1, c_2)$ en la Ec. 3.48).

$$sim_{Resnik}(c_1, c_2) = max_{c \in S(c_1,c_2)}[-log[p(c)]] \tag{3.48}$$

Enfoque de Lin

Lin también propuso un modelo de similitud basado en la teoría de la información [138]. En concreto, el autor utiliza una serie de suposiciones a la hora de evaluar la similitud entre dos conceptos dados y, a diferencia de Resnik, considera en su métrica tanto los atributos que tienen en común los conceptos comparados, como aquellos que los diferencian. Es obvio que cuanto mayor sea el número de atributos comunes y menores las diferencias apreciadas entre dos conceptos, más similares serán éstos.

De acuerdo a la propuesta de Lin, la similitud semántica entre dos conceptos en una taxonomía depende tanto del significado que comparten como del significado que tienen cada uno de ellos. En este tipo de estructuras jerárquicas, el significado compartido por dos conceptos se puede extraer fácilmente a partir de su antecesor común más cercano (LCA), ya que éste corresponde a la clasificación común más específica de ambos conceptos. Una vez identificado el LCA, Lin mide el significado común de los dos conceptos comparados como la cantidad de información necesaria para declarar lo que ambos tienen en común. Por su parte, el significado de cada concepto individual es cuantificado como la cantidad de información requerida para describirlo de una forma completa.

Ambas ideas aparecen reflejadas en la Ec. 3.49, donde $\sigma_s(c_1, c_2)$ es la similitud semántica entre dos conceptos c_1 y c_2, LCA_{c_1,c_2} es el LCA entre ambos conceptos y $Pr(c)$ representa la probabilidad de que una instancia esté clasificada bajo el concepto c en la taxonomía considerada.

$$\sigma_s(c_1, c_2) = \frac{2 \cdot log[Pr(LCA_{c_1,c_2})]}{log[Pr(c_1)] + log[Pr(c_2)]} \tag{3.49}$$

Otros autores se inspiraron en las métricas presentadas en esta sección para proponer nuevos enfoques [108, 41, 210, 31]. A pesar de esta diversidad, es posible identificar una carencia común debida a que todas estas propuestas consideran únicamente las relaciones jerárquicas definidas en una taxonomía, omitiendo la inferencia de relaciones más complejas entre los conceptos comparados.

Similitud semántica entre los conceptos de una misma ontología

Si bien, tal como se ha descrito en los enfoques mencionados hasta ahora, la evaluación de la similitud semántica en taxonomías ha sido objeto de investigación en

muchos trabajos, la cuantificación de esta métrica en grafos genéricos seguía siendo un problema abierto en el estado del arte.

Precisamente es ésta la dirección que explora la propuesta presentada en [141], en la que se asume como modelo de representación del conocimiento el directorio Web ODP (*Open Directory Project*). ODP clasifica millones de documentos Web en una ontología temática, empleada para inferir relaciones semánticas entre los documentos clasificados bajo las distintas categorías definidas en la misma. Dicha ontología puede ser interpretada como un grafo organizado jerárquicamente, donde cada uno de sus nodos representa un tema y contiene el conjunto de objetos que tratan precisamente esta temática.

La ontología ODP es, por esta razón, un modelo de complejidad mayor que las taxonomías consideradas en los enfoques descritos hasta ahora, en las que sólo se establecen relaciones de herencia entre los nodos. Por el contrario, en ODP es posible identificar, además de este tipo de relaciones jerárquicas, enlaces simbólicos y relacionales entre los diferentes nodos del grafo.

Un enlace relacional entre dos objetos significa que ambos representan temas relacionados entre sí (lo que en el dominio de la Web se corresponde con los enlaces *ver también*). Por su parte, un enlace simbólico entre dos nodos significa que los objetos incluidos en uno de ellos también pertenecen al otro, a pesar de no estar contenidos en el mismo. En este sentido, conviene destacar que en ODP un objeto sólo puede estar incluido en uno de los nodos de la ontología. Precisamente uno de los motivos por los cuales los enlaces simbólicos fueron introducidos en ODP, fue evitar la repetición de objetos en la ontología y, aun así, poder clasificarlos bajo distintas perspectivas.

Conscientes de las diferencias estructurales y semánticas existentes entre las taxonomías y las ontologías, los autores de [141] extienden la propuesta de Lin —en la que únicamente se considera la similitud semántica entre conceptos de una estructura jerárquica (ver Ec. 3.49)— y cuantifican el parecido entre dos objetos incluidos en los nodos de la ontología ODP. La principal aportación de esta aproximación es que aprovecha tanto los componentes jerárquicos como los no jerárquicos (enlaces simbólicos y relacionales) durante este proceso de cuantificación de la similitud semántica.

Sin embargo, este trabajo también presenta una seria limitación, ya que, tal como adelantamos antes, únicamente es aplicable en ontologías cuyas instancias pertenecen a una única clase. Es por ello que el enfoque debería ser generalizado para poder considerar bases de conocimiento genéricas, en las que la clasificación múltiple de las instancias esté permitida.

Otro de los enfoques que incluye una métrica de similitud semántica entre los conceptos de una misma ontología aparece descrito en [160]. Este trabajo propone una estrategia de filtrado colaborativo basada en *item* (ver Sección 2.6.3), en el que en lugar de considerar únicamente las clasificaciones de los usuarios del sistema para comparar dos contenidos concretos, se incluyen también las descripciones semánticas de los mismos durante el proceso de cuantificación de su similitud semántica.

El objetivo es predecir el nivel de interés de un usuario activo en relación a un producto objetivo. Para ello, [160] selecciona los N productos incluidos en su perfil personal que son más similares a dicho producto. Tras esta selección, y tal como se describió en la Sección 2.6.3, el enfoque predice el nivel de interés antes citado como una suma ponderada de las clasificaciones que el usuario activo ha asignado en su perfil a los N contenidos vecinos del objetivo. Los pesos utilizados son los valores de similitud cuantificados entre dichos contenidos y el producto objetivo.

En la métrica de similitud definida en [160] es posible diferenciar dos contribuciones, tal como muestra la Ec. 3.50:

$$CombinedSim(i_p, i_q) = (1 - \alpha) \cdot RateSim(i_p, i_q) + \alpha \cdot SemSim(i_p, i_q) \quad (3.50)$$

- Similitud basada en clasificaciones ($RateSim(i_p, i_q)$): Ésta es la única componente que contemplan los enfoques colaborativos convencionales basados en *item*, en los que i_p e i_q son similares si aparecen simultáneamente en muchos de los perfiles del sistema.
- Similitud semántica ($SemSim(i_p, i_q)$): A la hora de comparar dos contenidos específicos, esta componente se basa en los principales atributos semánticos de los mismos, extraídos desde una ontología de conocimiento que conceptualiza el dominio de aplicación del sistema recomendador.

Una de las limitaciones más importantes de la métrica de similitud semántica propuesta en [160] está relacionada con el esquema de representación de conocimiento que utiliza. Este modelo incluye únicamente un conjunto de clases (cuyas instancias son los productos considerados durante el proceso de recomendación), así como una relación de los principales atributos semánticos de las mismas. De ello se desprende que este enfoque no explota toda la estructura de la ontología de dominio subyacente, obviando la existencia de relaciones más complejas tanto entre las clases de la conceptualización (herencia), como entre sus instancias (por ejemplo, hermandad entre los atributos de los productos considerados).

Precisamente el hecho de prescindir de parte del conocimiento representado en la ontología del sistema, limita de forma considerable las capacidades de razonamiento semántico del enfoque propuesto en [160], impidiendo la inferencia —y posterior aplicación en el proceso de cuantificación de la similitud semántica— tanto de las relaciones antes mencionadas, como de otras más complejas extraídas desde las mismas.

En lugar de inferir este tipo de relaciones semánticas sirviéndose de los atributos considerados en el enfoque, el trabajo presentado en [160] propone construir una matriz semántica S a partir de tales atributos. Las filas de S identifican cada uno de los n productos considerados durante el proceso de recomendación, mientras que sus columnas representan los posibles valores de los atributos semánticos de estos productos. Tal como describe [160], con el objetivo de aliviar los costes computacionales del algoritmo, este enfoque aplica las técnicas LSI descritas en la Sección 3.10.1 para reducir la dimensión de la matriz original S. Este proceso de reducción consiste en representar los contenidos iniciales mediante un conjunto reducido de variables latentes, que sustituyen a todos aquellos atributos semánticos originales que están altamente (cor)relacionados entre sí. Gracias a dichas variables latentes, el enfoque obtiene una matriz S' de dimensión inferior a la de S, en la que se representan productos y variables en lugar de productos y atributos semánticos. Una vez calculada la matriz S', este enfoque aplica el procedimiento matemático descrito en [160] y obtiene una matriz cuadrada de dimensión $n \times n$, en la que una entrada i, j corresponde al valor de la similitud semántica entre los contenidos i_i e i_j, incluida en la Ec. 3.50.

Otro de los enfoques que explota los beneficios de las ontologías de conocimiento en un sistema con requisitos de personalización aparece descrito en [86]. Este trabajo incluye una descripción de Bibster, un sistema P2P dirigido a aquellos investigadores

que están dispuestos a compartir sus metadatos bibliográficos, almacenados localmente en repositorios de conocimiento. Este sistema permite que los usuarios accedan, de forma personalizada, a los metadatos disponibles en la red P2P (típicamente ficheros BibTeX). Para este propósito, [86] emplea dos ontologías como fuentes de conocimiento del sistema: (i) la ontología SWRC (*Semantic Web Research Comunity*) que, como su nombre indica, modela una comunidad investigadora incluyendo, entre otros conceptos, investigadores, publicaciones, conferencias y áreas de conocimiento; y (ii) la jerarquía ACM, en la que se clasifican hasta 1287 categorías temáticas diferentes vinculadas al dominio de las ciencias de la computación.

Con el fin de poder seleccionar, de entre toda la información disponible en la red P2P, los metadatos que interesan a cada usuario, este enfoque propone una métrica de similitud semántica que valora el parecido entre publicaciones diferentes. Es decir, Bibster accede a las publicaciones que interesan a un usuario dado (cuyos metadatos bibliográficos están almacenados en su repositorio local) y las compara con los trabajos consultados por otros usuarios de la red. La métrica utilizada combina diferentes contribuciones, finalmente reunidas en una función de similitud global (basada en la suma ponderada de dichas contribuciones individuales):

- *Similitud sintáctica*: Esta componente cuantifica la similitud entre dos cadenas de texto (e.g. nombres de investigadores, títulos de publicaciones, etc.).

- *Similitud basada en grafo*: Esta contribución compara dos publicaciones dadas considerando únicamente el conjunto de autores asociados a cada una de ellas. Así, esta componente es máxima cuando los investigadores que crearon ambos trabajos son exactamente los mismos. De ello se desprende que esta métrica únicamente considera una de las propiedades explícitas de la ontología SWRC (la que identifica el autor de una publicación), en lugar de analizar el resto de atributos semánticos formalizados en la base de conocimiento.

- *Similitud jerárquica*: Esta componente adopta el enfoque propuesto por Rada *et al.* para así poder valorar el parecido entre las temáticas de la clasificación ACM presentes en las dos publicaciones comparadas. Como consecuencia, aquellos conceptos (temáticas) que ocupan posiciones cercanas a la raíz de la jerarquía ACM son semánticamente menos similares que los que pertenecen a niveles inferiores en la estructura, precisamente por ser estos últimos más específicos.

- *Conocimiento específico del dominio de aplicación del sistema*: En esta componente se consideran características específicas de los metadatos bibliográficos utilizados en Bibster, como el tipo de entrada BibTex (*misc*, *article*, *book*, etc.) o el año de publicación del artículo.

A modo de conclusión, destacar que la métrica de similitud propuesta en [86] no explora durante el proceso de cuantificación todo el conocimiento disponible en las ontologías del sistema. De hecho, de las componentes que acabamos de describir se desprende que el enfoque se basa principalmente en las relaciones de herencia y en algunas propiedades explícitamente representadas en la conceptualización, omitiendo la inferencia de relaciones implícitas más complejas. Esta misma limitación se puede extender al resto de propuestas revisadas en este apartado.

Similitud semántica basada en la definición de plantillas

Un enfoque interesante que también considera la noción de similitud semántica fue propuesto en el dominio del proyecto SemDis, descrito en la Sección 3.5. Este trabajo desarrolla un sistema llamado TRAKS (*Terrorism Related Assessment using Knowledge Similarity*) orientado al descubrimiento de actividades delictivas de diversa naturaleza [3].

En esta herramienta, el usuario debe modelar, mediante una plantilla, las entidades y relaciones que interesan en su consulta, de forma que el objetivo de TRAKS es descubrir en una base de conocimiento otros escenarios (semánticamente) similares a los descritos en ésta. Su aplicación en el dominio de la seguridad nacional es especialmente útil. Así, por ejemplo, un analista especializado en actividades delictivas podría describir las entidades y relaciones comunes en un determinado contexto (blanqueo de dinero, terrorismo, etc.), permitiendo que TRAKS identifique, de forma automática, escenarios similares a los descritos en esa plantilla, representados en bases de conocimiento en las que las consultas manuales serían impracticables.

A la hora de cuantificar la similitud semántica entre las entidades y relaciones de ontología considerada y las modeladas en la consulta del usuario, este enfoque emplea únicamente las relaciones de herencia definidas en la conceptualización. Así, una entidad es recuperada si: (i) pertenece a la misma clase que las definidas en la plantilla, o si (ii) existe un antecesor común en la jerarquía entre dicha clase y las especificadas en la consulta.

Los resultados mostrados al usuario —un conjunto de entidades interconectadas en la ontología de una forma (semánticamente) similar a la definida en la plantilla inicial—, son clasificados según su relevancia, utilizando como criterio precisamente la distancia jerárquica entre las entidades (y propiedades) definidas en la consulta y las recuperadas por TRAKS. Obviamente, el enfoque considera que aquellos elementos más cercanos en la taxonomía a los especificados en la consulta del usuario, son más representativos y relevantes para la misma, de ahí que ocupen las posiciones preferentes en la clasificación final.

Similitud semántica aplicada en sistemas de búsqueda de habilidades

Siguiendo con la revisión de los enfoques descritos en la literatura basados en la evaluación de la similitud semántica entre los conceptos de una misma ontología, encontramos la propuesta definida en [52]. Este trabajo propone un sistema cuyo objetivo es establecer correspondencias entre las habilidades ofrecidas por algunos usuarios y las demandadas por otros. Por ejemplo, una empresa podría querer contratar a un ingeniero con ciertos conocimientos para poder desarrollar un proyecto concreto. En este escenario, la empresa demanda ciertas habilidades que deben ser ofrecidas por los candidatos al puesto de trabajo.

En concreto, la herramienta soporta dos tipos de perfiles (*Oferta* y *Demanda*), especificados a partir de una misma ontología de conocimiento. El enfoque propuesto aprovecha las capacidades de los razonadores de la DL para asistir a los usuarios durante la definición de sus consultas. Dichos razonadores construyen, de forma automática, los citados perfiles (en función del tipo de petición solicitada) y verifican la consistencia de los mismos, asegurando que el sistema pueda resolver las consultas recibidas.

A la hora de valorar la similitud semántica entre dos perfiles dados (habilidades

ofertadas y habilidades demandadas), esta aproximación permite que las correspondencias entre ambos no sean perfectas, escenario en el que, por ejemplo, sería posible encontrar un perfil *Demanda* idéntico a alguno de los ofrecidos por los usuarios (o viceversa). En su lugar, el sistema emplea sus capacidades de razonamiento lógicas para entregar a los usuarios perfiles que, si bien no son idénticos a los disponibles en la ontología, satisfacen sus consultas sin dejar lugar a posibles inconsistencias. Así, aprovechando los razonamientos llevados a cabo por las herramientas de la DL (ver Sección 3.4.4), esta herramienta puede detectar si dos clases son equivalentes, o si dos instancias definidas en los perfiles comparados pertenecen o no a una misma clase, aunque esta declaración no esté explicitada en la ontología. Además, la herramienta ofrece la posibilidad de informar automáticamente al usuario en el momento en el que aparezca un determinado perfil que se ajuste mejor a su petición, esto es, que sea más similar (semánticamente) a la misma. Una vez descubiertos los perfiles de interés para cada usuario, el sistema debe clasificarlos en un orden adecuado, dependiente de su grado de similitud con los especificados en la consulta.

Esta aproximación de descubrimiento de habilidades mejora algunos de los enfoques previos surgidos en el mismo dominio de aplicación [237, 242], únicamente basados en la evaluación de relaciones jerárquicas. Sin embargo, este enfoque aún podría ser mejorado mediante capacidades inferenciales más sofisticadas que las soportadas en los razonadores DL.

Un enfoque con una funcionalidad muy similar a la propuesta en [52] se describe en [186]. En esta herramienta, el objetivo es localizar servicios Web en función de las capacidades que éstos ofrecen. En este escenario, el sistema valora la similitud semántica entre los servicios Web anunciados por algunos usuarios y los demandados por parte de otros. Ambos tipos de consultas se describen, en este caso, a partir de una ontología de conocimiento conforme al lenguaje DAML. Al igual que en los enfoques previos, la métrica de similitud propuesta en [186] considera únicamente relaciones jerárquicas, de ahí sus limitadas capacidades inferenciales.

Una vez evaluada la similitud semántica entre conceptos formalizados en una misma ontología, los investigadores dieron un paso lógico: generalizar este proceso de cuantificación para poder establecer correspondencias entre los conceptos de dos bases de conocimiento diferentes.

Similitud semántica entre los conceptos de dos taxonomías

En [78] se describe el sistema S-Match, una herramienta que toma dos taxonomías y cuantifica la similitud semántica entre cualquier par de nodos de dichas estructuras. Para ello, esta aproximación asume que dos nodos serán más similares cuanto más fuertes sean las relaciones semánticas inferidas entre los conceptos que éstos representan en cada taxonomía.

A diferencia de los enfoques que consideran una única conceptualización —basados en el *principio de inclusión*— este trabajo infiere varios tipos de relaciones entre los nodos, utilizando el conocimiento formalizado en las dos taxonomías comparadas. De acuerdo a dichas relaciones dos conceptos pueden ser equivalentes, uno más genérico (o específico) que el otro, disjuntos, o bien, puede haber cierto solapamiento semántico entre ambos sin llegar a ser equivalentes.

La principal limitación de este trabajo está relacionada con el tipo de bases de

conocimiento sobre las que puede ser aplicado: taxonomías en las que no es posible establecer relaciones de diversa naturaleza definidas a través de propiedades. Precisamente estas últimas podrían ser aprovechadas durante el proceso de cuantificación de similitud semántica propuesto en [78], permitiendo inferir relaciones complejas entre los nodos comparados. La extensión de esta aproximación y su aplicación en ontologías genéricas sigue siendo un problema abierto en el estado del arte.

En el dominio de la Web Semántica, también se han definido algunas (pocas) propuestas que combinan la búsqueda y recuperación de información desde ontologías con ciertos requisitos de personalización. Este reducido número de enfoques contrasta con la profusa cantidad de sistemas recomendadores definidos en la Web para aplicaciones muy diversas, tal como se ha descrito en el capítulo anterior.

3.10.3 Sistemas recomendadores

A pesar de la mencionada falta de diversidad, las herramientas de personalización propuestas para la Web Semántica son contribuciones relevantes que combinan las técnicas de modelado de usuarios y las estrategias de recomendación presentadas en el capítulo 2, con las capacidades de razonamiento automático revisadas en las secciones previas.

Enfoque basado en ontologías

El primer enfoque de personalización surgido en el dominio de la Web Semántica fue propuesto por Stuart Middleton en su tesis doctoral [149]. El autor propone un sistema llamado Quickstep que recomienda artículos a los investigadores, en función de sus áreas de interés y conocimiento. Para ello, el enfoque describe una estrategia híbrida que combina los métodos basados en contenido y el filtrado colaborativo.

Los resultados de este trabajo revelan efectos sinergéticos muy interesantes entre los sistemas recomendadores y las ontologías, manifestados en incrementos considerables tanto de la retirada como de la precisión de las sugerencias ofrecidas, con respecto a otras propuestas existentes que no consideran este tipo de bases de conocimiento.

En concreto, la ontología utilizada por Middleton modela el dominio académico, representando información (extraída de las bases de datos de diferentes departamentos de la Universidad de Southampton) sobre investigadores, publicaciones, proyectos, conferencias, intereses de investigación, etc. La piedra angular de dicha ontología es una jerarquía en la que se organizan las categorías temáticas (clases) sobre las que pueden versar las publicaciones (instancias) de los investigadores. Este enfoque emplea dicha ontología para extraer conocimiento sobre las preferencias de los usuarios y construir sus perfiles personales.

Dichos perfiles son modelados en QuickStep como vectores que almacenan las categorías temáticas de la jerarquía que son relevantes para el usuario, junto con su nivel de interés concreto. Con el propósito de considerar únicamente las preferencias más recientes del usuario, el enfoque establece que el interés de una categoría sea inversamente proporcional al tiempo transcurrido desde que éste consultó el artículo en el que aparece dicha categoría. Para poder actualizar los niveles de todas las clases definidas en el perfil, cada categoría propaga el 50 % de su valor a su supercategoría directa en la jerarquía.

Las categorías incluidas en los perfiles deben ser inferidas por el sistema a partir de los artículos que el usuario haya consultado recientemente en la Web. Para ello, Middleton representa cada publicación como un vector de los términos que aparecen en la misma, ponderados de acuerdo a su frecuencia de aparición[6]. Así, para identificar las categorías relevantes de un documento, el enfoque calcula la correlación Pearson-r (ver Ec. 2.2) entre dichos vectores y las categorías formalizadas en la jerarquía de la ontología. De esta forma, aquellas clases que acumulen valores de correlación más elevados son las incluidas en el perfil del usuario. Este enfoque ofrece la posibilidad al usuario de acceder a su perfil, pudiendo, por ejemplo, modificar las categorías asignadas por el sistema a cada uno de los artículos consultados en el pasado, en caso de que la clasificación automática no le parezca apropiada.

Una vez inferidas las preferencias de un usuario, el sistema Quickstep simplemente aplica una técnica de filtrado basada en contenido y recomienda aquellas publicaciones clasificadas bajo las categorías ontológicas que más le interesan.

Con el objetivo de evidenciar la efectividad del modelado de usuarios a partir de una ontología, Middleton realizó varias evaluaciones experimentales. En éstas, el autor compara su enfoque ontológico frente a propuestas que modelan los perfiles de los usuarios como simples listas planas, sin organización jerárquica, en las que se identifican sus intereses y preferencias. Los resultados de su evaluación revelan incrementos tanto en la retirada como en la precisión de las sugerencias ofrecidas a los usuarios, gracias a la combinación de las ontologías y los sistemas recomendadores.

Además del modelado de usuarios basado en ontologías, Middleton contribuyó de forma relevante al estado del arte de los recomendadores de la Web Semántica, proponiendo un mecanismo para aliviar algunas de las limitaciones asociadas a las técnicas de personalización existentes. En concreto, dichas limitaciones están relacionadas con la llegada de un nuevo usuario al sistema (el llamado *new user ramp-up* descrito en la Sección 2.6.3) y con las primeras etapas de funcionamiento del recomendador (*cold-start*).

En el primer escenario, las preferencias de un usuario que acaba de registrarse en el sistema son, típicamente, muy imprecisas y reducidas, de ahí que el sistema no disponga de conocimiento suficiente para asegurar recomendaciones de alta calidad. En el segundo de los casos, son los sistemas colaborativos los que no disponen de clasificaciones en base a las cuales seleccionar sugerencias personalizadas. Este problema es especialmente crítico: por una parte, los usuarios no confían en el sistema debido a que sus recomendaciones son muy imprecisas y, como consecuencia, el recomendador no registra un número suficientemente elevado de usuarios, de forma que no es posible mejorar la calidad de sus resultados.

Para combatir tales limitaciones, Middleton aprovecha las capacidades de representación de conocimiento de las ontologías. De hecho, el recomendador utiliza la ontología del sistema para consultar las publicaciones de todos los usuarios registrados en la misma. Este tipo de información es especialmente útil, dado que los intereses de un usuario concreto pueden ser inferidos fácilmente a partir de los temas tratados en sus publicaciones. De esta forma, para poder modelar los intereses personales de cada usuario, el sistema calcula la correlación entre sus publicaciones y los artículos clasificados en la ontología. Aquellas categorías ontológicas que estén definidas en ambos vectores, identifican las preferencias del usuario. Gracias al conocimiento aportado por

[6]Todos aquellos términos que no son muy frecuentes en el artículo o que no aportan información significativa (artículos, pronombres, etc.) son eliminados del mismo.

dichas preferencias, QuickStep dispone de información suficiente para ofrecer sugerencias precisas a los usuarios que vayan llegando al sistema, combatiendo los efectos adversos del citado *cold-start*.

Por otra parte, cuando un nuevo usuario se registra en el sistema, el enfoque propuesto por Middleton consulta sus publicaciones en la ontología académica y construye su perfil personal basándose en las preferencias de otros usuarios con intereses similares (de ahí que sea un método colaborativo). Para ello, el autor recurre a la herramienta ONTOCOPI [173] descrita en la Sección 3.10.1. Recordemos que este sistema permite explorar una ontología y encontrar aquellas instancias que están relacionadas de forma significativa con una dada. Este tipo de capacidades convierten a ONTOCOPI en una herramienta especialmente útil en el dominio del recomendador QuickStep, ya que permite seleccionar de forma automática los vecinos de un usuario concreto. Para ello, Middleton asigna un peso elevado a propiedades como *miembroDe* o *autorDe*, asumiendo que aquellos usuarios que pertenezcan a un mismo grupo de investigación y que hayan firmado juntos varias publicaciones, tendrán intereses similares y, por tanto estarán incluidos en el mismo vecindario.

Este enfoque permite que QuickStep pueda completar las preferencias imprecisas de aquellos usuarios que acaban de llegar al sistema. Para ello, Middleton calcula la correlación entre las preferencias de estos usuarios (disponibles en la ontología) y las de sus vecinos (identificados mediante ONTOCOPI), y añade a su perfil las categorías ontológicas resultantes. Éstas también intervienen en el proceso de recomendación, que consiste en seleccionar artículos clasificados bajo las categorías definidas en dicho perfil.

De lo comentado en esta sección se desprende que Middleton aprovecha las capacidades de las ontologías a la hora de modelar las preferencias de los usuarios, sin embargo no explota el potencial semántico de este tipo de conceptualizaciones durante el proceso de recomendación.

Enfoque basado en taxonomías

Otro de los enfoques que destaca la utilidad de los sistemas recomendadores en el campo de la Web Semántica es el descrito en [259]. Este trabajo propone una estrategia de personalización híbrida, en la que también se combinan los métodos basados en contenido y el filtrado colaborativo. Si bien el enfoque es suficientemente flexible para ser extendido a aplicaciones de diversa naturaleza, en [259] éste se utiliza en el dominio del comercio electrónico para sugerir nuevos libros a los usuarios, en función de sus historiales de compra más recientes.

Esta aproximación combate uno de los problemas asociados a los enfoques colaborativos en el dominio de la Web Semántica. Tal como describimos en la Sección 2.6.3, el elemento clave en los sistemas colaborativos es la formación del vecindario de cada usuario activo. Este proceso típicamente se basa en detectar simples solapamientos entre los productos clasificados por los usuarios comparados, de forma que es necesario que dos perfiles almacenen *exactamente* los mismos productos para así cuantificar similitud entre ellos. Es obvio que a medida que se incrementa el tamaño de la base de datos disponible en el recomendador, será menos probable que dos usuarios clasifiquen los mismos productos, dificultando así el proceso de formación del vecindario de cada uno de ellos.

Con el afán de reducir los efectos adversos de esta limitación, los autores de [259]

proponen representar el conocimiento sobre el dominio de aplicación del sistema y aprovecharlo durante el proceso de selección de los vecinos de cada usuario activo. En concreto, esta aproximación utiliza como base de conocimiento una taxonomía en la que se identifican las diferentes categorías (llamadas descriptores) bajo las que se pueden clasificar los libros, y en la que se imponen las siguientes restricciones:

- En primer lugar, la taxonomía debe tener estructura de árbol, lo que impide que una categoría dada tenga más de una supercategoría en el modelo.
- Por otra parte, los libros (instancias) considerados durante el proceso de recomendación deben estar clasificados necesariamente en los descriptores más específicos de la taxonomía, es decir, aquellos que ocupan los niveles más profundos de la misma.
- Por último, destacar que el enfoque permite que cada libro pueda estar asociado a varios descriptores de la taxonomía, lo que permite modelar con mayor precisión el dominio de aplicación concreto del sistema.

El enfoque aprovecha la estructura jerárquica de la taxonomía para generar solapamientos entre los perfiles de los usuarios, de forma que sea posible detectar que sus preferencias son similares, aun cuando éstos no hayan clasificado exactamente los mismos libros. Para ello, en [259] el perfil de cada usuario se representa mediante un vector, cuyas componentes identifican su nivel de interés en relación a cada una de las categorías definidas en la taxonomía de libros del sistema. El enfoque calcula los niveles correspondientes a los descriptores más específicos, y después los propaga por la jerarquía hasta alcanzar el nodo raíz. Para ello, [259] reparte una puntuación dada (representada como un parámetro de configuración s) entre las categorías definidas en cada perfil. Dado que cada libro clasificado por un usuario puede pertenecer a varias categorías en la taxonomía, su nivel de interés se distribuye uniformemente entre todos sus clasificadores de acuerdo a la Ec. 3.51.

$$int(d_i) = \frac{s}{|P_1| \cdot |D_{l_1}|} \tag{3.51}$$

donde l_1 es uno de los $|P_1|$ libros comprados por el usuario U_1, y d_i uno de los $|D_{l_1}|$ descriptores que dicho libro tiene en la taxonomía del sistema.

Una vez calculados los niveles de interés de los descriptores más específicos de la jerarquía, el enfoque propaga estos valores a las categorías más genéricas de la taxonomía, empleando la Ec. 3.52. Así, siendo d_j una supercategoría de d_i, su nivel de interés se calcula en función de la clasificación asignada a esta última ($int(d_i)$) y del número de hermanos que d_i tenga en la jerarquía ($\#sib(d_i)$).

$$int(d_j) = k \cdot \frac{int(d_i)}{1 + \#sib(d_i)} \tag{3.52}$$

Esta ecuación conduce a niveles de interés mayores para aquellas categorías más cercanas a los últimos niveles de la taxonomía (por ser los más específicos), y cuyas subcategorías tengan un número reducido de hermanos en la misma. El parámetro k permite adaptar la expresión en función de las características (por ejemplo, profundidad, número de categorías, etc.) de la taxonomía empleada.

Una vez construidos los vectores que representan cada perfil, este enfoque simplemente compara las preferencias de los usuarios calculando su correlación Pearson-r (ver Ec. 2.2). Precisamente los M usuarios cuyos niveles de correlación con el usuario activo sean más elevados, se incluyen en su vecindario. A la hora de generar las recomendaciones, el enfoque selecciona los libros que han comprado estos vecinos (y que el usuario activo no conoce), y los ordena de acuerdo a su relevancia para éste.

En virtud de su naturaleza híbrida, la relevancia de un libro concreto depende tanto de la correlación existente entre el usuario objetivo y aquellos vecinos que lo han comprado antes que él (filtrado colaborativo), como de la similitud existente entre dicho libro y los clasificados por el usuario activo en su perfil (métodos basados en contenido).

Obviamente, cuanto más significativos sean los valores de similitud cuantificados tanto en la etapa colaborativa como en la correspondiente al filtrado basado en contenido, más relevante será un determinado libro para el usuario activo y, por tanto, mejor será su posición en la recomendación final.

3.11 Sumario

En este capítulo, se han revisado los diferentes enfoques propuestos en la literatura en los campos de la representación y la organización del conocimiento, destacando la adopción generalizada de las ontologías en el dominio de la Web Semántica. Tras describir su arquitectura y presentar los formatos normalizados utilizados para la implementación de este tipo de formalizaciones, se han revisado diferentes propuestas en el ámbito del razonamiento semántico. En este campo, podemos establecer una doble clasificación atendiendo a la capacidad de razonamiento soportada en los enfoques existentes.

3.11.1 Razonamientos semánticos básicos

A este grupo pertenecen los razonadores lógicos vinculados al dominio de la DL y los lenguajes de consulta de ontologías propuestos en el contexto de la Web Semántica.

Razonadores DL

Estas herramientas realizan inferencias lógicas automáticas que permiten comprobar la consistencia de la ontología, clasificar automáticamente sus instancias o descubrir relaciones jerárquicas a partir de las ya formalizadas en la misma. Sin embargo, sus razonamientos no conducen a relaciones de naturaleza compleja ocultas en la ontología, que permitirían descubrir conocimiento útil en el dominio del sistema.

Lenguajes de consulta de ontologías

Los paradigmas de consulta considerados en los enfoques existentes permiten acceder a los conceptos y relaciones explícitamente representadas en la ontología, omitiendo la inferencia de relaciones complejas como las mencionadas anteriormente. Además, la utilización de estos lenguajes obliga al usuario a conocer en detalle la ontología de dominio del sistema para así poder formular correctamente sus consultas, en las

que debe identificar las entidades (clases e instancias) y propiedades formalizadas en la base de conocimiento. Tal requisito limita la utilidad de estos enfoques en algunos dominios de aplicación, precisamente por ser una exigencia abusiva para el usuario.

3.11.2 Razonamientos semánticos avanzados

La inferencia de relaciones complejas como las mencionadas antes ya ha sido materia de estudio en el ámbito de la Web Semántica. Junto a estas relaciones, acuñadas bajo el término de asociaciones semánticas complejas en el proyecto de investigación SemDis, se han definido otras propuestas que también permiten aplicar razonamientos avanzados. Nos referimos a las técnicas de propagación de activación (o técnicas SA) y a los enfoques basados en la cuantificación de la similitud semántica entre los conceptos de una base de conocimiento.

Asociaciones semánticas complejas

El proyecto de investigación SemDis nace con el objetivo de descubrir en una ontología RDFS, las asociaciones semánticas complejas establecidas entre dos intancias que debe especificar el usuario. Es precisamente este último requisito el que limita la aplicación de su mecanismo de inferencia en diversos dominios de aplicación (concretamente, aquellos en los que se pretenda razonar sobre la semántica sin haber identificado previamente dos instancias de la base de conocimiento). A pesar de tal limitación, el potencial del enfoque radica en su capacidad para inferir conocimiento oculto en una ontología, partiendo de las entidades y relaciones formalizadas en la misma.

Técnicas SA

Son mecanismos que permiten explorar, de forma eficiente, grandes redes de conocimiento formadas por un conjunto de nodos interconectados mediante enlaces. Los primeros tienen una relevancia implícita asociada (su nivel de activación), mientras que a los enlaces les corresponde un peso que cuantifica la fuerza de la relación existente entre los nodos que conecta. Su funcionamiento se basa en identificar un conjunto de nodos en la red y propagar sus niveles de activación a sus vecinos (a través de los enlaces que los unen directamente). Una vez finalizado este proceso de propagación, los mayores niveles corresponden a aquellos nodos que están relacionados de forma significativa con los identificados en un principio.

Al alcanzar nodos que no están directamente unidos con los seleccionados inicialmente, las técnicas SA permiten llevar a cabo razonamientos semánticos avanzados. La principal limitación de las propuestas existentes tiene su origen en el tipo de relaciones identificadas en los enlaces de la red de propagación. Éstos limitan la calidad de los razonamientos semánticos aplicados en las técnicas SA, ya que, al considerar únicamente relaciones muy directas[7] entre los nodos de la red, omiten gran cantidad de conocimiento que podría ser aprovechado durante el proceso de propagación.

[7] Por ejemplo, en el campo de la recuperación de información, se establece una relación entre dos términos cuando éstos aparecen simultáneamente en un mismo documento.

Similitud semántica

Son muchas las propuestas definidas en la literatura, especialmente en el campo de la búsqueda y recuperación de información, que utilizan procesos de razonamiento semántico para cuantificar la similitud (semántica) entre los conceptos de una base de conocimiento. Los primeros enfoques que surgieron cuantificaban la similitud entre los conceptos organizados en una estructura taxonómica, mientras que los más recientes han extendido su alcance hasta los conceptos de una misma ontología e, incluso, de bases de conocimiento diferentes.

En los primeros, la similitud entre dos conceptos dados se basa únicamente en las relaciones jerárquicas establecidas entre ellos en la taxonomía, mientras que los segundos utilizan también las propiedades representadas en la ontología. Sin embargo, ninguna de las propuestas existentes en el estado del arte define una métrica que considere la inferencia de relaciones semánticas complejas entre los conceptos comparados. Esto último provoca que los enfoques actuales desperdicien conocimiento útil sobre los conceptos considerados, a la hora de cuantificar su similitud.

Objetivos de la tesis

Tal como apuntamos en la Sección 1.2, el objetivo final de esta tesis es la incorporación de capacidades de razonamiento semántico en los sistemas recomendadores actuales. A la vista de las limitaciones identificadas en este ámbito, nuestro trabajo adopta los enfoques de razonamiento más avanzados y extiende sus capacidades, tal como se describe a continuación:

- Por su habilidad para descubrir conocimiento oculto en la ontología del sistema, la metodología descrita en esta tesis adopta las **asociaciones semánticas** propuestas en SemDis. Dado que este proyecto no contempla requisitos de personalización durante el proceso de razonamiento semántico, nuestro trabajo debe desarrollar un mecanismo de inferencia propio, capaz de descubrir tales asociaciones a partir de las preferencias de cada usuario. Asimismo, la metodología también debe garantizar la adaptación del razonamiento a medida que dichas preferencias evolucionan en el tiempo, ya que sólo en este caso se preservará el carácter personalizado de las recomendaciones.

  A pesar de sus limitaciones, los razonadores DL y los lenguajes de consulta de ontologías actuales también tienen cabida en nuestra metodología, ya que los primeros permiten complementar el conocimiento inicialmente formalizado en la ontología del sistema, mientras que los segundos posibilitan el acceso a las clases, propiedades e instancias utilizadas durante la inferencia de las asociaciones semánticas.

- El conocimiento aprendido a partir de tales asociaciones debe ser incluido en el proceso de recomendación del sistema. Para ello, nuestro trabajo utilizará **técnicas SA** enriquecidas en las que la propagación de activación no sólo recorre las propiedades y relaciones jerárquicas formalizadas en la ontología de dominio, sino también las asociaciones semánticas inferidas desde la misma.

- Asimismo, nuestra metodología también define una métrica de **similitud semántica** que extiende las propuestas en los enfoques existentes. Ésta explotará los

efectos sinergéticos derivados de la combinación del conocimiento explícito formalizado en la ontología (las propiedades y relaciones jerárquicas), y el conocimiento implícito inferido desde la misma (las asociaciones semánticas).

En los siguientes capítulos, abordamos la descripción de la metodología de razonamiento desarrollada en el trabajo, ilustrando su aplicación en el contexto de la DTV y, más concretamente, en el sistema recomendador AVATAR.

Parte III

Arquitectura del sistema AVATAR

Capítulo 4

Decisiones de diseño

Este capítulo presenta las principales decisiones consideradas durante el diseño de AVATAR, un sistema recomendador para entornos de Televisión Digital. Con el claro propósito de promover un uso generalizado de sus capacidades automáticas de personalización, este sistema adopta tecnologías normalizadas de libre disposición (como TV-Anytime, MHP y OWL). Asimismo, para asegurar una arquitectura abierta y modular, el recomendador AVATAR se compone de múltiples agentes —con funcionalidades perfectamente delimitadas— que cooperan activamente durante el proceso de recomendación.

4.1 Descripción genérica

Dado que nuestra metodología aplica procesos de razonamiento semántico sobre las preferencias de los usuarios y los contenidos de TV disponibles en el sistema recomendador AVATAR, son tres los pilares fundamentales sobre los que se asientan sus bases:

- la ontología que formaliza el conocimiento sobre el dominio de la DTV,
- los perfiles que registran las preferencias de los usuarios, y
- las estrategias de recomendación que, basándose en los razonamientos semánticos, seleccionan las sugerencias personalizadas.

Antes de abundar en los detalles de estos tres componentes (descritos en los capítulos 5, 6, 7 y 8), es necesario definir el contexto de ejecución del sistema para así poder identificar su ubicación en el mismo. Para ello, este capítulo presenta las principales decisiones de diseño adoptadas en AVATAR, revisa las tecnologías que utiliza este recomendador, y describe su arquitectura modular.

La primera de las decisiones de diseño está relacionada precisamente con la organización de la arquitectura del sistema. Tal como muestra la Fig. 4.1, ésta se ha distribuido entre el proveedor que ofrece el servicio de personalización y el STB del usuario, permanentemente conectados a través de un canal de retorno vía RTC, xDSL, cable o similar.

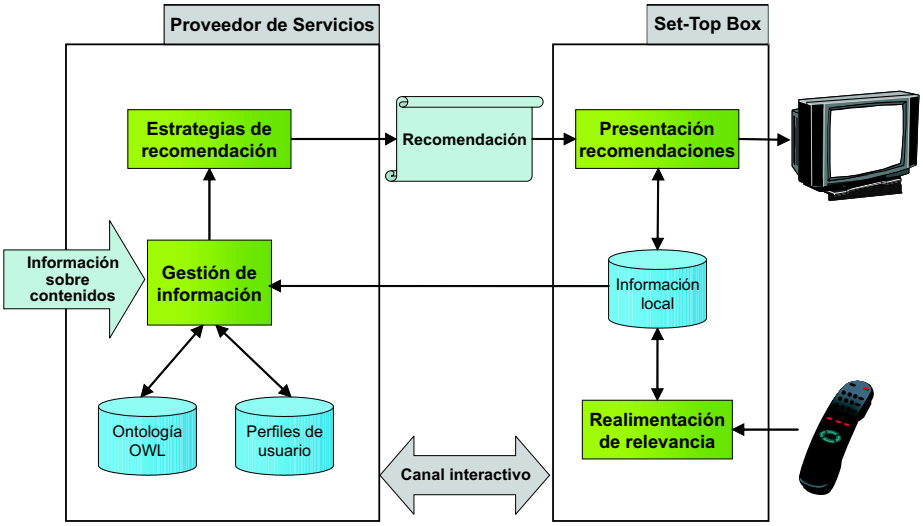

Figura 4.1: Visión genérica de la arquitectura de AVATAR

- En nuestro modelo, el proveedor de servicios mantiene en un servidor tanto los perfiles de los usuarios registrados en AVATAR, como la ontología de conocimiento sobre la que las estrategias de recomendación aplican los procesos de razonamiento semántico. Tales estrategias incorporan esta capacidad inferencial en las técnicas de filtrado de información adoptadas en los recomendadores convencionales. En concreto, nuestro trabajo desarrolla dos estrategias: la primera utiliza los métodos basados en contenido (ver Sección 2.6.2); la segunda es una técnica híbrida en la que dichos métodos se combinan con el filtrado colaborativo (ver Sección 2.6.4).

- En el STB simplemente se presenta al usuario la guía de programación electrónica (EPG) personalizada —en la que se incluyen los contenidos sugeridos por AVATAR— y se monitoriza la información que éste proporciona para poder actualizar su perfil en el proveedor.

Esta distribución responde a dos razones principales. Por una parte, las prestaciones[1] de los receptores actuales son incapaces de atender los costes computaciones asociados al mantenimiento de la ontología del sistema, el modelado de los perfiles de los usuarios y la ejecución de las estrategias de personalización. En su lugar, estas tareas se trasladan al proveedor de servicios, donde las capacidades de almacenamiento y procesado de información de sus servidores son suficientes para garantizar la viabilidad computacional. Por otra parte, la utilización del filtrado colaborativo obliga al sistema a conocer las preferencias de todos los usuarios para elaborar sus sugerencias, de ahí que las estrategias de recomendación deban residir en el proveedor de servicios.

[1] Actualmente, un STB medio dispone de un procesador de 80-130Mhz y hasta 16Mbytes de memoria RAM (fuente: *www.broadcastpapers.com*).

4.2 Tecnologías adoptadas

El segundo de los requisitos considerados durante el diseño de AVATAR es la utilización de tecnologías normalizadas que promuevan un uso generalizado de sus capacidades automáticas de personalización. Para ello, nuestro sistema adopta tres especificaciones que gozan de gran popularidad: MHP, TV-Anytime y OWL.

4.2.1 La especificación MHP

En los últimos años, se han normalizado un gran número de tecnologías vinculadas al dominio de la TV Digital que, aunque desarrolladas de manera aislada, han sido finalmente combinadas tras advertir un nivel de madurez propicio para su integración.

En este contexto, MHP es, sin duda, la especificación más relevante. Tal como apuntamos en la Sección 1.3, este estándar permite la ejecución de aplicaciones interoperables en receptores con hardware y software específico, desarrollado por cualquier fabricante. Para ello, MHP define un contexto de ejecución para las aplicaciones y un interfaz para que éstas puedan acceder a los recursos de cualquier tipo de receptor, desde STBs a televisores digitales o PCs multimedia (ver Fig. 4.2).

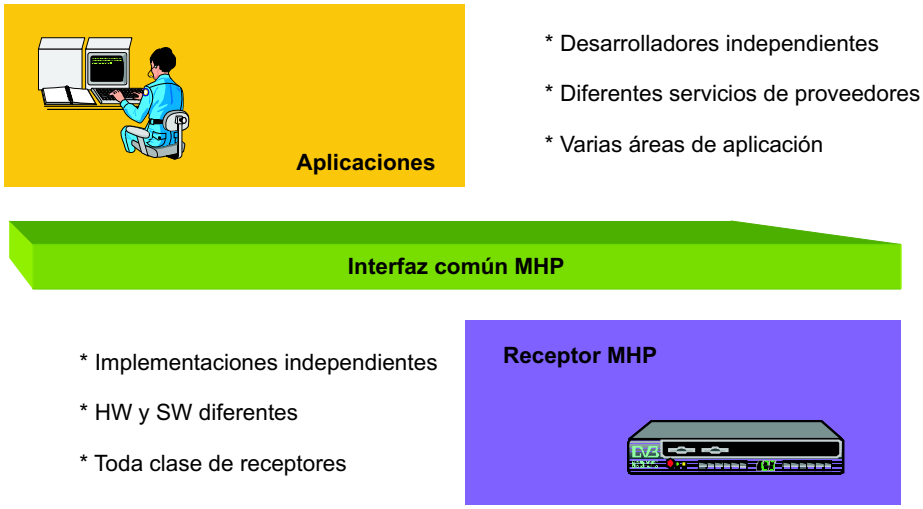

Figura 4.2: Interfaz MHP entre aplicaciones y receptores digitales

El flujo de transporte DVB decodificado en el receptor del usuario debe incluir la señalización oportuna para garantizar tanto la correcta ejecución de las aplicaciones como la posibilidad de interacción con el usuario. El ciclo de vida de estas aplicaciones está supervisado por un gestor que se encarga de iniciar y/o finalizar su ejecución, en respuesta a las órdenes del usuario o a indicaciones insertadas en la información de señalización. Este ciclo de vida permite la gestión eficiente de múltiples aplicaciones que se ejecutan simultáneamente en el receptor digital.

MHP en AVATAR

Habida cuenta de la implantación generalizada de MHP en los receptores actuales dotados de capacidades interactivas, nuestro trabajo establece que la presentación de las recomendaciones personalizadas y la monitorización de la información que proporcionan los usuarios, apuntadas en la Fig. 4.1, se implementen mediante sendas aplicaciones MHP. Este contexto no impide que el usuario descargue y ejecute otras aplicaciones interactivas de diversa naturaleza (e.g. servicios de información, votaciones, apuestas, juegos, etc.), controladas por el gestor que reside en su receptor digital.

En este punto, conviene destacar que si bien en la Fig. 4.1 asumimos que el receptor del usuario es un STB, la generalidad y flexibilidad de la metodología desarrollada permite vincular el servicio de personalización con receptores de diversa naturaleza. Así, por ejemplo, el usuario podría recibir las sugerencias personalizadas de AVATAR fuera de su casa, utilizando un dispositivo portátil (teléfono móvil, PDA, etc.) que suplantaría al STB.

La segunda iniciativa de normalización considerada en nuestro trabajo se centra en las descripciones de los contenidos de TV, utilizadas para formalizar la ontología de dominio del sistema sobre la que se aplica el razonamiento semántico incorporado en la metodología.

4.2.2 La especificación TV-Anytime

TV-Anytime es un estándar cuya primera fase [244] normaliza aplicaciones XML que describen contenidos audiovisuales genéricos, instancias específicas de los mismos, perfiles de usuario, información de segmentación de contenidos, e incluso políticas relacionadas con la protección de los contenidos y el acceso condicional a los mismos. Dichas descripciones son proporcionadas mediante diferentes tipos de metadatos, con el objetivo de poder dotar a los receptores digitales de la capacidad de almacenar contenidos con unas determinas características especificadas por los usuarios. Así, por ejemplo, la norma permite que el usuario pueda encontrar, navegar y gestionar contenidos procedentes de una gran variedad de fuentes internas y externas, incluyendo la difusión tradicional, la TV interactiva, Internet o el almacenamiento local en un PDR (*Personal Digital Recorder*).

Tras la publicación de la fase 1 de TV-Anytime en 2003, las especificaciones de su segunda fase también han sido normalizadas por ETSI en abril de 2006 [245]. Entre sus principales áreas de interés, destacan la compartición y distribución de ficheros, la sincronización de múltiples fuentes de contenidos, el empaquetado de varios tipos de programas dotados, en algunos casos, de capacidades interactivas, y la selección de contenidos adecuados para los usuarios en función de sus preferencias personales. Hoy en día, es esta última área la que despierta mayor interés entre los proveedores de servicios, ya que sienta las bases de nuevos modelos de negocio en el campo del almacenamiento y la entrega de contenidos personalizados, de indudable utilidad para los usuarios de la DTV dada la sobrecarga de información disponible en este dominio.

Son dos las características distintivas de la especificación TV-Anytime: (i) separa totalmente la descripción de los contenidos multimedia de su localización (dada por el canal y la hora de difusión) y (ii) es independiente del protocolo de difusión utilizado. En relación a esto último, destacar que ya se ha propuesto una especificación para transmitir metadatos TV-Anytime sobre flujos de transporte DVB [243].

Referenciación de contenido y resolución de localización

Para poder lograr la independencia entre la descripción de los programas y su localización, TV-Anytime propone un proceso llamado *referenciación de contenido* (*Content Referencing*). El elemento clave en este proceso son los identificadores únicos (llamados CRIDs - *Content Reference IDentifiers*) asignados tanto a contenidos individuales como a grupos de contenidos. Los CRIDs no son más que URIs (*Universal Resource Identifiers*) con el siguiente formato:

```
CRID://<autoridad>/<datos>
```

donde el campo `autoridad` corresponde a un dominio DNS (*Domain Name Service*) y el campo `datos` permite diferenciar los diferentes CRIDs creados por una misma autoridad.

Las autoridades no sólo publican los CRIDs de los contenidos, sino que también se encargan de resolver su localización espacio-temporal concreta. Para ello, tal como muestra la Fig. 4.3, dado el CRID de un programa, las autoridades extraen el localizador DVB que especifica la hora, la fecha y el canal en el que se emite. De esta forma, a la hora de solicitar contenidos concretos, los usuarios únicamente deben especificar sus CRIDs y, en el momento en que éstos estén disponibles, el receptor resuelve su localización para poder conocer la información de emisión asociada a los mismos.

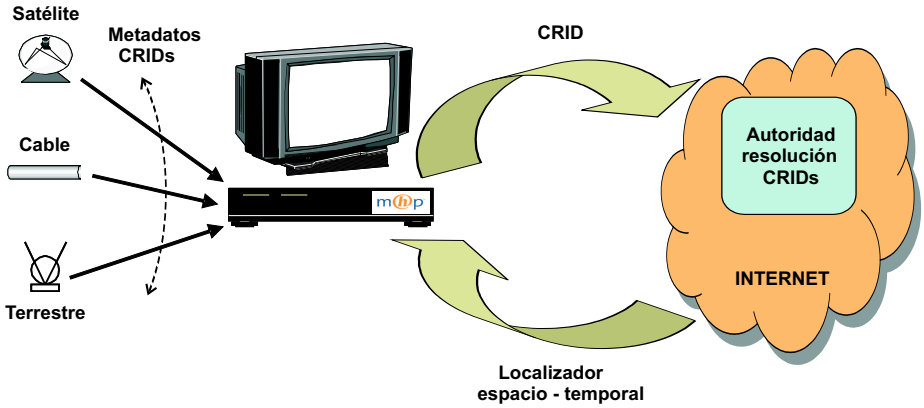

Figura 4.3: Proceso de resolución de localización utilizado en TV-Anytime

Metadatos TV-Anytime

Son cuatro los tipos de metadatos definidos en la especificación TV-Anytime:

- **Metadatos de descripción de contenidos**: Proporcionan características de los contenidos a los que están asociados, como género, idioma, título, resumen, palabras claves, créditos, clasificación por edades, fecha y país en el que se produjeron, etc. Este tipo de datos (junto al CRID de cada programa) son los que deben especificar los usuarios a la hora de solicitar contenidos concretos.

- **Metadatos de descripción de instancias**: Describe instancias concretas de un programa, indicando, entre otros campos, la hora y el canal en el que se emite, la duración, si se transmite en directo o en diferido, si es la primera o la última vez que está disponible, etc.

- **Metadatos sobre el consumidor**: Permiten identificar las preferencias de cada usuario y su historial de visionado, es decir, los programas que ha visto en el pasado. Estos metadatos definen la información necesaria para poder dirigir los servicios ofrecidos a una audiencia específica.

- **Metadatos de segmentación**: Cada flujo audiovisual puede descomponerse en varios segmentos. La información relativa a cada uno de ellos se proporciona mediante estos metadatos. Este mecanismo de segmentación permite que el usuario pueda acceder tanto al contenido completo como a cada uno de los segmentos identificados en el mismo.

TV-Anytime en AVATAR

Tal como apuntamos al comienzo del capítulo, para poder razonar sobre la semántica de los contenidos de TV y las preferencias de los usuarios, la metodología desarrollada debe disponer de una base de conocimiento en la que se formalicen los conceptos y relaciones típicas en el dominio de la DTV.

Dado que la especificación TV-Anytime (y más concretamente los metadatos de descripción de contenidos) proporciona descripciones semánticas muy detalladas de los contenidos audiovisuales, nuestro trabajo adopta esta norma como punto de partida en el proceso de formalización semántica. Desde esta especificación, es posible extraer jerarquías de clases (e instancias específicas) referidas a atributos semánticos de naturaleza muy diversa, así como propiedades que identifican las relaciones existentes entre algunos programas. Tal como describiremos en el Capítulo 5, tales entidades y relaciones son la piedra angular del proceso de razonamiento en el que se basan las recomendaciones elaboradas en AVATAR.

4.2.3 La especificación OWL

Para representar en la ontología del sistema la semántica de los contenidos de TV, es necesario elegir un lenguaje de implementación concreto. Nuestro trabajo opta por OWL, un formato cuya capacidad expresiva (ver Sección 3.4.3) no sólo permite formalizar —mediante clases, propiedades e instancias— los conceptos y relaciones extraídos de los metadatos TV-Anytime, sino también procesarlos posteriormente mediante las estrategias de razonamiento semántico incorporados en la metodología.

Este procesado incorpora tanto razonamientos lógicos sencillos —aplicados sobre la información geográfica y temporal asociada a los contenidos de TV—, como otros de naturaleza más compleja, cuyo objetivo es inferir relaciones ocultas entre los contenidos formalizados en la ontología OWL. Tal como describiremos en los capítulos 7 y 8, tales relaciones permiten completar el conocimiento disponible sobre las preferencias de los usuarios y, por ende, favorecen procesos de personalización más precisos que los aplicados en los enfoques actuales.

4.3 Arquitectura del sistema AVATAR

La modularidad es también uno de los requisitos perseguidos durante el diseño de la arquitectura de AVATAR. Dicha característica facilita no sólo la incorporación de nuevas estrategias de recomendación en el sistema, sino también la adopción de estándares futuros que, junto a MHP, TV-Anytime y OWL, promuevan la reutilización de sus capacidades de personalización. Para lograr una arquitectura abierta y modular, tanto las aplicaciones MHP que se ejecutan en el receptor del usuario como el servidor que implementa el servicio de recomendación, albergan un conjunto de agentes software con roles perfectamente delimitados, que cooperan entre sí en la tarea de la personalización.

Con el objetivo de concretar la visión genérica de la arquitectura del sistema mostrada en la Fig. 4.1, en esta sección describimos tanto la funcionalidad de cada uno de los agentes incluidos en la misma, como la información que intercambian durante sus interacciones.

4.3.1 Set-Top Box del usuario

Tal como se aprecia en la Fig. 4.4, las aplicaciones ejecutadas en el receptor del usuario incluyen tres agentes:

- **Agente de registro**: Como su nombre indica, la tarea encomendada a este agente está relacionada con la fase de registro del usuario en el sistema. En AVATAR, cada espectador está identificado de forma unívoca mediante una referencia numérica (en adelante, ID), asignada en el servidor del proveedor de servicios y utilizada para indexar la base de datos que almacena los perfiles de todos los usuarios registrados. Junto a dicho ID, el agente de registro recibe la información necesaria para que el usuario pueda seleccionar un conjunto de preferencias con las que inicializar su perfil personal; dicha información incluye una relación de géneros y contenidos de TV específicos.

- **Agente de monitorización del usuario**: Este agente registra las acciones llevadas a cabo por los usuarios, incluyendo los programas que seleccionan como preferencias iniciales, los contenidos que aceptan o rechazan cuando el sistema les ofrece una recomendación, etc. De esta forma, el recomendador considera preferencias permanentemente actualizadas, lo que redunda en la calidad de las sugerencias ofrecidas y en la confianza de los usuarios.

- **Agente gestor de la comunicación entre usuario y proveedor**: Tal como muestra la Fig. 4.4, esta agente atiende las solicitudes de los dos anteriores:

    - Por una parte, solicita al servidor la información requerida durante la fase de registro de los usuarios en el sistema.
    - Por otro lado, una vez que el usuario ha seleccionado sus preferencias iniciales, este agente envía dicha información al servidor del proveedor para proceder a la creación de su perfil. En este sentido, conviene destacar la posibilidad de que un usuario especifique como preferencias iniciales ciertos contenidos que no están incluidos en la información mostrada por el agente de registro. En este caso, el agente que gestiona la comunicación entre el usuario y el proveedor, envía al proveedor la relación de los CRIDs que

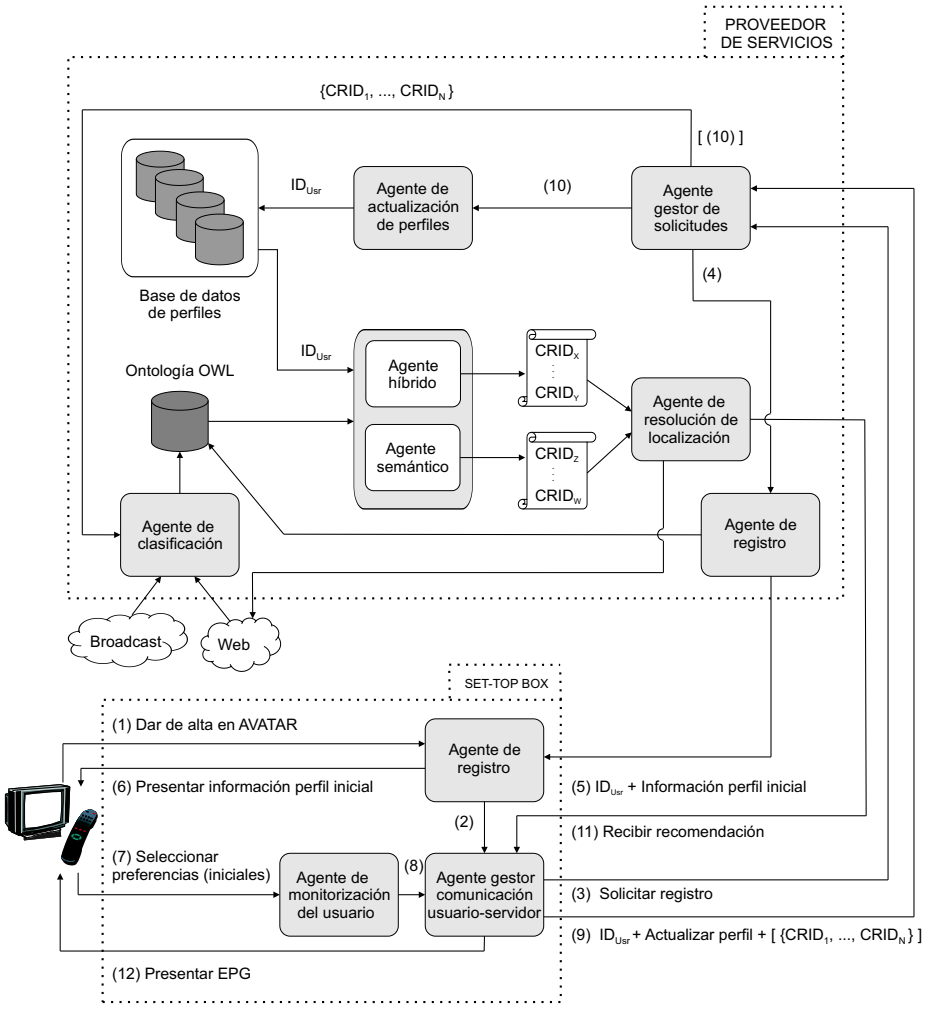

Figura 4.4: Arquitectura del sistema AVATAR

identifican dichos contenidos, para que éstos puedan ser incorporados tanto a la ontología del sistema como al perfil inicial del usuario. La descripción de este proceso se retomará en el Capítulo 6.

- Por último, este agente también se encarga de recibir la lista de programas sugeridos a cada usuario, incluidos, junto a la hora, fecha y canal de emisión, en la EPG personalizada mostrada al mismo.

4.3.2 Servidor del proveedor de servicios

Los agentes que implementan el servicio de personalización en el proveedor son:

- **Agente de registro**: Este agente se encarga de proporcionar a su homónimo en el receptor del usuario la información requerida en la fase de registro de los usuarios en el sistema: el ID único de cada espectador y una relación de géneros y contenidos específicos, extraída de la ontología de conocimiento del sistema (ver Fig. 4.4).

- **Agente de clasificación**: Dado que la oferta de contenidos de TV varía con mucha frecuencia, nuestro sistema debe actualizar la relación de programas representados en la base de conocimiento para, de esta forma, ofrecer recomendaciones novedosas a los usuarios.

  Es el agente de clasificación quien se encarga de este proceso de actualización. Para ello, tal como muestra la Fig. 4.4, este agente explora fuentes de diversa naturaleza (difusión, Web, etc.) e incorpora a la base de conocimiento OWL los nuevos programas y descripciones semánticas disponibles en las mismas.

  Para acometer tal tarea, el agente de clasificación debe identificar las clases de la ontología bajo las que debe clasificar cada contenido y sus respectivos atributos TV-Anytime, e incluirlos como instancias de las mismas. Tal como describiremos en el siguiente capítulo, este agente también se encarga de añadir a la base de conocimiento OWL nuevas relaciones (propiedades) inferidas mediante razonamientos (geográficos y temporales) automáticos.

- **Agente gestor de solicitudes**: De acuerdo a lo explicado previamente, son tres las solicitudes que puede recibir este agente:

  - Como respuesta a la primera de ellas, debe avisar al agente de registro para que éste envíe a su homónimo en el receptor del usuario la información requerida para formalizar su alta en el sistema.
  - En caso necesario, también avisa al agente de clasificación para que incorpore a la ontología OWL aquellos contenidos que el usuario haya declarado como preferencias iniciales.
  - Por último, notifica la necesidad de actualizar los perfiles de los usuarios al agente encargado de tal tarea, tanto en la fase de registro del usuario en el sistema como cuando éste proporciona información de relevancia sobre una recomendación concreta.

- **Agente de actualización de perfiles**: Este agente indexa la base de datos que almacena las preferencias de todos los usuarios y, mediante su ID específico, accede al perfil personal de cada espectador. Sus tareas incluyen la incorporación

de nueva información sobre las preferencias de los usuarios, la eliminación de aquellas que se hayan quedado obsoletas, y la actualización tanto de los niveles de interés de los espectadores como de la información que facilita la adaptación del sistema ante los cambios ocurridos en dichas preferencias.

- **Agentes recomendadores**: Se encargan de seleccionar una relación de contenidos personalizados para cada usuario, basándose en sus preferencias y en el conocimiento formalizado en la ontología de TV. Tal como muestra la Fig. 4.4, nuestro enfoque incluye un agente para cada una de las estrategias implementadas en la metodología de razonamiento semántico desarrollada.

  - **Agente semántico**: Tal como describe [27], este agente, que utiliza los métodos basados en contenido como técnica de filtrado de información, sugiere al usuario contenidos de TV que están asociados semánticamente con aquellos que le interesaron en el pasado. Esta estrategia se ejecuta de forma automática cuando varían las preferencias del usuario, ofreciendo con ello, sugerencias permanentemente actualizadas.
  - **Agente híbrido**: Tal como se describe en [29], el objetivo del agente híbrido es identificar los usuarios del sistema a los que podrían interesar un conjunto de contenidos previamente seleccionados. En este proceso, el sistema utiliza una estrategia híbrida que combina los métodos basados en contenido, el filtrado colaborativo y las capacidades de inferencia semántica de la metodología desarrollada en este trabajo.

- **Agente de resolución de localización**: Una vez que los agentes recomendadores han identificado los contenidos que deben ser sugeridos a los usuarios, es necesario localizar la hora, la fecha y el canal en el que éstos se emiten. Dado que esta información no está representada en la ontología OWL, es el agente de resolución de localización quien obtiene los localizadores espacio-temporales de los contenidos recomendados a partir de sus respectivos CRIDs (ver Fig. 4.3).

  Junto a estos datos, el agente incluye la sinopsis de cada programa, así como una breve explicación textual en la que se describen las relaciones semánticas en las que se basó el sistema para seleccionar los programas sugeridos a partir de las preferencias del usuario. Con esta información se construye la EPG personalizada que el agente que gestiona la comunicación entre el usuario y el proveedor muestra finalmente al espectador.

Una vez presentada la arquitectura de AVATAR, los siguientes capítulos de la memoria abundan en los agentes más relevantes en la metodología de razonamiento semántico desarrollada. Comenzamos describiendo el agente que actualiza la base de conocimiento OWL del sistema (Capítulo 5); a continuación, nos centramos en el agente que modela las preferencias de los usuarios (Capítulo 6) y en aquellos que, basándose en sus perfiles, ejecutan las dos estrategias de recomendación desarrolladas en este trabajo (Capítulos 7 y 8, respectivamente).

Capítulo 5

Una ontología sobre el dominio de la TV

Este capítulo presenta la ontología de dominio utilizada en el sistema AVATAR. Una vez justificada la elección del lenguaje de implementación adoptado en la metodología, se describe la información formalizada en la base de conocimiento del recomendador, extraída a partir de la especificación de metadatos TV-Anytime. Por último, el capítulo introduce las dos ontologías que utiliza el sistema para representar la información geográfica y temporal asociada a los contenidos, y describe los razonamientos automáticos aplicados sobre la misma.

5.1 Lenguaje de implementación

Tal como se describió en la Sección 3.2.4, la expresividad de los lenguajes de implementación de ontologías está estrechamente relacionada con el tipo de razonamiento que puede ser llevado a cabo sobre las mismas. En este sentido, la riqueza expresiva de OWL —que permite formalizar conceptos y relaciones no soportados en otros formatos, como RDF y RDFS— ha sido decisiva para elegir este lenguaje de implementación en nuestro metodología.

Más concretamente, de los tres sublenguajes incluidos en OWL —Lite, DL y Full— nos hemos decantado por OWL DL. Esta elección responde a dos razones: en primer lugar, OWL Lite carece de ciertas funcionalidades útiles en el proceso de formalización semántica (e.g. restricciones de cardinalidad genéricas, disyunción de clases, tipos enumerados, etc.), lo que limitaría el conocimiento representado en la ontología; por otra parte, actualmente, no existen razonadores lógicos decidibles para el lenguaje OWL Full, lo que impediría recurrir a estas herramientas para comprobar la consistencia de nuestra ontología de TV.

Tal como se describe en el Apéndice B, la ontología de TV ha sido implementada mediante el editor Protégé [197], cuya interfaz gráfica simplifica en gran medida el proceso de conceptualización semántica. Asimismo, por su compatibilidad con modelos OWL DL y su implementación altamente optimizada, el razonador lógico RACER ha sido incluido como herramienta de soporte en las fases de diseño y desarrollo de la ontología, garantizando, de esta manera, la consistencia de la base de conocimiento del sistema.

5.2 Estructura de la ontología

Para poder razonar sobre la semántica de los contenidos de TV, la ontología OWL debe representar descripciones semánticas muy detalladas de los mismos, incluyendo atributos (en adelante llamados *características semánticas*) como la temática, los créditos que participan en los programas, información geográfica y temporal asociada a ellos, etc.

Tal como describimos en la Sección 4.2.2, la especificación TV-Anytime [244] define este tipo de información semántica mediante los llamados metadatos de descripción de contenido, de ahí que nuestro trabajo se base en tales descripciones a la hora de definir las clases, propiedades e instancias formalizadas en la ontología de dominio de AVATAR. La actualización de esta base de conocimiento es tarea del agente de clasificación del sistema, encargado de incluir nuevos contenidos de TV (y sus atributos semánticos) a medida que estén disponibles.

Conviene destacar que nuestra metodología únicamente formaliza en la ontología OWL aquellos metadatos TV-Anytime que son útiles a la hora de aplicar procesos de razonamiento sobre la semántica de los contenidos de TV y las preferencias de los usuarios; características como la tasa de transmisión, los formatos de los ficheros de audio y vídeo relativos a los contenidos considerados, o las relaciones de aspecto de estos últimos, entre otras, son obviadas durante el proceso de representación de conocimiento.

En concreto, los atributos TV-Anytime formalizados en nuestra conceptualización OWL son:

- CRID: Según lo explicado en la Sección 4.2.2, el CRID permite identificar de forma única cada programa representado en la ontología OWL, propiedad que aprovechará nuestra metodología de razonamiento.
- Título del contenido de TV.
- Palabras clave: Describen las características más destacables del programa.
- Sinopsis: Es una breve descripción textual acerca de cada contenido.
- Género: Este atributo describe la temática de cada programa.
- Idiomas en los que se difunde el programa.
- Formato de producción: Mediante este atributo es posible clasificar los contenidos según su formato estructural (debates, documentales, entrevistas, representaciones teatrales, etc.), con independencia de la temática abordada en los mismos.
- Alertas de contenido: Permiten poner sobre aviso a los usuarios sobre la presencia de determinados tipos de contenidos en los programas difundidos: sexo, violencia, etc.
- Temática relativa a los anuncios publicitarios: La especificación proporciona una clasificación detallada sobre la temática de los contenidos comerciales difundidos en TV.
- Créditos: Este atributo hace referencia a los roles que pueden desempeñar las personas que participan en un programa de TV: directores, actores, productores, etc.

- Fecha y localización de producción: Mantienen información temporal y geográfica sobre la producción del contenido.
- Fecha y localización de creación.
- Fecha y localización relativa al estreno.
- Intención con la que fue creado un programa: Describe la intención del proveedor a la hora de emitir un determinado contenido (informar a los usuarios, entretenerlos, educarlos, etc).
- Audiencia del contenido: La especificación clasifica la audiencia potencial de un programa específico en función de su edad, profesión, nivel cultural, etc.
- Origen técnico: TV-Anytime identifica la fuente técnica de la que procede cada contenido (creado en un estudio, vinculado a la industria del cine, etc).
- Atmósfera: Este atributo describe ciertas características emocionales o psicológicas vinculadas a los contenidos de TV (romanticismo, felicidad, diversión, etc).
- Relación existente con otros contenidos audivisuales: TV-Anytime define diferentes tipos de relaciones entre los contenidos audiovisuales. Tales relaciones se establecen, por ejemplo, entre una película y su tráiler, entre dos episodios de una misma serie, etc.

Para cada uno de los atributos que acabamos de mencionar, la norma TV-Anytime define un conjunto completo de posibles valores, proporcionando así descripciones muy precisas sobre los contenidos considerados. La ontología OWL incorporada en nuestra aproximación ha sido diseñada precisamente a partir de tales valores, facilitando al máximo el proceso de traducción de programas TV-Anytime a clases, instancias y propiedades OWL, abordado por el agente de clasificación del sistema.

En concreto, cada uno de los contenidos considerados en nuestro enfoque —identificado de forma unívoca mediante su CRID— se corresponde con una instancia de la ontología OWL. De acuerdo a lo explicado en la Sección 3.3.4, dichas instancias pertenecen a clases organizadas jerárquicamente. Para definir esta jerarquía de contenidos, nuestra metodología emplea la clasificación de géneros establecida en TV-Anytime, en la que es posible identificar los distintos tipos de programas existentes.

El resto de atributos mencionados anteriormente se utilizan para definir las características semánticas de los programas considerados en el sistema. Los valores de dichos atributos, así como los géneros definidos en la norma, son revisados en las siguientes secciones.

5.2.1 Jerarquía de contenidos de TV

La especificación TV-Anytime establece una clasificación muy detallada en la que se describen los posibles géneros bajo los que se clasifican los contenidos audiovisuales. En concreto, tal como muestra la Tabla 5.1, esta norma define siete categorías muy genéricas, a partir de las cuales es posible identificar otras más específicas. Desde tal estructura, la construcción de la jerarquía de contenidos de TV perseguida en nuestra ontología OWL parece inmediata.

Géneros
Non-fiction
Sports
Leisure
Fiction
Amusement
Music
Interactive games

Tabla 5.1: Clasificación de géneros definidos en TV-Anytime (I)

Cada uno de los géneros TV-Anytime está identificado de forma unívoca mediante un código numérico [244]. A partir de estos códigos resulta muy sencillo inferir la especialización de algunos de los géneros a partir de otros más abstractos. Así, tal como muestra la Tabla 5.2, la norma asigna el código 3.1 a aquellos contenidos que no pertenecen al género ficticio (*Non-fiction*) e identifica mediante los códigos 3.1.1, 3.1.6 y 3.1.7 tres especializaciones de la misma, referidas a noticias (*News*), programas científicos (*Science*) y contenidos de interés general (*Human interest*), respectivamente [1].

Debido a su carácter unívoco, nuestra metodología emplea los códigos numéricos definidos en TV-Anytime para identificar cada una de las clases de la jerarquía a la que pertenecen los contenidos de TV en la ontología OWL implementada. Así, por ejemplo, a partir de los géneros mostrados en la Tabla 5.2, nuestro trabajo define, de manera inmediata, el extracto de la jerarquía de contenidos representado en la Fig. 5.1.

Nuestro enfoque impone las siguientes restricciones sobre la jerarquía de géneros implementada:

- En primer lugar, los contenidos de TV representados en la base de conocimiento OWL pertenecen siempre a las clases hoja de esta jerarquía. Tal como detallaremos en el siguiente capítulo, esta condición permite que el sistema infiera el nivel de interés del usuario en relación al género al que pertenece cada programa, y también en relación a todos aquellos que lo anteceden en la jerarquía definida en la ontología.

- Por otro lado, cada contenido puede pertenecer a una o a varias clases de la Fig. 5.1. Esta condición es especialmente relevante en un sistema basado en razonamiento semántico como el propuesto, ya que la ausencia de la clasificación múltiple podría conducir a una pérdida de precisión en lo que a representación del conocimiento se refiere.

- Por último, destacar que la jerarquía de contenidos de nuestro enfoque debe tener estructura de árbol. De acuerdo a lo explicado en la Sección 3.3.1, tal condición

[1] El lector puede consultar en el Apéndice A (Tabla A.1) la relación completa de los códigos numéricos asignados a cada uno de los géneros contemplados en la especificación TV-Anytime.

Género	Código numérico
NON-FICTION	3.1
News	3.1.1
Daily news	3.1.1.1
Special news	3.1.1.2
...	...
Current affairs	3.1.1.16
...	...
Sciences	3.1.6
Applied sciences	3.1.6.1
Nature sciences	3.1.6.2
...	...
Archaelogy	3.1.6.14
Human interest	3.1.7
Reality	3.1.7.1
Gossip	3.1.7.2
...	...
Religious buildings	3.1.7.7
...	...
Events	3.1.9
Anniversary	3.1.9.1
Fair	3.1.9.2
...	...
Exhibition	3.1.9.5
...	...
FICTION	3.4
General light drama	3.4.1
...	...
Action	3.4.6
Adventure	3.4.6.1
Disaster	3.4.6.2
...	...
Epic	3.4.6.13
...	...
Psychological drama	3.4.19
...	...
INTERACTIVE GAMES	3.7

Tabla 5.2: Clasificación de géneros definidos en TV-Anytime (II)

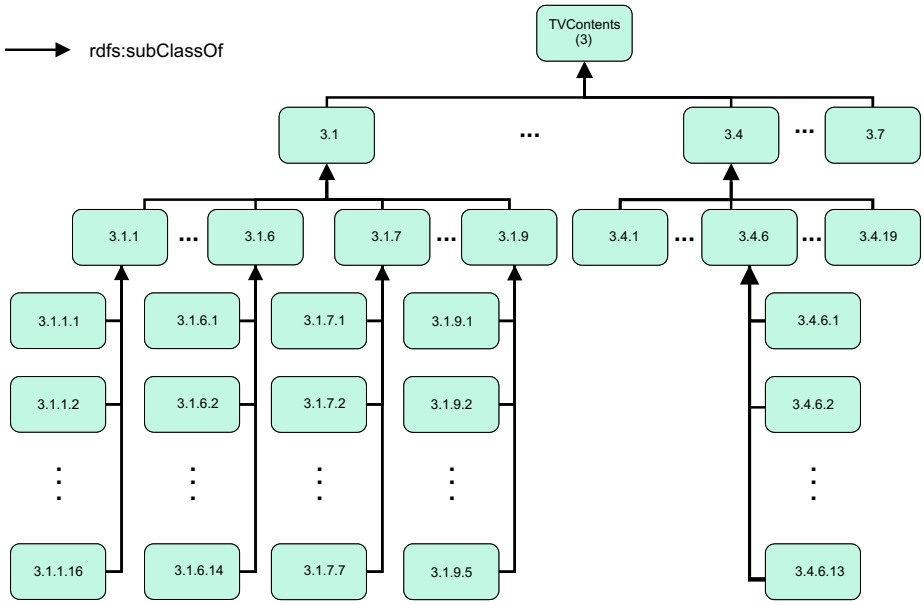

Figura 5.1: Extracto de la jerarquía de géneros definida en la ontología OWL (ver Tabla 5.2)

supone que cada clase tenga una única superclase directa en la estructura jerárquica.

5.2.2 Características semánticas de los contenidos de TV

Una vez clasificados los programas de TV concretos bajo las categorías representadas en la jerarquía de géneros, es necesario describir sus características semánticas. Para poder asociar dichas características a los programas correspondientes, nuestra ontología recurre a los dos tipos de propiedades definidas en el lenguaje OWL DL (ver Sección 3.4.2).

- Un pequeño subconjunto de las mismas son propiedades *Datatype* e identifican atributos como el CRID, el título, su sinopsis y un conjunto de palabras clave. Estas dos últimas características son especialmente importantes en nuestro sistema semántico, ya que permiten inferir conocimiento adicional muy útil sobre los contenidos de TV considerados. En concreto, nuestro enfoque explora ambos atributos e infiere a partir de los mismos información temporal y geográfica asociada a los programas, aplicando para ello los procesos de razonamiento semántico que describiremos en las secciones 5.3.2 y 5.4.2, respectivamente.

- El resto de la información semántica se formaliza mediante propiedades *Object* (para las que se definen, además, sus respectivas propiedades inversas), de ahí que la mayoría de las características semánticas descritas en la Sección 5.2.2 (alertas de contenido, créditos, temática, información geográfica y temporal asociada al programa, etc.) correspondan a instancias concretas en la ontología OWL.

Nuestro trabajo identifica tales instancias, así como la jerarquía de clases a las que éstas pertenecen, inspeccionando los metadatos TV-Anytime. Esta especificación define una clasificación jerárquica de posibles valores para cada una de las características semánticas presentadas en la Sección 5.2.2. Conviene destacar que, para algunas características semánticas, la norma define jerarquías con varios niveles de especialización. A partir de tales jerarquías, nuestro proceso de formalización semántica extrae las clases e instancias de la base de conocimiento OWL con arreglo al siguiente criterio: los valores definidos en el nivel más específico de la norma se incluyen en la base de conocimiento OWL como instancias, mientras que los valores establecidos en los niveles más genéricos son los que identifican las clases a las que pertenecen dichas instancias. Por otra parte, dichas clases respetan la estructura jerárquica definida en la propia especificación TV-Anytime. Tal como describiremos en los siguientes capítulos, este enfoque resulta especialmente beneficioso para las dos estrategias de personalización que incluye nuestra metodología, ya que permite descubrir contenidos y atributos que potencialmente pueden interesar al usuario, por estar relacionados (semánticamente) con las preferencias definidas en su perfil personal.

Una vez presentadas las propiedades definidas en nuestra ontología OWL, describimos cada una de las características semánticas formalizadas en la misma.

Alertas de contenido

La primera jerarquía de valores TV-Anytime que presentamos, define una clasificación de categorías para alertar a los usuarios sobre la presencia de determinados tipos de contenidos en los programas difundidos.

Un extracto de esta clasificación —descrita de forma íntegra en el Apéndice A— se muestra en la Tabla 5.3. A partir de ésta, nuestro enfoque extrae las clases e instancias representadas en la Fig. 5.2, aplicando el criterio explicado anteriormente. Conviene destacar que, en este caso, la norma define un único nivel de especialización para cada una de las diez alertas de contenido contempladas (identificadas mediante los códigos 6.1 a 6.10), de ahí que la jerarquía de alertas de contenido formalizada en nuestra ontología OWL tenga el aspecto mostrado en la Fig. 5.2. Dicha figura incluye diez clases genéricas y un conjunto de instancias específicas en cada una de ellas, extraídas del único nivel de especialización definido en la norma. Asimismo, nuestro enfoque también incorpora una clase raíz *Alerts* a la que se ha asignado un código de valor 6, no utilizado en ninguna de las descripciones semánticas incluidas en la norma.

Por último, destacar que los diferentes tipos de alertas considerados en la formalización OWL se asocian a cada contenido de TV mediante una propiedad *Object* denominada *hasAlert*.

Idioma de los contenidos

La especificación TV-Anytime establece una clasificación muy detallada de los idiomas en los que se emiten los contenidos audiovisuales, incluida en el Apéndice A de esta memoria (Tabla A.3). Dichos idiomas estás identificados de forma unívoca mediante los códigos de dos letras definidos en la especificación ISO/IEC 639-1 (por ejemplo, “gl” identifica el idioma gallego).

Alertas de contenido	Código numérico
ALERTS	6
SEX	6.1
No sex descriptor	6.1.1
Implied sexual activity	6.1.2
Sexuality	6.1.3
Explicit sexual activity	6.1.4
Sexual violence	6.1.5
NUDITY	6.2
No nudity descriptors	6.2.1
Partial nudity	6.2.2
Full frontal nudity	6.2.3
VIOLENCE - HUMAN BEINGS	6.3
No violence descriptor	6.3.1
Deliberate infliction of pain to human	6.3.2
Strong psychological pain to human	6.3.3
Deliberate killings of human	6.3.4
VIOLENCE - ANIMALS	6.4
...	...
Deliberate infliction pain to animals	6.4.2
...	...
VIOLENCE - FANTASY CHARACTERS	6.5
...	...
Deliberate killings of fantasy characters	6.5.3
...	...
LANGUAGE	6.6
Occasional use of swear words	6.6.2
Frequent use of swear words	6.6.3
...	...
DISTURBING SCENES	6.7
...	...
Gore	6.7.3
...	...
DISCRIMINATION	6.8
...	...
Deliberate discrimination (on the basis of gender, race, etc.	6.8.2
ILLEGAL DRUGS	6.9
...	...
Illegal drugs use with instructive detail	6.9.3
STROBING	6.10
...	...
Strobing on photosensitive epilepsy	6.10.2

Tabla 5.3: Clasificación de alertas de contenido definidas en TV-Anytime

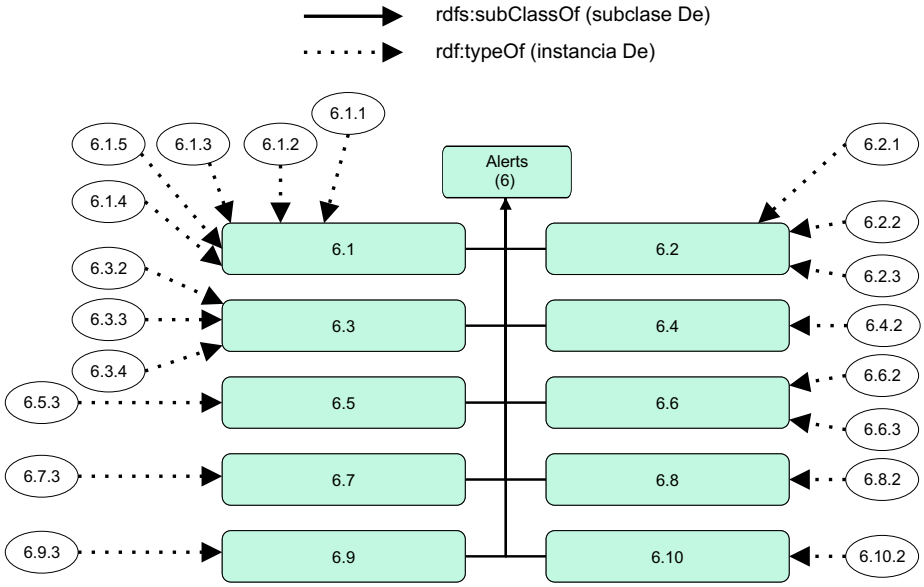

Figura 5.2: Extracto de las clases e instancias referidas a alertas de contenido en la ontología OWL (ver Tabla 5.3)

Para poder representar este tipo de conocimiento en nuestra ontología OWL, hemos incluido una clase genérica *Languages*, cuyas instancias específicas son todos los valores considerados en TV-Anytime (identificadas mediante sus respectivas combinaciones alfabéticas). Cada una de dichas instancias está relacionada con los contenidos de TV representados en la ontología mediante una propiedad *Object* denominada *hasLanguage*.

Créditos que participan en los contenidos

TV-Anytime también establece una relación muy precisa de los diferentes roles que una persona puede desempeñar en un contenido de TV. De forma análoga a la comentada en el apartado anterior, la especificación enumera un conjunto de posibles valores, omitiendo cualquier tipo de organización jerárquica. Nuestro enfoque define una clase OWL para cada uno de estos roles: *Actors*, *Directors*, *Participants*, etc. Éstas se definen como subclases de una clase genérica *Credits* y se asocian a los contenidos de TV incluidos en la ontología mediante diferentes propiedades: *hasActor*, *hasDirector*, *hasParticipant*, etc. Un pequeño subconjunto de dichas clases se muestra en la Tabla 5.4; la clasificación íntegra está disponible en el Apéndice A (Tabla A.4). Obviamente, las instancias de estas clases serán los individuos concretos que desempeñan los diferentes roles contemplados en un programa de TV.

Origen técnico del contenido

La norma TV-Anytime incorpora entre sus descripciones semánticas un atributo referido al origen técnico de los contenidos de TV. La relación de los posibles valores

Créditos	Código numérico
CREDITS	V7
Dubber	V708
Key character	V709
Participant	V43
Stunts	V710
Choreographer	V80
Costume designer	V484
Floor manager	V487
Director	V42
Dresser	V486
Interviewed guest	V97
...	...

Tabla 5.4: Clasificación de los créditos de los contenidos en TV-Anytime

de este atributo también puede consultarse en el Apéndice A, donde figuran las seis categorías genéricas definidas por la norma, así como las más específicas establecidas a partir de ellas. De acuerdo a lo explicado en las características semánticas previas, las primeras se representan en nuestra ontología OWL mediante clases (e.g. 5.1 y 5.2 en la Tabla 5.5) y las segundas mediante instancias específicas de las mismas (5.1.1, 5.1.2, y 5.1.3, entre otras, en dicha tabla).

Origen técnico del contenido	Código numérico
ORIGINATION	5
STUDIO	5.1
Live	5.1.1
As live	5.1.2
Edited	5.1.3
MADE ON LOCATION	5.2
Live	5.2.1
As live	5.2.2
Edited	5.2.3
...	...

Tabla 5.5: Clasificación del origen técnico de los contenidos en TV-Anytime

Relación existente entre diferentes contenidos

El fórum TV-Anytime también define algunas de las relaciones que es posible identificar entre dos contenidos de TV. Este tipo de relaciones son especialmente útiles en una metodología como la desarrollada en nuestro trabajo, en la que es posible aprovechar las correspondencias establecidas entre los programas definidos en la ontología del sistema, para mejorar —gracias al razonamiento semántico— las recomendaciones

ofrecidas en los enfoques tradicionales.

A fin de aprovechar el conocimiento que aportan tales relaciones, nuestra ontología OWL las representa en la ontología OWL del sistema mediante un conjunto de propiedades *Object*. Dado que dichas propiedades se establececen entre programas de TV concretos, el rango y dominio de las mismas coincide con la clase *TVContents* mostrada en la Fig. 5.1. En la Tabla 5.6 identificamos estas propiedades y describimos las relaciones que las originan.

Descripción genérica	Propiedades OWL
Un contenido es el tráiler de otro.	*TrailerOf*
Un contenido es el tráiler de un grupo de contenidos.	*GroupTrailerOf*
Un contenido es "hermano" de otro (por ejemplo, dos capítulos de una misma serie).	*SiblingOf*
Dos programas tienen el mismo contenido pero uno de ellos es una versión alternativa con ciertas características mejoradas (por ejemplo, calidad, sonido, etc.).	*AlternativeOf*
Un programa es el "padre" de otro (por ejemplo, un contenido que agrupa todos los episodios de una misma serie).	*ParentOf*
Un contenido es recomendado al usuario por haber visto antes otro programa.	*RecommendationOf*
Un grupo de contenidos es recomendado al usuario tras haber visto otro programa.	*GroupRecommendationOf*
Un programa incluye anuncios comerciales sobre el contenido de otro.	*CommercialAdventOf*
Un programa permite comprar productos que aparecen en otro contenido.	*DirectProductPurchaseOf*
Un contenido proporciona información adicional sobre otro programa.	*ForMoreDetail*
Un contenido incluye una crítica sobre otro programa.	*ReviewOf*
Un contenido contiene un resumen de otro programa.	*RecapOf*
Un programa cuenta cómo se ha creado otro contenido.	*MakingOf*
Un programa contiene información sobre los temas tratados en otro contenido.	*SupportOf*

Tabla 5.6: Relaciones entre contenidos definidas en TV-Anytime

Tipo de audiencia objetivo de un contenido

La norma TV-Anytime también establece una clasificación del tipo de audiencia al que va dirigido un determinado programa, atendiendo a múltiples criterios. Entre estos últimos, podemos destacar la edad de los espectadores, el grupo social al que pertenecen, su profesión, su nivel cultural o incluso la situación geográfica que ocupan en el momento de ver dicho contenido. Aunque esta información se refiere a los usuarios, nuestra conceptualización la asocia directamente a los programas de TV, debido a que la ontología implementada no incluye datos sobre los espectadores sino únicamente descripciones de contenidos audiovisuales. Este enfoque no impide, sin embargo, in-

ferir ciertas características personales de los usuarios (como las enumeradas anteriormente) a partir de los programas que éstos han visto en el pasado, de ahí su relevancia en el proceso de personalización basado en razonamiento llevado a cabo en AVATAR.

Siguiendo las mismas pautas explicadas en las secciones previas, en la Tabla 5.7 se muestran algunas de las clases e instancias específicas de las mismas, extraídas a partir de los valores definidos en TV-Anytime (disponibles en la Tabla A.5 en el Apéndice A).

Audiencia objetivo del contenido	Código numérico
INTENDED AUDIENCE	4
AGE GROUPS	4.2
Children	4.2.1
age 4-7	4.2.1.1
age 8-13	4.2.1.2
age 14-15	4.2.1.3
Young Adults	4.2.2
age 16-17	4.2.2.1
ge 18-20	4.2.2.2
age 21-24	4.2.2.3
Adults	4.2.3
age 25-34	4.2.3.1
...	...
age 65+	4.2.3.5
All Ages	4.2.4
SOCIAL GROUPS	4.3
Ethnic	4.3.1
Religious	4.3.2
GENDER	4.6
Primarily for men	4.6.1
Primarily for women	4.6.2
For males and females	4.6.3
EDUCATIONAL LEVEL	4.8
Primary	4.8.1
Secondary	4.8.2
LIFESTYLE STAGES	4.9
Single	4.9.1
Couple	4.9.2
Family with children 16+	4.9.6

Tabla 5.7: Clasificación del tipo de audiencia objetivo de un contenido

Tal como adelantábamos en la Sección 5.2.2, la norma define varios niveles de especialización a la hora de clasificar la audiencia objetivo de los programas: el primero de ellos identifica nueve categorías genéricas (de 4.1. a 4.9). Tal como muestra la Tabla 5.7, al segundo nivel pertenecen, por ejemplo, las categorías 4.2.1 a 4.2.4, que identifican cuatro grupos de edades. Para estas últimas, la norma también define un

tercer nivel, en el que se establecen categorías más específicas como las relativas a los código 4.2.1.1 a 4.2.1.3, correspondientes al colectivo de los niños entre 4 y 15 años.

De acuerdo al criterio de formalización semántica establecido al comienzo del capítulo, estas últimas categorías identifican instancias específicas en nuestra base de conocimiento OWL, mientras que las pertenecientes a los niveles superiores corresponden a clases organizadas jerárquicamente. Tal organización se representa en la Fig. 5.3, donde se aprecia que las categorías 4.2.1.1 a 4.2.1.3 son instancias específicas de la clase 4.2.1 y, por tanto, también de su superclase 4.2.

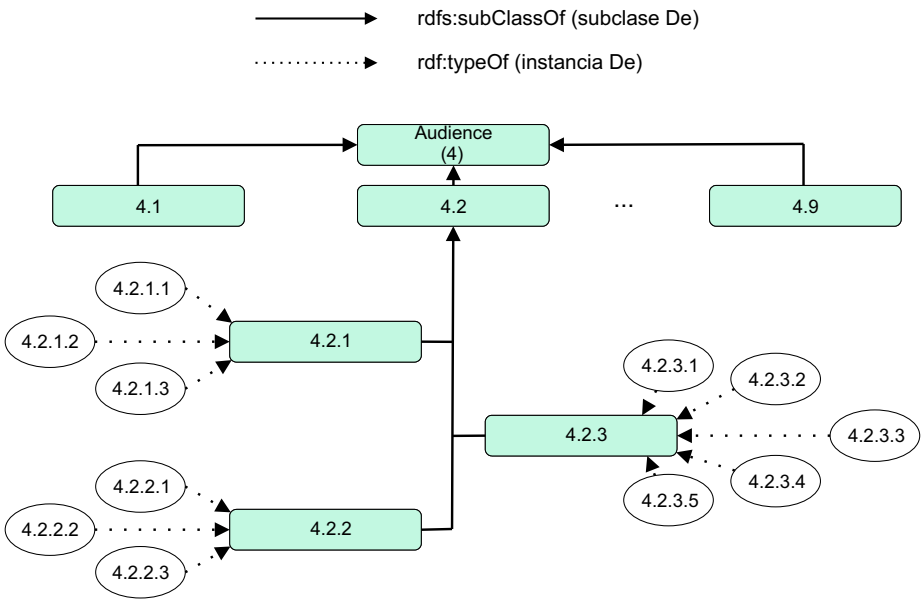

Figura 5.3: Extracto de las clases e instancias referidas al tipo de audiencia de los contenidos en la ontología OWL (ver Tabla 5.7)

Por último, destacar que las instancias referidas al tipo de audiencia de un determinado contenido están relacionadas con éste mediante una propiedad *Object* denominada *hasIntendedAudience*.

Intención del contenido

La especificación TV-Anytime también incluye entre sus descripciones semánticas la intención principal con la que ha sido creado un determinado contenido. Tal como muestra la Tabla 5.8, la clasificación definida por la norma incluye propósitos como entretenimiento, información, educación o fines promocionales, entre otros.

Las clases e instancias específicas extraídas a partir de la relación de valores mostrados en la Tabla 5.8 se asocian a los contenidos de TV de la ontología mediante la propiedad *hasIntention*.

Intención del contenido	Código numérico
INTENTION	1
ENTERTAINMENT	1.1
Pure entertainment	1.1.1
Informative entertainment	1.1.2
INFORMATION	1.2
Government	1.2.1
Pure information	1.2.2
Infotainment	1.2.3
Advice	1.2.4
EDUCATION	1.3
School programmes	1.3.1
Primary	1.3.1.1
Secondary	1.3.1.2
Tertiary	1.3.1.3
Further education	1.3.2
...	...

Tabla 5.8: Clasificación de la intención de un contenido definida en TV-Anytime

Formato estructural del contenido

Además de considerar la temática abordada en cada contenido, la especificación TV-Anytime también permite clasificar los programas atendiendo únicamente a su estructura. Para ello, tal como muestra la Tabla 5.9, la norma define un conjunto de categorías genéricas con sus respectivas especializaciones, traducidas a la ontología OWL mediante clases e instancias concretas, respectivamente. De forma análoga a la descrita en las características semánticas previas, dichas instancias están asociadas a los programas de TV mediante propiedades *hasFormat*. La clasificación completa de formatos estructurales definida en TV-Anytime puede consultarse en el Apéndice A (Tabla A.7).

Características psicológicas/emocionales asociadas a los contenidos

Este atributo semántico permite describir ciertas características psicológicas o emocionales asociadas a los contenidos de TV. La especificación simplemente define una relación de cincuenta y tres posibles valores, sin establecer ningún tipo de organización jerárquica. Precisamente debido a esto último, los códigos numéricos que identifican dichos valores se incorporan a la ontología OWL como instancias concretas de una clase genérica *Atmospheres*, y se asocian a cada programa mediante la propiedad *hasAtmosphere*. Para consultar la relación completa de valores definidos en la especificación, se remite al lector al Apéndice A (Tabla A.8).

Tipo de productos ofertados en contenidos con fines comerciales

Este atributo semántico describe el tipo de productos o servicios ofrecidos en un determinado anuncio comercial. En concreto, la norma diferencia diez categorías ge-

Formato del contenido	Código numérico
FORMAT	2
STRUCTURED	2.1
Bulletin	2.1.1
Magazines	2.1.2
Commented event	2.1.3
Documentary	2.1.4
Interview/Debate	2.1.5
Lecture/Speech	2.1.6
Textual	2.1.7
Phone-in	2.1.8
DJ with discs	2.1.9
REPRESENTATION/PLAY	2.2
Fictional portrayal of life	2.2.1
Readings	2.2.2
Representation with puppets	2.2.3
CARTOON/ANIMATION	2.3
Anime	2.3.1
Computer	2.3.2
Cartoon	2.3.3
SHOW	2.4
Hosted show	2.4.1
Single Game show	2.4.1.1
Big Game show	2.4.1.2
Panel show	2.4.2
Single Game show	2.4.2.1
Big Game show	2.4.2.2
Non-hosted show	2.4.3
Stand-up comedian(s)	2.4.4
...	...

Tabla 5.9: Clasificación del formato estructural de un contenido en TV-Anytime

néricas y define especializaciones en cada una de ellas, desde las cuales se extraen, de manera inmediata, las clases e instancias específicas mostradas en la Tabla 5.10.

De forma análoga a la comentada en los atributos anteriores, el lector puede consultar la relación completa de valores descritos en la especificación TV-Anytime en el Apéndice A (Tabla A.9).

Fecha y localización de producción, creación y estreno de los contenidos

Por último, para modelar el conocimiento relativo a la fecha y localización de producción, creación y estreno de los contenidos de TV, nuestro enfoque utiliza dos ontologías en las que se formaliza información geográfica y temporal, cuya descripción se detalla en las siguientes secciones. Dichas ontologías permitirán mejorar las capacidades de razonamiento semántico del sistema, recurriendo a inferencias lógicas automáticas.

5.3 Información temporal

La ontología OWL-Time [96], desarrollada por miembros del consorcio W3C, permite llevar a cabo procesos de razonamiento automáticos de especial utilidad en un sistema inteligente como el propuesto en esta tesis.

5.3.1 La ontología OWL-Time

OWL-Time es una ontología que surgió con el claro objetivo de facilitar descripciones y propiedades temporales sobre los documentos y servicios disponibles en la Web. Esta conceptualización OWL proporciona un vocabulario formal para expresar las relaciones existentes entre instantes e intervalos temporales, así como información sobre la duración de estos últimos, fechas y horas concretas.

Si bien la ontología en su totalidad puede ser consultada en [96], en esta sección simplemente describimos aquellas clases y propiedades que están dentro del ámbito de la tesis, y que resultan especialmente útiles a la hora de aplicar procesos de razonamiento automáticos basados en la información temporal asociada a los contenidos de TV.

A la hora de definir conceptos temporales, OWL-Time incluye una clase genérica *TemporalEntity* y dos subclases de la misma, que identifican tanto los instantes como los intervalos temporales (*Instant* e *Interval*, respectivamente). Con el fin de establecer relaciones entre ambas entidades, la conceptualización define tres propiedades específicas:

- *begins*: Esta relación permite identificar el instante temporal en el que comienza un determinado intervalo. Es una propiedad OWL *Object* cuyo dominio y rango son las clases *TemporalEntity* e *Instant*, respectivamente.

- *ends*: Esta propiedad define el instante en el que finaliza un determinado intervalo temporal.

Formato del contenido	Código numérico
COMMERCIAL/PRODUCTS	3.50
Agriculture, forestry and fishery products	3.50.1
Products of agriculture, horticulture and market gardening	3.50.1.1
Live animals and animal products	3.50.1.2
...	...
ORES AND MINERALS; ELECTRICITY, GAS AND WATER	3.50.2
Coal and lignite	3.50.2.1
Crude petroleum and natural gas	3.50.2.2
...	...
FOOD PRODUCTS, BEVERAGES AND TOBACCO; TEXTILES AND LEATHER PRODUCTS	3.50.3
Meat, fish, fruit, vegetables, oils and fats	3.50.3.1
Dairy products	3.50.3.2
...	...
OTHER TRANSPORTABLE GOODS	3.50.4
Products of wood, cork, straw and plaiting materials	3.50.4.1
Paper products	3.50.4.2
...	...
METAL PRODUCTS, MACHINERY AND EQUIPMENT	3.50.5
Basic metals	3.50.5.1
General purpose machinery	3.50.5.3
...	...
INTANGIBLE ASSETS; LAND; CONSTRUCTION SERVICES	3.50.6
Land	3.50.6.2
Constructions	3.50.6.3
...	...
DISTRIBUTIVE TRADE SERVICES; LODGING; TRANSPORT SERVICES	3.50.7
Wholesale trade services	3.50.7.1
Air transport services	3.50.7.6
...	...
FINANCIAL AND RELATED SERVICES; REAL ESTATE SERVICES; AND LEASING SERVICES	3.50.8
Insurance services	3.50.8.1
Leasing and rental services	3.50.8.3
...	...

Tabla 5.10: Clasificación de productos ofrecidos en contenidos comerciales definida en TV-Anytime

- *inside*: Esta relación se establece entre un instante temporal y el intervalo en el que está incluido.

Además, para poder asociar información horaria y temporal concreta a un instante, OWL-Time define la propiedad *Datatype inXSDDateTime*, cuyo dominio es la clase *Instant* y cuyo rango corresponde al tipo de datos XML Schema *xsd:dateTime*[2]. Todas las clases y propiedades mencionadas hasta ahora aparecen resumidas en la Fig. 5.4.

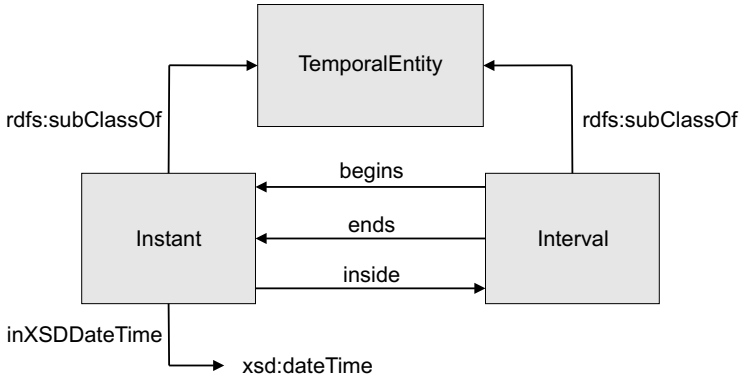

Figura 5.4: Extracto de las clases y propiedades de la ontología OWL-Time

OWL-Time formaliza las clases y propiedades en relación a la información temporal, pero no incluye instancias específicas. Estas últimas deben ser incorporadas a la base de conocimiento en función de las necesidades de cada tipo de aplicación en la que se emplea dicha ontología. En nuestro dominio de televisión personalizada, las fechas de creación, producción y estreno de los contenidos TV-Anytime mencionados en la Sección 5.2.2 son instancias de la clase *TemporalEntity*, pudiendo, por tanto, representar tanto instantes concretos como intervalos que se extienden a lo largo del tiempo. Dichas instancias son extraídas directamente a partir de las descripciones semánticas proporcionadas por la especificación, y asociadas a cada programa mediante las propiedades *hasCreationDate*, *hasProductionDate* y *hasPremiereDate*. Mediante este proceso, nuestro sistema puede aprovechar el conocimiento formalizado en la ontología OWL-Time, y con ello mejorar sus capacidades de razonamiento temporal sobre los atributos semánticos TV-Anytime.

Junto a la información semántica proporcionada por TV-Anytime, nuestro enfoque añade nuevas instancias a la ontología OWL-Time. En concreto, este proceso de instanciación incorpora diferentes eventos temporales a la base de conocimiento, modelándolos como intervalos temporales que se extienden entre dos instantes concretos[3]. Pongamos un ejemplo para clarificar este procedimiento. Supongamos el evento correspondiente a la Primera Guerra Mundial, un acontecimiento histórico ocurrido entre 1914 y 1918. El código OWL que permite formalizar esta instancia en la ontología temporal se muestra a continuación:

[2]Por ejemplo, el instante correspondiente a las 10:30 de la mañana del día 1 de enero de 2003, se expresaría mediante este tipo de datos como *2003-01-01T10:30:00*.

[3]Junto a las clases de las ontologías temporal y geográfica (descrita en la Sección 5.4.1), nuestro enfoque define una clase genérica *Topics*, cuyas instancias son términos que figuran en las sinopsis y las palabras clave de los contenidos de TV.

```
<time:Interval rdf:ID="WorldWarI">
 <time:begins rdf:resource="#beginning_WorldWarI">
 <time:ends rdf:resource="#end_WorldWarI">
</time:Interval>

<time:Instant rdf:ID="beginning_WorldWarI">
 <time:inXSDDataTime rdf:datatype="&xsd:dateTime">1914
 </time:inXSDDataTime>

<time:Instant rdf:ID="end_WorldWarI">
 <time:inXSDDataTime rdf:datatype="&xsd:dateTime">1918
 </time:inXSDDataTime>
```

La inclusión de este tipo de intervalos en la ontología OWL favorece la inferencia de nuevo conocimiento, gracias a los procesos de razonamiento temporal llevados a cabo por el sistema, descritos en las siguientes secciones.

5.3.2 Razonamiento temporal

Tal como hemos descrito en la Sección 5.3.1, los contenidos de TV en nuestra base de conocimiento están relacionados con instantes e intervalos temporales concretos. Precisamente aprovechando las propiedades establecidas en la ontología OWL-Time (*begins*, *ends* e *inside*), es posible inferir relaciones de interés entre diferentes contenidos, cuyo origen es la asociación de éstos con un mismo intervalo temporal.

Además de actualizar la base de conocimiento OWL del sistema AVATAR, incluyendo nuevos contenidos y descripciones semánticas, el agente de clasificación también se ocupa de los procesos de razonamiento temporal descritos en esta sección. Para ello, el agente no sólo emplea la información temporal relativa a las fechas de creación, producción y estreno de los contenidos representados en la ontología de TV del sistema, sino que procesa las sinopsis y palabras claves que la especificación asocia a los contenidos de TV, a fin de extraer nuevo conocimiento sobre los mismos.

En lo relativo a la sinopsis, el agente elimina aquellas palabras que no aportan información relevante de cara al proceso de inferencia semántica (artículos, pronombres, terminaciones comunes, etc.). Una vez realizado esta fase de preprocesado, explora la ontología temporal buscando instancias que se correspondan con los conceptos restantes de la sinopsis y las palabras claves. Así, por ejemplo, si la descripción textual de una película establece que está ambienta en la Primera Guerra Mundial (información no proporcionada en ninguno de los atributos TV-Anytime descritos en la Sección 5.2.2, aunque definida como instancia en OWL-Time), el agente de clasificación puede utilizar los intervalos definidos en la ontología temporal para descubrir nuevas relaciones a partir de las ya formalizadas en la conceptualización.Para poder formalizar este tipo de relaciones en la ontología, nuestro enfoque define una propiedad *Object* denominada *isAbout*.

Las reglas en las que se basa este proceso de inferencia temporal son enumeradas a continuación. Para ello, supongamos que p_1 y p_2 son dos contenidos concretos, $instante_1$ y $instante_2$, dos instantes definidos en OWL-Time y, finalmente, $intervalo_1$ e $intervalo_2$, dos intervalos temporales incluidos en dicha ontología.

- Supongamos, en primer lugar, que p_1 y p_2 están asociados con $intervalo_1$ e $intervalo_2$, respectivamente. En este escenario, si los instantes que delimitan el principio y el final de $intervalo_2$ están incluidos en $intervalo_1$, el agente de clasificación relaciona el contenido p_2 con el intervalo más extenso, basándose precisamente en dicha relación de inclusión. Para ello, nuestro enfoque establece una instancia de la propiedad *isAbout* entre aquellas que identifican al programa y al intervalo, tal como muestra la Fig. 5.5.

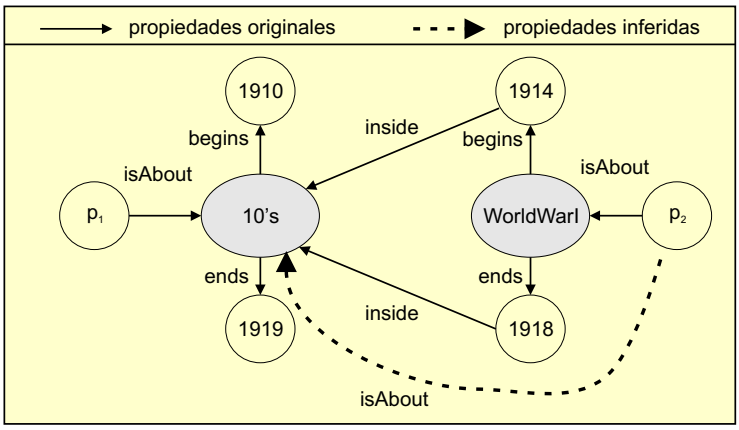

Figura 5.5: Ejemplo de razonamiento temporal (i)

- Supongamos ahora que p_2 sigue estando relacionado con $intervalo_2$, mientras que p_1 está asociado a $instante_1$. En este caso, si dicho instante está incluido en el intervalo temporal de p_2, es posible relacionar ambos contenidos con dicho intervalo, tal como muestra la Fig. 5.6.

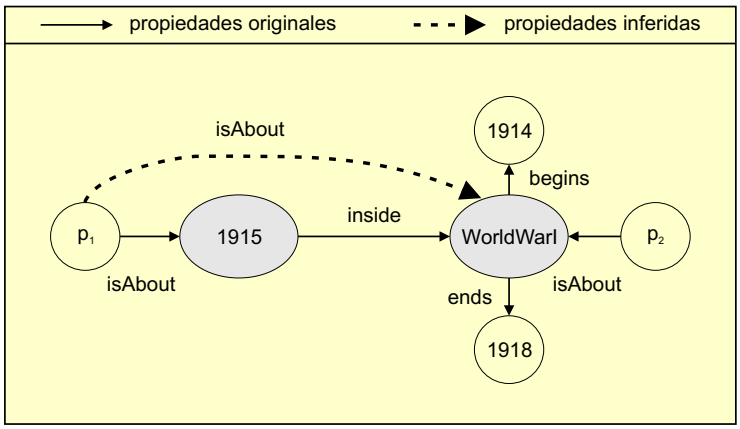

Figura 5.6: Ejemplo de razonamiento temporal (ii)

- Por último, asumimos que ambos contenidos están relacionados con sendos instantes temporales. En este escenario, es necesario descubrir un intervalo que con-

tenga ambos instantes para así poder relacionar los dos contenidos. Un ejemplo de este tipo de razonamiento temporal se muestra en la Fig. 5.7.

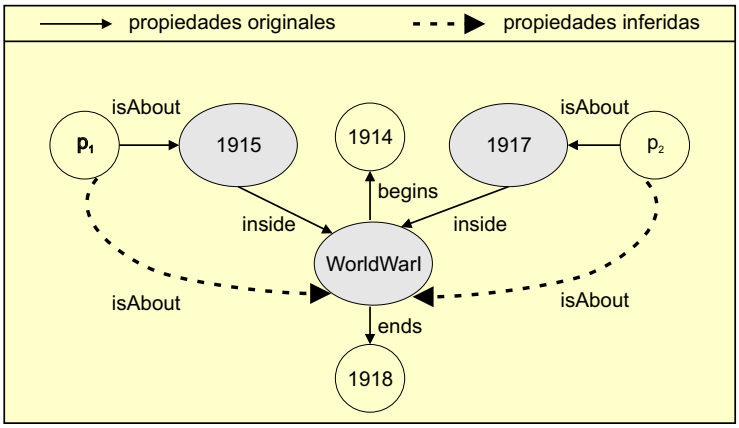

Figura 5.7: Ejemplo de razonamiento temporal (iii)

5.4 Información geográfica

Con el fin de incluir procesos de razonamiento semántico sobre la información geográfica asociada a los contenidos de TV, nuestro enfoque implementa una pequeña ontología OWL (en adelante, GEO) en la que se representan conceptos como continentes, países, provincias, ciudades y pueblos, así como instancias específicas de los mismos. Conviene destacar que esta propuesta todavía no reutiliza las ontologías geográficas disponibles públicamente [179], debido a que éstas representan una información demasiado detallada (coordenadas, latitud, altitud, etc.), excesiva para los procesos inferenciales perseguidos en nuestro sistema. En su lugar, nuestra base de conocimiento incluye únicamente los conceptos mencionados al comienzo.

5.4.1 La ontología GEO

La ontología GEO incluye cinco clases genéricas (*Continents*, *Countries*, *Regions*, *Cities*, *Villages*) relacionadas entre sí mediante la propiedad transitiva *partOf*, tal como muestra la Fig. 5.8[4].

Las instancias de dichas clases identifican las distintas regiones geográficas en las que un contenido ha sido creado, producido y/o estrenado, de forma similar a la comentada en la ontología temporal OWL-Time. Esta información se representa en la ontología de TV mediante las propiedades *hasCreationPlace*, *hasProductionPlace* y *hasPremierePlace*. Con el fin de poder asociar a cada una de ellas cualquiera de las regiones geográficas definidas en la Fig. 5.8, nuestra ontología crea una nueva clase a partir de la unión de las mismas (denominada *GEOPlaces*), empleada como rango

[4]Destacar que, en virtud de la transitividad de esta última propiedad, existen relaciones entre las clases de la ontología que no han sido representadas en la figura (por ejemplo, las ciudades y los pueblos también son parte de los países y de los continentes).

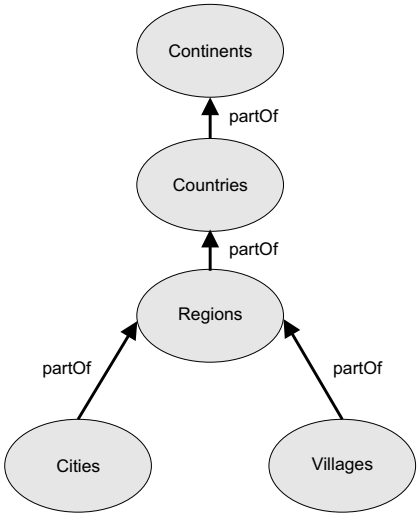

Figura 5.8: Clases y propiedades definidas en la ontología OWL GEO

de las mencionadas propiedades. El código OWL relativo a *hasPremierePlace* se representa a continuación, donde la clase *TVContents* es el nodo raíz de la jerarquía de géneros mostrada en la Fig. 5.1.

```
<owl:Class rdf:ID="GEOPlaces">
 <owl:unionOf rdf:parseType="Collection"
  <owl:Class rf:about="#Continents"/>
  <owl:Class rf:about="#Countries"/>
  <owl:Class rf:about="#Regions"/>
  <owl:Class rf:about="#Cities"/>
  <owl:Class rf:about="#Villages"/>
 </owl:unionOf>
</owl:Class>

<owl:ObjectProperty rdf:ID="hasPremierePlace">
 <rdfs:domain rf:resource="#TVContents">
 <rdfs:range rf:resource="#GEOPlaces">
</owl:ObjectProperty>
```

De forma análoga a la explicada en la Sección 5.3.2, el agente de clasificación incluido en nuestro sistema razona sobre las instancias de la clase *GEOPlaces* con el fin de inferir nuevo conocimiento sobre los contenidos de TV disponibles en la ontología OWL. Este conocimiento permitirá mejorar los procesos inferenciales en los que se basan las recomendaciones semánticas elaboradas por nuestro sistema.

5.4.2 Razonamiento geográfico

Al igual que en el proceso de razonamiento temporal descrito en la Sección 5.3.2, el agente de clasificación también procesa las sinopsis y palabras clave de los contenidos de TV a fin de inferir información geográfica sobre los mismos. Una vez localizadas

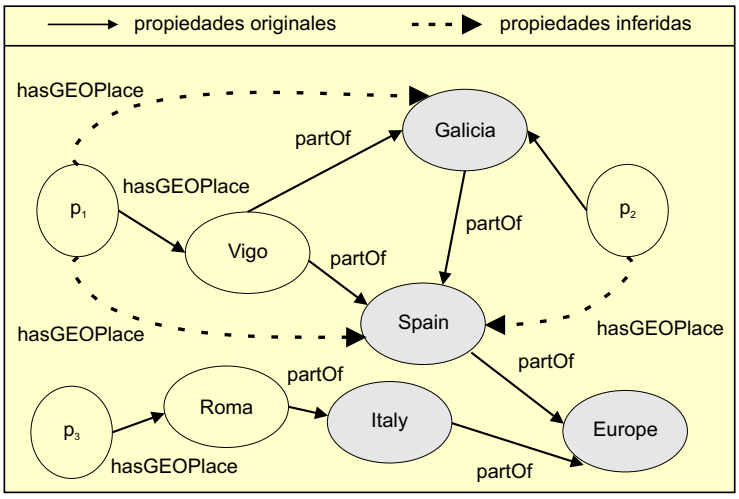

Figura 5.9: Ejemplo de razonamiento geográfico

las instancias específicas de la ontología GEO entre los términos incluidos en ambos campos TV-Anytime, y asociadas a los programas correspondientes mediante propiedades *hasGEOPlace*, el agente infiere relaciones semánticas entre dichos contenidos a partir de esta información geográfica.

La idea es relacionar entre sí aquellos programas que estén asociados con una misma región geográfica. Dado que esta premisa podría conducir al descubrimiento de relaciones demasiado indirectas, las reglas en las que se basa el proceso de razonamiento geográfico propuesto limitan las regiones comunes a provincias, comunidades y países, omitiendo los continentes.

Sean p_1 y p_2 sendos contenidos de TV relacionados con las instancias geo_1 y geo_2, respectivamente, mediante la propiedad *hasGEOPlace*.

- En este escenario, si geo_1 es una región geográfica incluida dentro de la provincia, comunidad o país geo_2, es posible relacionar los dos programas p_1 y p_2 con esta última mediante la propiedad *hasGEOPlace*. Obviamente, un razonamiento análogo puede ser aplicado en el caso de que sea geo_2 la región incluida en geo_1.
- Por otra parte, si existe una provincia, comunidad o país geo_x que incluya tanto a geo_1 como a geo_2, es posible relacionar ambos contenidos con dicha región común.

La Fig. 5.9 ilustra ambos tipos de razonamiento geográfico. Tal como apuntábamos anteriormente, en el escenario planteado, el agente de clasificación no infiere relación alguna entre los contenidos p_1 (vinculado a Vigo) y p_3 (ligado a Roma), al considerar que la pertenencia de dos ciudades a un mismo continente es una asociación geográfica demasiado indirecta, que no aporta conocimiento útil al proceso de recomendación.

Una vez implementada la ontología de dominio, el sistema AVATAR dispone de la base de conocimiento sobre la que poder aplicar los procesos de razonamiento semántico, en los que, obviamente, las preferencias de los usuarios juegan un papel decisivo.

Capítulo 6

Modelado de Usuarios

En este capítulo se presenta la técnica de modelado de usuarios utilizada en la metodología de razonamiento desarrollada en la tesis. Esta técnica representa las preferencias de cada usuario a partir del conocimiento formalizado en la ontología de dominio del sistema, incluyendo tanto los productos específicos que interesan (o no) a este usuario, como sus respectivas descripciones semánticas. Además de asignar niveles de interés adecuados a dichas preferencias, la técnica empleada incorpora un mecanismo de adaptación ante los (posibles) cambios ocurridos en éstas. Tal mecanismo permite que el sistema se adapte a los intereses más recientes del usuario y olvide los más antiguos, identificando así preferencias obsoletas que deben ser eliminadas de su perfil.

6.1 Perfiles de usuario basados en ontologías

Para poder razonar sobre las preferencias de los usuarios, la metodología desarrollada requiere una representación formal de las mismas, en la que se incluyan descripciones semánticas tanto de los contenidos que agradan al espectador (denominados de forma genérica *preferencias positivas*) como de aquellos que no le interesan (*preferencias negativas*). El cometido de tales descripciones es facilitar el aprendizaje de gran cantidad de conocimiento sobre las preferencias del usuario, para después poder utilizarlo de forma eficiente durante las tareas de personalización automática.

Así, junto a los contenidos de TV específicos que identifican las preferencias del usuario, los perfiles manejados en AVATAR registran el nivel de interés asociado a los mismos (positivo o negativo según sea o no relevante para el espectador), y una relación de sus características semánticas más significativas: créditos, temática, información temporal y geográfica asociada a dichos contenidos, etc.

Este tipo de requisitos impide que la mayoría de las técnicas de modelado de usuarios propuestas en el estado del arte puedan ser utilizadas en nuestro trabajo. En concreto, a excepción de las redes semánticas y asociativas, el resto de los métodos revisados en la Sección 2.2 no son válidos en nuestro dominio, ya que definen modelos de usuarios demasiado simples, basados en listas de palabras claves o en las clasificaciones dadas por los usuarios en relación a cada uno de los productos específicos incluidos en sus perfiles. Esta limitación en lo que a la representación del conocimiento sobre las preferencias de los usuarios se refiere, se traduce en capacidades de inferencia muy pobres, insuficientes para explotar el potencial de sus descripciones semánticas.

Por su parte, tal como introdujimos en la Sección 2.2, las redes semánticas y asociativas permiten representar el conocimiento identificando tanto los conceptos (nodos) relevantes en un determinado dominio, como las relaciones establecidas entre éstos, convenientemente etiquetadas (propiedades). Son precisamente tales capacidades expresivas, las que convierten a estas redes en herramientas de modelado válidas en el dominio de nuestro sistema semántico.

Teniendo en cuenta que la información relativa a las preferencias de los usuarios está ya formalizada en la ontología[1] de TV desarrollada, nuestra propuesta de modelado [25] se basa en construir los perfiles a partir de la misma —de ahí que reciban el nombre de *perfiles-ontología*—, aprovechando con ello la capacidad de reutilización de conocimiento asociada a este tipo de conceptualizaciones.

Esta propuesta recuerda al enfoque de modelado de usuarios propuesto por Middleton en el sistema QuickStep [149] (ver Sección 3.10.3). A diferencia de esta aproximación, nuestra metodología no emplea únicamente la estructura jerárquica de la ontología de TV a la hora de modelar las preferencias de los usuarios, sino que considera una representación más compleja, extraída a partir del conocimiento formalizado en la ontología OWL descrita en el capítulo previo. La diferencia entre ambas propuestas está relacionada con la naturaleza de las estrategias de recomendación empleadas en cada una de ellas. Por su parte, Middleton basa sus recomendaciones en el cálculo de correlaciones entre los perfiles jerárquicos de cada usuario, de ahí que la representación de los mismos como vectores de las categorías ontológicas que interesan a cada usuario sea suficiente. Por el contrario, AVATAR razona sobre el conocimiento de la ontología para elaborar sugerencias mejoradas a partir de la inferencia de asociaciones semánticas, razón por la cual es necesario disponer de una representación más compleja de las preferencias de los espectadores.

Si bien esta representación permite que la metodología propuesta descubra relaciones semánticas de interés para los usuarios, y con ello mejorar sus recomendaciones personalizadas, no es menos cierto que también es el origen de ciertas limitaciones. Éstas están relacionadas fundamentalmente con las capacidades de almacenamiento requeridas en el servidor del sistema (ver Fig. 4.4) a la hora de registrar los *perfiles-ontología* de todos los usuarios. Con el propósito de aliviar tales requisitos, nuestra metodología adopta dos medidas:

- Por una parte, dado que la información semántica sobre los contenidos de TV está ya formalizada en la ontología OWL, nuestra metodología no replica en cada perfil las clases, instancias y propiedades que identifican las preferencias del usuario. En su lugar, cada perfil mantiene únicamente referencias a las entidades formalizadas en la conceptualización OWL y, a la vez, registra los niveles de interés asociados a las mismas (para identificar las preferencias positivas y negativas del espectador). Así, es posible mantener, de forma eficiente, el modelo semántico completo sobre los intereses del usuario requerido en nuestra metodología de razonamiento.

- Por otro lado, nuestro enfoque considera una función de olvido gradual (ver Sección 2.5) que permite que el proceso de recomendación se base únicamente en las preferencias más recientes del usuario, prescindiendo de todas aquellas clases e instancias referidas a intereses pasados. Además de la reducción del ta-

[1] Recordar que las ontologías no son más que redes semánticas en las que se ha definido, de forma consensuada, una semántica formal (ver Sección 3.2.1).

maño de los perfiles de los usuarios, esta medida conlleva otra ventaja importante para el sistema, pues facilita que éste se adapte de forma flexible a los posibles cambios en las preferencias de los espectadores. Una de las características más atractivas de la función de olvido gradual incluida en nuestra metodología es su naturaleza dinámica; ésta permite variar la selección de las preferencias actuales de cada usuario, en función de la precisión de las últimas recomendaciones ofrecidas por el sistema.

6.2 Composición de los *perfiles-ontología*

En esta sección nos centramos en el proceso de inicialización del *perfil-ontología* de cada espectador a partir de la información que éste proporciona durante su registro en el sistema. Asimismo, describimos cómo infiere el recomendador nuevo conocimiento sobre las preferencias de sus usuarios, basándose en las acciones (explícitas e implícitas) llevadas a cabo por los mismos. Esta realimentación de relevancia resulta especialmente ventajosa en el dominio de la personalización, pues permite que el recomendador mantenga los perfiles de los usuarios permanentemente actualizados.

6.2.1 Generación del perfil inicial

Uno de los objetivos perseguidos en el diseño de AVATAR es la minimización de la interacción entre el espectador y el sistema, en lo que a aporte de información explícita se refiere. Tal como describimos en la Sección 2.3.2, este propósito es incompatible con dos de los métodos de generación del perfil inicial de cada usuario allí revisados: la inicialización manual por parte del espectador y la utilización de estereotipos. La primera de ellas requiere que los usuarios inviertan mucho tiempo rellenando tediosos formularios, en los que deben incluir tanto sus preferencias como datos personales. En relación a estos últimos, conviene recordar además que los espectadores no siempre están dispuestos a desvelar información de carácter personal, lo que dificulta la definición de los estereotipos.

Ambas razones han decantado nuestra opción de inicialización del perfil de usuario hacia los métodos basados en conjuntos de entrenamiento. En concreto, AVATAR presenta una jerarquía de géneros (de contenidos de TV), para que el usuario seleccione sus preferencias positivas y negativas. Entre las ventajas de este método frente a los mencionados previamente, cabe destacar que esta organización jerárquica favorece una exploración rápida y cómoda, a la vez que representa información que los espectadores pueden identificar fácilmente (géneros como deportes, música, noticias y sus respectivas clasificaciones más específicas). Así, por ejemplo, si a un usuario le interesan únicamente los programas del género ficticio, no es necesario que explore el conjunto de entretenimiento en su totalidad, sino únicamente aquellas categorías localizadas bajo la clase *Fiction* en la jerarquía mostrada. Es obvio que este ahorro en el tiempo de búsqueda del espectador no sería posible en los métodos tradicionales basados en un conjunto de entretenimiento plano, carente de cualquier organización jerárquica. En realidad, esta jerarquía —construida por el agente de registro incluido en el proveedor de servicios (ver Fig. 5.1)—, contiene únicamente el subconjunto más representativo de las clases definidas en la ontología OWL de TV, ya que la presentación de todas ellas conllevaría tiempos de navegación quizás inasumibles para muchos de los usuarios.

A la vez, el sistema incorpora en el conjunto de entrenamiento mostrado al espectador algunas instancias (programas específicos) bajo las categorías más específicas de la jerarquía. De nuevo, destacar que la organización jerárquica de dicho conjunto permite que los usuarios puedan identificar rápidamente tanto aquellos contenidos que les interesan, como otros que representan sus preferencias negativas.

Con el fin de no saturar al espectador mostrando gran cantidad de información durante la fase de registro en el sistema, el enfoque propuesto presenta un número reducido de instancias bajo cada clase ontológica. Como consecuencia, es posible que el usuario no localice en el conjunto de entrenamiento presentado alguno(s) de los contenidos que identifican sus preferencias positivas y negativas. Tal como adelantamos en el capítulo 4, en esta situación, el agente que gestiona la comunicación entre el STB del usuario y el servidor del proveedor, notifica la ausencia de dichos contenidos al agente que procesa las solicitudes recibidas. Como consecuencia de esta notificación, se desencadenan dos acciones: en primer lugar, el agente de clasificación localiza los contenidos especificados por el usuario y los incorpora a la ontología OWL del sistema; a continuación, es el agente que actualiza los perfiles de usuario quien incorpora este programas como preferencias iniciales del espectador.

El proceso de inicialización del perfil propuesto combina tanto la selección de programas desde la jerarquía mostrada al espectador, como la solicitud explícita de contenidos, dado que ambos métodos se adaptan de forma flexible a las necesidades de cada usuario. En concreto, los usuarios más pasivos pueden inicializar sus perfiles indicando explícitamente algunos contenidos que le han interesado (o no) en el pasado, sin necesidad de explorar la jerarquía mostrada durante la fase de registro en el sistema. Por el contrario, aquellos usuarios que prefieren asumir un rol más participativo pueden completar este conjunto de programas, identificando rápidamente en la estructura jerárquica otros contenidos que podrían haberle pasado desapercibidos.

A partir de los contenidos seleccionados por el espectador, el sistema puede descubrir gran cantidad de información personal sobre el usuario, especialmente útil durante el proceso de recomendación. En concreto, tal como se describió en el capítulo 5, la ontología OWL implementada incluye diferentes características semánticas para los programas considerados, como alertas de contenido, intención de cada programa, tipo de audiencia a la que va dirigido, etc. A través de esta información semántica, el sistema aprende conocimiento sobre sus preferencias y datos de carácter personal (por ejemplo, tipo de profesión del usuario, su nivel cultural, audiencia presente en el momento de solicitar una recomendación, etc.), sin necesidad de que éste tenga que definirla explícitamente, tal como ocurriría en las técnicas basadas en estereotipos o en el proceso de generación manual.

Los programas seleccionados por el usuario, ya sean preferencias positivas o negativas, son incorporados a sus perfiles mediante el procedimiento descrito en la siguiente sección.

6.2.2 Realimentación de relevancia

El enfoque propuesto modela las preferencias de los usuarios utilizando tanto la información aportada por los mismos al registrarse en el sistema, como el conocimiento disponible en la ontología de TV. A partir de los contenidos especificados por el usuario durante la fase de registro, nuestra metodología construye su perfil manteniendo únicamente referencias a las instancias correspondientes a estos programas en la

conceptualización OWL.

Dichas referencias no son más que los CRIDs que identifican de forma unívoca cada contenido TV-Anytime definido en la ontología de TV (ver Sección 4.2.2). De esta forma, almacenando únicamente el identificador del contenido, es posible consultar la base de conocimiento del sistema y, una vez localizada la instancia correspondiente a dicho CRID, acceder tanto a las características semánticas del programa como a la jerarquía de géneros a la que éste pertenece (ver Fig. 5.1), sin necesidad de replicar dicha información en cada modelo de usuario.

Como consecuencia, y tal como muestra la Fig. 6.1, si varios usuarios del sistema especifican los mismos contenidos, la metodología de modelado propuesta simplemente mantiene las referencias oportunas desde sus respectivos perfiles a las entidades formalizadas en un único repositorio de conocimiento (la ontología OWL).

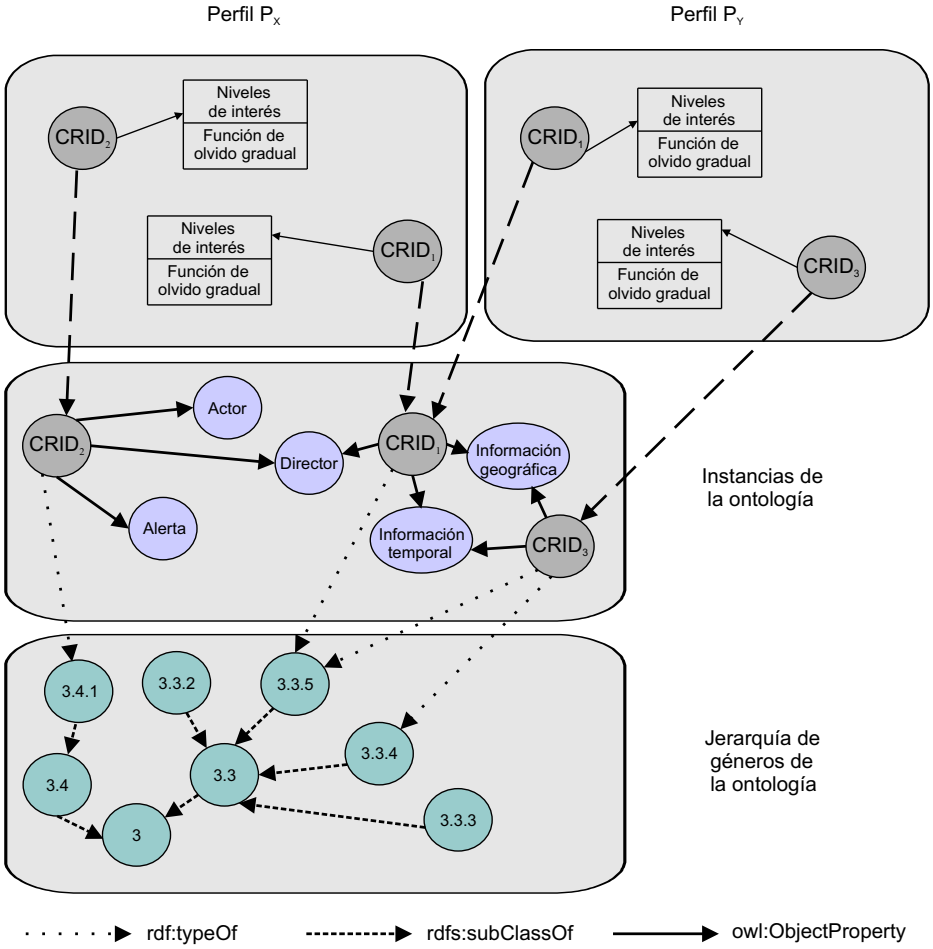

Figura 6.1: Técnica de modelado de usuarios empleada en AVATAR

Junto al CRID de cada programa, el perfil mantiene los niveles de interés de este contenido, de sus características semánticas y de los géneros a los que pertenece, así como los índices que permiten manejar la función de olvido gradual incorporada en

nuestra metodología[2].

Las características semánticas son especialmente interesantes en un dominio como el de la televisión personalizada y, en concreto, en un sistema semántico como AVATAR, ya que pueden estar asociadas a varios de los programas definidos en el perfil del usuario, permitiendo identificar relaciones entre sus preferencias (por ejemplo, es común que a un espectador le interesen varias películas donde interviene un mismo actor o que pertenecen a un mismo género temático). Además, mediante este modelo, el sistema puede conocer no sólo los contenidos que interesan (o no) al usuario, sino también sus actores, directores o temáticas favoritas, atributos de innegable utilidad en un entorno de personalización. Pongamos un ejemplo para clarificar el proceso de construcción de cada *perfil-ontología*.

Ejemplo: En este escenario de aplicación, consideramos que un usuario U_X se registra en el sistema y declara tanto su interés por las películas *Top Gun*, *Cocktail*, *Mejor imposible* y *Náufrago*, como su desinterés en relación al documental *Katrina*.
Una vez identificados los CRIDS de los cinco contenidos en la ontología de TV, el sistema incorpora a su perfil personal P_X dichos identificadores, referenciando (de forma indirecta) la información representada en Fig. 6.2 y Fig. 6.3, en las que se incluyen las instancias y clases de la ontología OWL, respectivamente.
Conviene destacar, tal como se aprecia en el perfil P_X, la posibilidad de que un mismo programa esté clasificado bajo varias categorías en la jerarquía de géneros de la ontología OWL (por ejemplo, *Top Gun* pertenece al género dramático (3.4.1), al romántico (3.4.3) y al género de acción (3.4.6.2)), así como el hecho de que una misma característica semántica esté asociada a varios de los contenidos de TV almacenados en el mismo (por ejemplo, el actor *Tom Cruise* interviene tanto en la película *Top Gun* como en *Cocktail*).
Tal como adelantábamos antes, esta última propiedad permite al sistema identificar fácilmente las preferencias positivas del espectador. En nuestro ejemplo concreto, AVATAR puede inferir el interés del usuario U_X en relación al actor *Tom Cruise*, la actriz *Helen Hunt* o los géneros *dramático* y *romántico*, entre otros, precisamente por estar estos últimos relacionados con varios de los contenidos que resultaron atractivos para el usuario en el pasado.

Una vez construido el perfil inicial del usuario, el sistema puede empezar a elaborar recomendaciones personalizadas basadas en las preferencias definidas en el mismo. Una vez mostradas estas sugerencias, el espectador tiene la posibilidad de aceptar o rechazar los contenidos seleccionados, así como de especificar, opcionalmente, niveles de interés concretos en relación a cada uno de ellos. Esta realimentación de relevancia (ver Sección 2.4) proporcionada por el usuario —implícita en el primer caso y explícita en el segundo— es aprovechada por el sistema para actualizar su *perfil-ontología*. Así, de forma análoga a la comentada al principio de la sección, en este escenario el sistema añade al perfil todos aquellos programas que el usuario haya aceptado o rechazado, así como los índices relativos a dichos contenidos, a sus características semánticas y a la jerarquía de géneros correspondiente. Como consecuencia, el sistema construye los perfiles de los usuarios de forma incremental, a medida que recibe conocimiento adicional sobre sus preferencias personales.

Dado que la realimentación de relevancia explícita no es una exigencia para los

[2]En adelante, siempre que aludamos a las clases e instancias definidas en los perfiles de usuario, nos referimos a las entidades referenciadas desde los mismos y almacenadas únicamente en la ontología OWL.

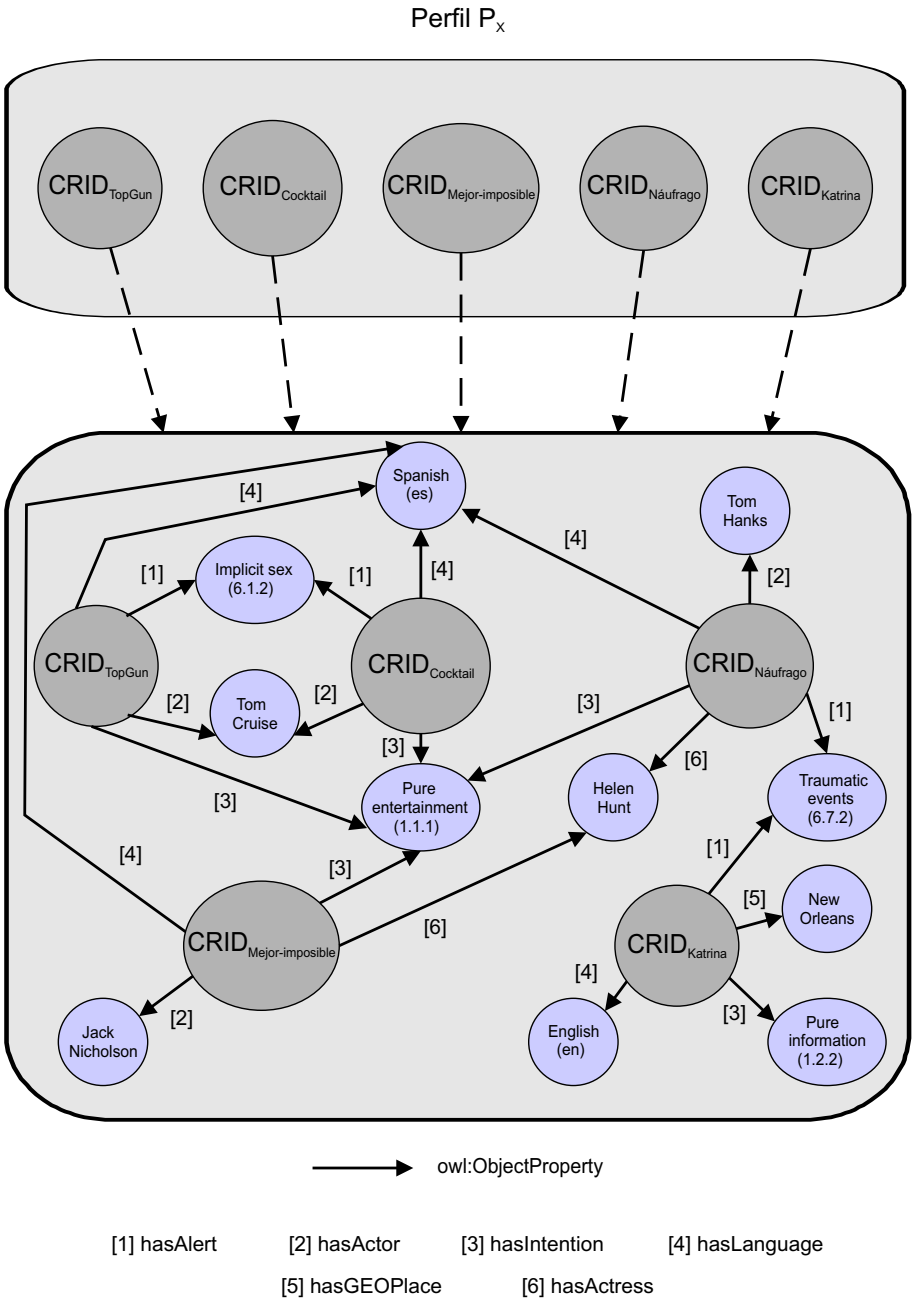

Figura 6.2: Instancias relativas al *perfil-ontología* P_X

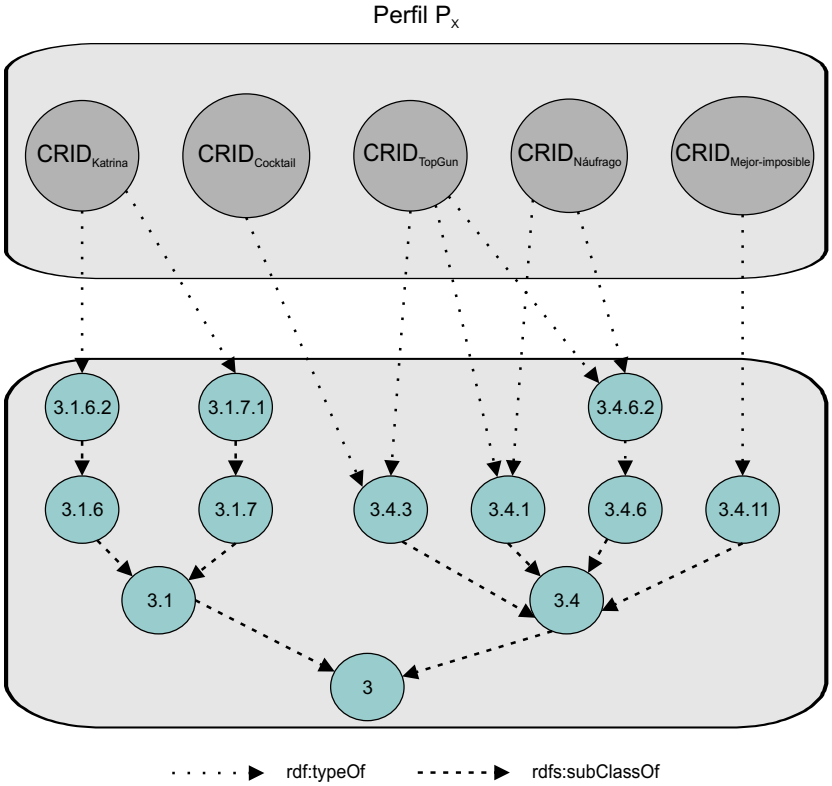

Figura 6.3: Clases relativas al *perfil-ontología* P_X

usuarios de nuestro sistema, es necesario recurrir a algún mecanismo que cuantifique el nivel de interés que tiene cada espectador en relación a los programas registrados en su perfil, distinguiendo así las preferencias positivas de las negativas.

6.3 Cálculo de los niveles de interés

Nuestra aproximación asigna un índice numérico, llamado DOI (*Degree Of Interest*), tanto a las instancias referenciadas desde el *perfil-ontología* (programas de TV y características semánticas), como a aquellas clases que identifican los géneros representados en la Fig. 5.1. En concreto, los valores negativos de dicho índice (en el rango [-1,0)) identifican las preferencias negativas del usuario, correspondiendo el valor -1 al nivel de desinterés más significativo; en relación a las clasificaciones positivas, el usuario puede expresar un nivel de interés máximo mediante el valor +1.

Con el propósito de asegurar los requisitos de interacción mínima entre el usuario y el sistema, nuestro enfoque no exige a los espectadores que declaren explícitamente sus niveles de interés, sino que dichos valores son inferidos por el recomendador tomando como referencia las acciones que éstos llevan a cabo.

Nuestra aproximación utiliza los niveles de interés inferidos para cada programa, para inicializar tanto los correspondientes a sus respectivas características semánticas como los relativos a los géneros de la Figura 5.1 a los que estos pogramas pertenecen.

En concreto, el enfoque calcula los índices DOI de cada una de las clases hoja y, a continuación, propaga estos valores por toda la jerarquía de géneros, hasta alcanzar el nodo raíz de la misma.

6.3.1 Nivel de interés de los contenidos de TV

El comportamiento de visionado del usuario es el elemento clave a la hora de cuantificar su nivel de interés en relación a cada uno de los programas registrados en su perfil. En relación a estos últimos, podemos identificar dos escenarios de aplicación diferentes. Uno de ellos está vinculado a la fase de registro del espectador en el sistema, y el otro, a la realimentación de relevancia proporcionada por el mismo ante las recomendaciones de AVATAR.

En el primer supuesto, es posible que los usuarios no sólo identifiquen programas concretos que le han interesado (o no) en el pasado, sino también clasificaciones explícitas para los mismos. Si el usuario declara estas clasificaciones, nuestro enfoque incorpora a su perfil estos contenidos, e inicializa sus respectivos índices DOI con dichos valores. En caso contrario, asumimos que el (des)interés del usuario en relación a los programas identificados es máximo, de ahí que los índices DOI asignados sean 1 y -1 para las preferencias positivas y negativas, respectivamente.

En el segundo de los escenarios planteados, el sistema debe inferir el nivel de interés de los usuarios a partir de la realimentación de relevancia implícita que éstos proporcionan cuando el sistema les ofrece una recomendación. En este caso, parámetros como la aceptación o el rechazo del espectador ante los programas sugeridos o la proporción de los mismos que éste ha visto, suponen un aporte de información relevante en el proceso de inferencia. Dichos parámetros se incluyen en el modelo del usuario mediante los índices presentados a continuación:

- **Índice de realimentación** (IOF- *Index Of Feedback*): Este factor permite que el espectador evalúe cualitativamente los programas que ha aceptado o rechazado en la recomendación ofrecida por el sistema[3]. En concreto, nuestro enfoque permite identificar si dichos contenidos son *muy interesantes*, *interesantes*, *poco interesantes* o *muy poco interesantes*. Para cada una de estas valoraciones, el índice IOF toma los valores 1, 0.5, -0.5 y -1, respectivamente. Si el usuario simplemente selecciona o rechaza un programa sin proporcionar esta evaluación, los valores del índice son 0.75 y -0.75, respectivamente.
- **Índice de visionado** (IOV- *Index Of Viewing*): El valor de este índice es la relación entre el tiempo que invierte el usuario en ver un contenido recomendado por el sistema y su duración total.

Reuniendo las contribuciones de ambos parámetros resulta la Ec. 6.1, en la que se cuantifica el nivel de interés del contenido p recomendado a un usuario U_X, cuyo perfil es P_X:

$$DOI_{P_X}(p) = \begin{cases} IOV_{P_X}(p) \cdot IOF_{P_X}(p) & \text{si } U_X \text{ acepta el programa } p \\ IOF_{P_X}(p) & \text{en otro caso} \end{cases} \tag{6.1}$$

[3]Conviene resaltar que el sistema AVATAR acompaña cada programa sugerido de una descripción textual que permite que el usuario pueda valorar (aceptar o rechazar) el contenido aunque no lo conozca.

Ejemplo: Consideremos de nuevo el usuario U_X cuyo perfil se representa en la Fig. 6.2. En este ejemplo, asumiremos que los programas definidos en P_X no han sido explícitamente declarados por el usuario durante la fase de registro, sino que han sido recomendados por nuestro sistema. De esta forma, podemos evaluar el efecto de los índices descritos en esta sección y explicar cómo AVATAR infiere un nivel de interés concreto para cada uno de los contenidos considerados, a partir de las acciones que lleva a cabo este espectador. Los valores de los citados índices aparecen recogidos en la Tabla 6.1. Desde estos resultados, es fácil apreciar la influencia de cada parámetro. En concreto, en este escenario se asume que todos los programas sugeridos han sido aceptados por el usuario U_X, salvo el documental *Katrina*. A excepción de la película *Cocktail*, el espectador ha declarado realimentación de relevancia explícita para todos los programas mostrados en la sugerencia, de ahí que los índices IOF de los contenidos aceptados tomen valor 1 y el correspondiente al documental rechazado, valor -1. Finalmente, los valores de los factores IOV revelan que el espectador ha visto el 40 % de la película *Mejor imposible*, y el contenido íntegro de los restantes programas. La aplicación de la Ec. 6.1 conduce de forma inmediata a los niveles de interés definidos en la Tabla 6.1.

Contenidos de TV	IOF	IOV	DOI
Mejor imposible	1	0.4	0.4
Náufrago	1	1	1
Top Gun	1	1	1
Cocktail	0.75	1	0.75
Katrina	-1	-	-1

Tabla 6.1: Niveles de interés de los contenidos definidos en el perfil P_X (Fig. 6.2)

6.3.2 Nivel de interés de las características semánticas

Conviene destacar que la clasificación calculada mediante la Ec. 6.1 no sólo se asigna al programa p definido en el perfil del usuario U_X, sino también a cada una de sus características semánticas.

Tal como mencionamos en la Sección 6.2.2, es común en el dominio de la TV que este tipo de atributos estén asociados a varios programas a la vez. Por esta razón, es posible que algunas de las características del contenido p cuyo CRID va a registrar el sistema en P_X, estén ya asociadas a otros programas de TV incluidos anteriormente en dicho perfil. En este caso, tales características tienen ya asignado un índice DOI inicial en P_X que será promediado con el calculado mediante la Ec. 6.1, obteniendo así su nivel de interés final.

Ejemplo: Considerando de nuevo el perfil P_X representado en la Fig. 6.2, detallamos el cálculo de los niveles de interés correspondientes a las características semánticas de los contenidos definidos en el mismo. Los resultados aparecen resumidos en la Tabla 6.2. Merece la pena destacar que las características asociadas a un único contenido de TV heredan directamente su índice DOI. Por el contrario, tal como decíamos antes, los niveles de interés de aquellas características relacionadas con varios programas (por ejemplo, *Tom Cruise* y *Helen Hunt*) resultan del promedio de los índices DOI de cada uno de éstos.

Características semánticas	Contenidos de TV	Índices DOI
Jack Nicholson	*Mejor imposible*	0.4
Helen Hunt	*Mejor imposible* *Náufrago*	0.5 · (0.4 + 1) = 0.7
Tom Hanks	*Náufrago*	1
Tom Cruise	*Top Gun* *Cocktail*	0.5 · (1+0.75)=0.875
New Orleans	*Katrina*	-1
Entertainment *(1.1.1)*	*Mejor imposible* *Náufrago* *Top Gun* *Cocktail*	0.25 · (0.4+1+1+0.75) = 0.788
Spanish (es)	*Mejor imposible* *Náufrago* *Top Gun* *Cocktail*	0.25 · (0.4+1+1+0.75) = 0.788
Implicit sex *(6.1.2)*	*Top Gun* *Cocktail*	0.5 · (1 + 0.75) = 0.875
Traumatic events *(6.7.2)*	*Náufrago* *Katrina*	0.5 · (1-1) = 0
Pure information *(1.2.2)*	*Katrina*	-1
English (en)	*Katrina*	-1

Tabla 6.2: Niveles de interés de las características semánticas incluidas en el perfil P_X

6.3.3 Nivel de interés de las clases hoja

Tal como describimos en el capítulo previo, nuestro enfoque permite que un mismo contenido de TV pertenezca a varias de las clases hoja representadas en la jerarquía de géneros establecida en la ontología OWL. En la metodología propuesta, cuando el sistema infiere el índice de interés de un programa concreto, los niveles de cada una de sus clases hoja se actualizan automáticamente. Para ello, el índice DOI de una clase hoja se calcula promediando las clasificaciones de todos los programas definidos en el perfil del usuario que pertenezcan a dicha categoría, tal como muestra la Ec. 6.2. En ésta, se asume que el perfil P_X contiene N programas clasificados bajo el género C_p, siendo p_i cada uno de los mismos.

$$DOI_{P_X}(C_p) = \frac{1}{N} \cdot \sum_{i=1}^{N} DOI_{P_X}(p_i) \tag{6.2}$$

Ejemplo: Al igual que en los ejemplos anteriormente descritos, en la Tabla 6.3 se muestran los niveles de interés correspondientes a cada una de las clases hoja referenciadas desde el perfil P_X de la Fig. 6.3.

Si bien la Ec. 6.2 conduce a un mecanismo de asignación de índices de interés

Clases hoja	Contenidos de TV	Índices DOI
General light drama (3.4.1)	*Top Gun* *Náufrago*	0.5 · (1+1) = 1
Comedy (3.4.11)	*Mejor imposible*	0.4
Romance (3.4.3)	*Top Gun* *Cocktail*	0.5 · (1+0.75) = 0.875
Disaster (3.4.6.2)	*Náufrago* *Top Gun*	0.5 · (1+1) = 1
Nature (3.1.6.2)	*Katrina*	-1
Reality (3.1.7.1)	*Katrina*	-1

Tabla 6.3: Niveles de interés de las clases hoja definidas en el perfil P_X (Fig. 6.3)

muy sencillo, es necesario revisar este procedimiento teniendo en cuenta algunas situaciones excepcionales. Para ello, imaginemos el escenario de aplicación descrito a continuación. Un usuario tiene en su perfil una clase con un índice DOI muy elevado (próximo a uno) y, de repente, rechaza programas de esa categoría. Es evidente que esta clasificación debería ser fuertemente decrementada, con el objetivo de evitar que el sistema incurra en nuevos errores en las próximas recomendaciones al sugerir programas de este género.

El criterio elegido en AVATAR a la hora de resolver este tipo de conflictos se basa en utilizar un umbral, de forma que todas aquellas clases con niveles de interés superiores a éste, y cuyos programas sean rechazados por el usuario, experimentarán una reducción en sus índices DOI. En concreto, el nuevo nivel de interés de dichas clases se calcula promediando los correspondientes a aquellas clases del perfil cuyos índices DOI son positivos y menores que el umbral establecido. Por su naturaleza, el valor de este umbral debe ser elevado (próximo a 1) y no necesariamente dependiente del dominio de aplicación del sistema[4].

Ejemplo: Tal como muestran los niveles de interés representados en la Tabla 6.3, el usuario U_X parece estar muy interesado en los contenidos de ficción pertenecientes al género dramático (3.4.1). Supongamos que el sistema sugiere un contenido perteneciente a esta categoría ontológica. En este escenario, si el usuario rechaza el programa recomendado, la técnica de modelado propuesta decrementaría su índice de interés hasta alcanzar un valor DOI_{P_X} (3.4.1) = 0.4, ya que, de acuerdo al criterio propuesto y a los niveles de interés mostrados en la Tabla 6.3, el único género que, teniendo un índice de interés positivo, no supera el umbral $0{,}85$ establecido, es el correspondiente a las comedias de ficción (3.4.11).

Nuestro criterio está basado en el promedio de los niveles de interés, debido a que esta función permite igualar el índice DOI de la clase rechazada frente a otros géneros que también interesan al usuario. Así, si el espectador sigue rechazando programas de esta clase, su índice DOI seguirá decreciendo y pasará a formar parte de sus preferencias negativas. Por el contrario, si el rechazo del espectador es sólo un hecho puntual

[4]En los experimentos realizados mediante la herramienta descrita en el Apéndice B, hemos utilizado un umbral en torno a $0{,}85$.

y, por tanto, sigue estando interesado en los contenidos de esta clase hoja, su nivel de interés puede incrementarse de nuevo, hasta volver a posicionar esta categoría entre las favoritas del usuario.

Por otra parte, merece la pena mencionar también la situación contraria a la descrita anteriormente; es decir, aquella en la que el usuario acepta programas pertenecientes a categorías que no le han interesado en el pasado y, que, por ello, registran índices de interés negativos en su perfil. Este hecho no se considera una situación excepcional en nuestro enfoque, debido a que, en este caso, el sistema no selecciona programas de géneros que podrían no interesar al usuario, como ocurría en el escenario descrito antes. De hecho, si el espectador continúa aceptando contenidos de esta categoría, su índice DOI se irá incrementando de acuerdo a la Ec. 6.2, hasta pasar a forma parte de sus preferencias positivas; en ese momento, el sistema ya puede incorporar instancias concretas de dicha clase a las recomendaciones elaboradas. Por tanto, este escenario únicamente provoca que los programas clasificados bajo algunos géneros sean sugeridos al usuario con cierto retraso; sin embargo, no decrementa la precisión de las recomendaciones del sistema.

Una vez cuantificado el índice DOI de las clases hoja definidas en el perfil —ya sea mediante la Ec. 6.2 o bien a través del criterio especificado antes—, ese valor se propaga por la jerarquía representada en la Fig. 5.1, tal como se describe a continuación.

6.3.4 Propagación de los niveles de interés a través de la jerarquía

Nuestra metodología adopta el enfoque propuesto por Cai N. Ziegler en su recomendador para la Web Semántica (descrito en la Sección 3.10.3), a la hora de propagar el nivel de interés de una clase hoja C_p a través de la jerarquía de géneros que identifica las preferencias del usuario. Tal como muestra la Ec. 6.3, los mayores niveles de interés corresponden a aquellas superclases más específicas en la jerarquía, es decir, más cercanas a las categorías hoja de la misma. Asimismo, el índice DOI de una superclase C_{p+1} será más significativo cuanto mayor sea el índice de interés de la clase C_p que propaga su nivel, y cuanto menor sea el número de hermanos de ésta.

$$DOI_{P_X}(C_{p+1}) = \frac{DOI_{P_X}(C_p)}{1 + \#sib(C_p)} \tag{6.3}$$

La Ec. 6.3 sólo permite cuantificar el nivel de interés que una clase propaga a su superclase directa en la jerarquía. Dado que de esta última pueden depender varias subclases, es necesario reunir las contribuciones de todas ellas para así cuantificar su índice DOI final. En el enfoque propuesto, el nivel de una superclase se calcula promediando los índices que propagan (mediante la Ec. 6.3) cada una de las categorías que dependen directamente de ella en la jerarquía de géneros.

Ejemplo: Suponiendo la taxonomía de clases representada en la Fig. 5.1, recogemos en la Tabla 6.4 los niveles de interés que cada clase hoja del perfil P_X propaga a sus superclases en esta estructura jerárquica. Para cada una de estas clases, representadas en la Fig. 6.3, la tabla muestra las contribuciones recibidas de sus subclases directas en la jerarquía, finalmente promediadas en su índice DOI.

Clases	Contribución de subclases	DOI
Sciences (3.1.6)	$\frac{DOI_{P_X}(3.1.6.2)}{1+\#sib(3.1.6.2)} = \frac{-1}{1+13} = -0{,}071$	-0.071
Human interest (3.1.7)	$\frac{DOI_{P_X}(3.1.7.1)}{1+\#sib(3.1.7.1)} = \frac{-1}{1+6} = -0{,}142$	-0.142
Non-fiction (3.1)	$\frac{DOI_{P_X}(3.1.6)}{1+\#sib(3.1.6)} = \frac{-0{,}071}{1+8} = -0{,}0079$ $\frac{DOI_{P_X}(3.1.7)}{1+\#sib(3.1.7)} = \frac{-0{,}142}{1+8} = -0{,}016$	-0.012
Action (3.4.6)	$\frac{DOI_{P_X}(3.4.6.2)}{1+\#sib(3.4.6.2)} = \frac{1}{1+12} = 0{,}077$	0.077
Fiction (3.4)	$\frac{DOI_{P_X}(3.4.1)}{1+\#sib(3.4.1)} = \frac{1}{1+18} = 0{,}052$ $\frac{DOI_{P_X}(3.4.3)}{1+\#sib(3.4.3)} = \frac{0{,}875}{1+18} = 0{,}046$ $\frac{DOI_{P_X}(3.4.6)}{1+\#sib(3.4.6)} = \frac{0{,}077}{1+18} = 0{,}004$ $\frac{DOI_{P_X}(3.4.11)}{1+\#sib(3.4.11)} = \frac{0{,}4}{1+18} = 0{,}021$	0.031

Tabla 6.4: Propagación de niveles de interés en la jerarquía de géneros de P_X (Fig. 6.3)

En un entorno de personalización como el asociado a AVATAR, son obvias las necesidades de adaptación del sistema recomendador ante los posibles cambios que se produzcan en las preferencias de los usuarios. Con el afán de asegurar tal objetivo, nuestra metodología incorpora una función de olvido gradual que permite que la herramienta identifique las preferencias actuales del usuario.

6.4 Adaptación ante cambios en las preferencias de los usuarios

Tal como describimos en la Sección 2.5.3, las llamadas funciones de olvido gradual proporcionan a los recomendadores mecanismos muy sencillos para poder adaptar sus sugerencias, rápida y cómodamente, a medida que cambian las preferencias de los usuarios. Dichas funciones se basan en asignar un índice numérico de valor 1 (denominado índice de olvido o FI, de *Forgetting Index*) a cada nueva preferencia añadida al perfil del usuario, a la vez que se decrementan los valores correspondientes a los intereses almacenados previamente en el mismo. De esta forma, el sistema puede identificar fácilmente información actual sobre el usuario, sin más que seleccionar aquellas

preferencias cuyos pesos numéricos son más significativos.

6.4.1 Función de olvido gradual basada en ventanas temporales

En concreto, nuestra metodología incorpora una función de olvido gradual basada en ventanas temporales. En este escenario, el enfoque únicamente considera, durante el proceso de elaboración de las recomendaciones, aquellas instancias cuyos índices de olvido FI están dentro de una ventana con un ancho limitado Ω.

La característica más destacable de esta ventana, representada en la Fig. 6.4, es que Ω es un parámetro ajustado dinámicamente en función de la precisión del recomendador en las sugerencias previamente elaboradas para el usuario, definida como la proporción de contenidos sugeridos por la herramienta y aceptados por éste.

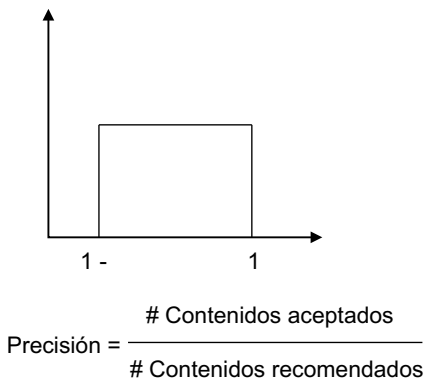

Figura 6.4: Ventana temporal utilizada en la función de olvido gradual

De esta forma, cuando el sistema detecta que su precisión se ha reducido, automáticamente estrecha la ventana temporal, considerando únicamente aquellas instancias referenciadas en el perfil cuyos índices FI toman un valor cercano a 1 (es decir, las más recientes). Por el contrario, si la precisión es alta, la ventana permite que el sistema pueda explorar gran parte del conocimiento disponible sobre las preferencias del usuario, incluyendo un número mayor de instancias durante el proceso de elaboración de recomendaciones.

Este mecanismo de redimensionamiento dinámico es especialmente interesante en nuestra metodología. Es obvio que un índice de olvido poco significativo no tiene por qué identificar necesariamente preferencias no válidas para un espectador, sino únicamente aquellas almacenadas en su perfil hace ya algún/mucho tiempo. Precisamente por esta razón, una ventana temporal de ancho fijo no parece una elección acertada, ya que, en ese caso, las instancias asociadas a índices reducidos dejarían de ser representativas para el sistema y, en consecuencia, serían omitidas durante el proceso de personalización. Por el contrario, nuestra metodología utiliza de forma eficiente el conocimiento aprendido sobre cada usuario. Así, gracias al proceso de ajuste dinámico propuesto, las instancias mencionadas antes volverán a ser consideradas por el sistema cuando se incremente la precisión (y, en consecuencia, se ensanche la ventana temporal). Si dichas preferencias son válidas para el espectador, la tasa de aciertos del sistema no se verá decrementada; en caso contrario, la ventana se estrechará y el recomenda-

dor trabajará únicamente sobre los intereses más recientes del usuario, adaptándose de forma flexible a su comportamiento y a sus intereses personales.

6.4.2 Actualización de los índices de olvido

Tal como hemos adelantado antes, cuando el sistema incorpora un nuevo CRID al perfil del usuario inicializa a 1 tanto el índice de olvido de este contenido, como el de sus respectivas características semánticas, a la vez que decrementa los índices asociados al resto de preferencias registradas en el mismo[5]. En concreto, siendo i_j cada una de las instancias referenciadas desde dicho perfil, nuestro enfoque decrementa su índice de olvido en una proporción dependiente de la precicisión del sistema, tal como muestra la Ec. 6.4.

$$FI_{P_X}(i_j) = \begin{cases} FI_{P_X}(i_j) - \frac{1}{1+\Omega} & \text{si } FI_{P_X}(i_j) - \frac{1}{1+\Omega} \geq 0 \\ 0 & \text{en otro caso} \end{cases} \tag{6.4}$$

Así, cuando la precisión del sistema es elevada, el sistema tiende a recordar todas las preferencias que conoce sobre un usuario. En este escenario, la ventana temporal tiene un ancho considerable, dando cabida a gran parte de las instancias definidas en el perfil del usuario, cuyos niveles de olvido habrán sufrido decrementos poco significativos. Por el contrario, cuando se aprecia una pérdida de precisión, el sistema no sólo reduce la ventana temporal, sino que además decrementa fuertemente los índices de olvido de las preferencias del perfil, permitiendo que sólo las más recientes puedan ser consideradas durante el proceso de recomendación.

Dado que los índices de olvido de las instancias referenciadas en el perfil son decrementadas continuamente, es posible que alcancen un valor nulo, tal como muestra la Ec. 6.4. En este caso, el sistema podría considerar que identifican preferencias obsoletas y eliminarlas del perfil; sin embargo, tal como apuntábamos en la sección previa, es posible que algunas de ellas continúen siendo válidas para el usuario. Por esta razón, nuestra metodología únicamente elimina del perfil la información sobre aquellas instancias que no son suficientemente significativas para el proceso de personalización. El criterio elegido a la hora de hacer tal selección considera que las preferencias que aportan más conocimiento al recomendador son aquellas que tienen asociados niveles de interés muy positivos o muy negativos, ya que éstas identifican tanto los contenidos (y atributos) que resultan más atractivos para el usuario, como aquellos que no le interesan en absoluto. Precisamente para poder prescindir de aquellas preferencias con niveles de interés moderados, nuestra metodología incorpora un umbral ϵ; así, todos aquellos índices de olvido nulos que estén asociados a instancias con niveles de interés menores que dicho umbral, son eliminados del perfil del usuario de forma permanente. Tal como muestra la expresión 6.5, nuestro enfoque también elimina los índices de interés de dichas instancias y, como consecuencia, estas últimas dejan de identificar las preferencias actuales del usuario activo.

$$\textbf{Si } (FI_{P_X}(i_j) = 0) \textbf{ Y } (|DOI_{P_X}(i_j)| < \epsilon) \textbf{ Entonces } \text{eliminar índices de } i_j \tag{6.5}$$

[5]En este sentido, destacar que si una característica semántica ya estaba definida en el perfil y, a la vez, también está asociada al programa añadido al mismo, su índice de olvido no se decrementa, sino que se inicializa a 1.

Tipo de instancia	Situación en el perfil	Índice de olvido FI	Índice DOI
Contenido de TV	Añadido al perfil	1	f (IOF, IOV).
	Almacenado en el perfil	Restar $\frac{1}{1+\Omega}$	
Característica semántica	Asociada al programa que se añade al perfil	1	Promediar DOI de los programas (con los que está relacionada) con índices FI en la ventana actual.
	Asociada a otros programas almacenados en el perfil	Restar $\frac{1}{1+\Omega}$	

Tabla 6.5: Efecto de la ventana temporal sobre los índices DOI de las instancias del perfil del usuario

Dado que los índices de interés DOI oscilan entre -1 y 1, el umbral ϵ debe tomar un valor elevado para así satisfacer el criterio de selección de instancias mencionado antes. Si bien nuestra metodología emplea el mismo valor para todos los usuarios del sistema, creemos que éste debería ser ajustado dinámicamente a lo largo del tiempo, consiguiendo con ello que la cantidad de información eliminada del perfil varíe en función de las capacidades de almacenamiento disponibles en el servidor del proveedor de servicios (ver Fig. 4.4).

De acuerdo a la expresión 6.5, el valor de ϵ podría incrementarse cuando sea necesario liberar una gran cantidad de información en el servidor, ya que en este escenario, el sistema filtraría más instancias con índices de olvido nulo. Por el contrario, el sistema podría eliminar menos información del perfil del usuario, y con ello considerar más conocimiento sobre las preferencias del usuario, siempre que la capacidad de almacenamiento de dicho servidor así lo permita. De lo comentado se desprende que el umbral ϵ depende claramente del entorno de ejecución de cada sistema recomendador en el que se aplique la metodología de razonamiento propuesta en este trabajo[6].

Una vez redimensionada la ventana temporal —considerando la tasa de aciertos obtenida en la última recomendación mostrada al usuario— y actualizados los índices de olvido, nuestra metodología identifica el conjunto de instancias que ya no representan las preferencias actuales del usuario. El conocimiento que el sistema debe olvidar está codificado en los niveles de interés definidos en el perfil, de ahí que sea necesario actualizar los índices DOI registrados en el mismo, considerando que parte de la información utilizada para calcular su valor actual ya no es válida en este momento.

[6]En los experimentos realizados en el sistema AVATAR hemos utilizados un umbral ϵ en torno a 0.8.

Este proceso de actualización aparece resumido en sendas tablas, donde describimos cómo se calculan los índices de olvido gradual, y cómo afectan éstos al cómputo de los niveles de interés una vez revisado el tamaño de la ventana temporal considerada.

Tal como se aprecia en la Tabla 6.5, los índices DOI de los programas añadidos al perfil de un usuario no experimentan cambio alguno tras redimensionar la ventana temporal. Así, su valor se calcula a partir de los índices de realimentación (IOF) y visionado (IOV), tal como se describió en la Sección 6.3.1. Por el contrario, los índices de las características semánticas incluidos en los perfiles sí pueden experimentar ciertas alteraciones. En concreto, tal como mencionamos en la Sección 6.3.2, los niveles de interés de las características semánticas asociadas a varios programas en un mismo perfil, se calculan promediando los índices DOI de cada uno de éstos. Una vez redimensionada la ventana temporal, dicho promedio considera únicamente aquellos contenidos cuyos índices de olvido estén incluidos en la misma, eliminando con ello la contribución de las preferencias obsoletas. Además, si la característica semántica no estaba ya definida en el perfil del usuario, hereda el nivel de interés del contenido que se añade al mismo y se inicializan sus respectivos índices de olvido a 1.

Un procedimiento análogo al que acabamos de describir se aplica sobre las clases hoja referenciadas desde el perfil del usuario. Tal como muestra la Tabla 6.6, en este caso su índice DOI se calcula considerando únicamente los niveles de interés de todos los programas de dicha clase que pertenezcan a la ventana temporal actual.

Tipo de clases	Índice de olvido FI	Niveles de interés DOI
Clases hoja	No existe	Promediar DOI de las instancias con índices FI en la ventana actual.
Clases restantes	No existe	Propagar DOI desde las clases hoja mediante Ec. 6.3.

Tabla 6.6: Efecto de la ventana temporal sobre los índices DOI de las clases del perfil

Ejemplo: Supongamos que, tras dimensionar la ventana temporal correspondiente al usuario U_X, el sistema descubre que el índice FI de la película *Top Gun* almacenada en su perfil P_X no está incluida en la misma. En este caso, los niveles de interés de todas aquellas características semánticas que esta película comparte con otros contenidos del perfil (*Tom Cruise*, *Implicit sex* (6.1.2), *Pure entertainment* (1.1.1) y *Spanish* (es)) deberían ser modificadas. Tal como muestra la Tabla 6.7, el índice DOI de *Tom Cruise* se calcularía considerando únicamente la película *Cocktail*, ya que ésta, al pertenecer a la ventana temporal, sí identifica preferencias actuales del usuario U_X. El mismo razonamiento es aplicable al resto de las características semánticas mencionadas.
De forma análoga, el nivel de interés de aquellas clases hoja a las que pertenezca la película *Top Gun* (3.4.1, 3.4.3 y 3.4.6.2 en la Tabla 6.3) también deberían ser actualizadas. En este caso, en el promedio realizado para calcular sus respectivos índices DOI se suprimiría la contribución de este contenido.
Finalmente, destacar que al variar el nivel de interés de dichas clases hoja, también lo harán los índices DOI de sus respectivas superclases en la jerarquía de géneros (3.4.6 y 3.4). Todos los valores mencionados aparecen representados en la Tabla 6.7.

Instancias/Clases	Contribuciones	DOI
Tom Cruise	*Cocktail*	0.75
Implicit sex (6.1.2)	*Cocktail*	0.75
Entertainment (1.1.1)	*Mejor imposible* *Náufrago* *Cocktail*	$\frac{0,4+1+0,75}{3} = 0,72$
Spanish (es)	*Mejor imposible* *Náufrago* *Cocktail*	$\frac{0,4+1+0,75}{3} = 0,72$
General light drama (3.4.1)	*Náufrago*	1
Romance (3.4.3)	*Cocktail*	0.75
Disaster (3.4.6.2)	*Náufrago*	1
Action (3.4.6)	$\frac{DOI_{P_X}(3.4.6.2)}{1+\#sib(3.4.6.2)} = \frac{1}{1+12} = 0,077$	0.077
Fiction (3.4)	$\frac{DOI_{P_X}(3.4.1)}{1+\#sib(3.4.1)} = \frac{1}{1+18} = 0,052$ $\frac{DOI_{P_X}(3.4.3)}{1+\#sib(3.4.3)} = \frac{0,75}{1+18} = 0,039$ $\frac{DOI_{P_X}(3.4.6)}{1+\#sib(3.4.6)} = \frac{0,077}{1+18} = 0,004$ $\frac{DOI_{P_X}(3.4.11)}{1+\#sib(3.4.11)} = \frac{0,4}{1+18} = 0,021$	0.029

Tabla 6.7: Actualización de los niveles de interés del perfil P_X

6.4.3 Eliminación de preferencias obsoletas

Tal como hemos descrito en la sección previa, una de las características más ventajosas de nuestro proceso de ajuste dinámico de la ventana temporal es que, a la vez que permite identificar rápidamente las preferencias más recientes del usuario, no conlleva una eliminación permanente de los intereses menos recientes. Si bien, con ello el recomendador puede utilizar de forma eficiente todo el conocimiento que ha aprendido sobre las preferencias del usuario, no es menos cierto que los requisitos de capacidad de almacenamiento requeridos en el servidor del sistema (ver Fig. 4.4), pueden llegar a ser críticos si estos perfiles alcanzan tamaños demasiado grandes.

A la vista de tal limitación, nuestra metodología propone un mecanismo basado en seleccionar todos aquellos *perfiles-ontología* que alcancen un determinado tamaño umbral, y procesarlos de acuerdo a las siguientes reglas:

- Todos los programas de TV definidos en los perfiles seleccionados cuyos índices de olvido no estén incluidos en la ventana temporal considerada, y cuyos niveles

de interés sean menores (en valor absoluto) que el umbral ϵ introducido en la Sección 6.4.2, son eliminados de forma permanente.

- Los índices de olvido e interés de las características semánticas de dichos contenidos también son eliminados, siempre que no estén asociadas a otros programas almacenados en estos perfiles.
- Finalmente, también se suprimen de los *perfiles-ontología* los índices de interés de todas aquellas clases (géneros) para las que no existan instancias concretas, como resultado del proceso de eliminación de información aquí descrito.

El tamaño umbral de los perfiles es un parámetro claramente dependiente del contexto de ejecución del sistema recomendador en el que se utilice la técnica de modelado de usuarios propuesta. Por su naturaleza, su valor debe aumentarse si tanto el número de usuarios registrados como la cantidad de información sobre sus preferencias es moderada, y disminuirse cuando algunos de estos parámetros se incremente de forma considerable. De acuerdo a lo mencionado en la Sección 6.4.2, en este escenario, nuestra metodología también ajustaría dinámicamente el umbral ϵ, disminuyendo su valor en el primer caso e incrementándolo en el segundo.

Gracias al conocimiento formalizado en la ontología del sistema, la técnica de modelado de usuarios descrita en este capítulo aporta descripciones (semánticas) muy detalladas sobre sus preferencias, aprovechadas, de forma efectiva, en las estrategias de personalización incluidas en la metodología.

Parte IV

Estrategias de recomendación

Capítulo 7

Estrategia basada en asociaciones semánticas

Este capítulo describe la primera de las estrategias de personalización incluidas en la metodología desarrollada en esta tesis, en la que se combinan los principios del filtrado tradicional basado en contenido con procesos de razonamiento semántico. Como resultado de dicha combinación, la estrategia sugiere a cada usuario productos que están relacionados semánticamente con aquellos que le interesaron en el pasado. Con el propósito de enriquecer el conocimiento disponible sobre las preferencias del usuario, nuestro trabajo considera tanto las relaciones representadas en la ontología como otras que están ocultas en la misma. Para descubrir tales relaciones, la estrategia recurre a mecanismos que exploran, de forma exhaustiva y eficiente, las entidades y propiedades formalizadas en la base de conocimiento OWL. Estos mecanismos garantizan tanto la relevancia de las relaciones inferidas, como la adaptación del proceso de razonamiento a medida que las preferencias del usuario evolucionan en el tiempo. Son precisamente estas capacidades inferenciales las que permiten que la estrategia mejore la calidad de las sugerencias seleccionadas y, a la vez, combata limitaciones reconocidas en los métodos basados en contenido que, hasta ahora, eran un problema abierto en la literatura. Junto a los detalles algorítmicos, el capítulo presenta un ejemplo de aplicación de la estrategia en el contexto del sistema AVATAR, y describe sus ventajas más destacables.

7.1 Introducción

7.1.1 Descripción genérica

Tal como describimos en la Sección 2.6.2, los recomendadores que utilizan el filtrado basado en contenido sugieren a los usuarios productos (contenidos de TV en AVATAR) similares a los que le interesaron en el pasado. Las métricas de similitud adoptadas tradicionalmente —basadas en clasificadores automáticos en los enfoques más sofisticados— seleccionan productos excesivamente parecidos a los que el usuario ya conoce, debido a que sólo consideran los valores que toman sus respectivos atributos o características en un conjunto de entrenamiento predefinido, prescindiendo de la semántica asociada a los mismos. El carácter *sobreespecializado* de tales sugerencias es el motivo que ha llevado a muchos recomendadores a abandonar los métodos basados

en contenido en favor del filtrado colaborativo y los sistemas híbridos.

Sin embargo, junto a este tipo de limitaciones, el filtrado basado en contenido también presenta ciertas características muy interesantes en un sistema recomendador. Por su propia naturaleza, es una estrategia precisa, que reproduce fielmente las preferencias del usuario en las recomendaciones elaboradas; además, a diferencia de los enfoques colaborativos e híbridos, los sistemas basados en contenido no necesitan disponer de las clasificaciones de muchos usuarios para poder ofrecer sugerencias personalizadas a un usuario dado, hecho que origina muchas de las limitaciones reconocidas en dichos enfoques (ver Sección 2.6.3).

Con el ánimo de aprovechar los beneficios comentados, nuestra metodología incorpora el filtrado basado en contenido como estrategia de recomendación, aprovechando los procesos de razonamiento semántico para combatir, de forma efectiva, la *sobreespecialización* de sus sugerencias, un problema que aún no ha encontrado solución en la literatura. Estos procesos se sirven de las descripciones semánticas de los contenidos de TV formalizados en la ontología OWL para, más allá de realizar simples comparaciones entre sus valores concretos, descubrir relaciones que ayuden a diversificar los programas sugeridos a cada usuario. En concreto, nuestro enfoque basado en contenido sugiere programas *relacionados con* las preferencias del usuario, en vez de *similares a* dichos intereses.

Con el claro objetivo de descubrir la mayor cantidad posible de conocimiento útil sobre las preferencias del usuario, nuestra metodología de razonamiento no sólo considera los enlaces jerárquicos y las propiedades explícitamente formalizadas en la ontología OWL, sino que explora la base de conocimiento del sistema, tratando de descubrir —a partir de sus clases, propiedades e instancias— relaciones semánticas ocultas en la misma. La utilización de este tipo de relaciones en un entorno de personalización redunda en la calidad de las recomendaciones ofrecidas a los usuarios, ya que, por una parte, permiten descubrir correspondencias útiles entre sus preferencias y los contenidos sugeridos que pasarían inadvertidas en los enfoques actuales y, por otra, combaten de forma efectiva gran parte de las limitaciones identificadas en los mismos.

Tal como describimos en el capítulo 3, Anyanwu y Sheth ya contemplaron este tipo de relaciones —a las que denominaron asociaciones semánticas complejas— en el contexto del proyecto SemDis. Dado que este proyecto no considera los requisitos de personalización presentes en los sistemas recomendadores, nuestro trabajo debe adaptar la clasificación de asociaciones propuesta por sus autores [5] a las necesidades de la metodología desarrollada. En concreto, nuestro enfoque utiliza únicamente algunas de estas asociaciones, prescindiendo de aquellas que, por ser más indirectas, no aportan conocimiento útil para la elaboración de las recomendaciones personalizadas.

Una vez identificadas las asociaciones semánticas consideradas en la metodología, es necesario definir el mecanismo de inferencia que permite trazar tales relaciones en la ontología del sistema. Este mecanismo no puede aprovechar la experiencia del proyecto SemDis, debido a que su propósito y el perseguido en nuestro trabajo son completamente diferentes. Tal como como describimos en la Sección 3.5, el objetivo en SemDis es descubrir las asociaciones semánticas establecidas entre dos instancias concretas que especifica el usuario, explorando para ello el conocimiento profundo de una ontología RDFS. Por el contrario, nuestra metodología de razonamiento debe descubrir contenidos que estén asociados semánticamente con las preferencias positivas del usuario, asegurando, a la vez, su relevancia para el mismo. Así, estableciendo un paralelismo con el paradigma de consulta propuesto por Anyanwu y Sheth, diremos

que nuestro trabajo sólo conoce una de las dos instancias especificadas en SemDis (las preferencias de los usuarios), debiendo descubrir la segunda de dichas instancias (los contenidos recomendados) y las asociaciones existentes entre ellas.

Para abordar este proceso de razonamiento semántico, nuestro enfoque identifica en la ontología de TV los programas que interesaron al usuario (registrados en su perfil con un nivel de interés positivo) y avanza, sucesivamente, a través de las secuencias de propiedades que permiten relacionar dichos contenidos con otras instancias de la base de conocimiento. Para garantizar la viabilidad computacional del proceso, nuestra metodología incluye un mecanismo de búsqueda controlada que, cada vez que alcanza un conjunto de instancias desde una dada, cuantifica la relevancia de las mismas y filtra aquellas que no son interesantes para el usuario activo considerado. De esta forma, el enfoque explora la base de conocimiento avanzando únicamente a través de las propiedades que permiten alcanzar nuevas instancias desde las que son relevantes para el usuario, y omitiendo las que no aportan conocimiento útil al proceso de recomendación por no estar relacionadas de forma significativa con sus preferencias personales.

Una vez finalizado el proceso de **filtrado**, la metodología dispone de un conjunto de secuencias de propiedades compuesto por instancias relevantes para el usuario. La estrategia desarrollada debe analizar tales secuencias, inferir las asociaciones semánticas entre aquellas instancias que identifiquen contenidos de TV concretos y, finalmente, seleccionar los programas que recomendará al usuario. En nuestro trabajo, este último proceso recibe el nombre de **validación semántica**, pues su cometido es gestionar el conocimiento inferido sobre los intereses de los usuarios y, posteriormente, validarlo a fin de asegurar el carácter personalizado de las recomendaciones ofrecidas.

Como resultado del proceso de filtrado antes mencionado, es posible que una misma instancia forme parte de varias de las secuencias de propiedades obtenidas, debido a que está relacionada con varios de los contenidos registrados en el perfil del usuario. Este tipo de relaciones *intersecuenciales* deben ser aprovechadas a la hora de seleccionar las sugerencias personalizadas. Para ello, el proceso de validación semántica debe explorar el conocimiento representado en la red que forman tales secuencias, en la que los nodos identifican las instancias incluidas en las mismas y los enlaces se refieren tanto a las propiedades de la ontología como a las asociaciones semánticas inferidas desde ésta.

Tal como se describió en la Sección 3.9, las técnicas de propagación de activación son mecanismos eficientes que permiten explorar e inferir conocimiento a partir de las relaciones conceptuales representadas en redes genéricas, de ahí su indudable utilidad en el proceso de validación semántica propuesto en nuestra metodología. En concreto, dado un conjunto de nodos relevantes en la red, las técnicas SA son capaces de identificar, tras el proceso de propagación de activación, aquellos que están relacionados con los primeros de forma significativa. En nuestro escenario de aplicación, el proceso de recomendación consiste en identificar en la red los contenidos que interesaron al usuario, seleccionar aquellos programas que, una vez finalizado el procedimiento de propagación, registran los niveles de activación más elevados y, finalmente, ordenarlos de acuerdo a su relevancia para el usuario.

7.1.2 El agente semántico

Una vez descritas las dos fases —filtrado y validación— incluidas en la estrategia de recomendación desarrollada, localizamos su ubicación en la arquitectura del sistema

AVATAR. En concreto, es el agente semántico representado en la Fig. 4.4 quien asume la ejecución de esta técnica. Dicho agente se compone de varios módulos, en los que se han delegado las diferentes tareas que integran el proceso de razonamiento semántico aplicado en la estrategia, tal como muestra la Fig. 7.1.

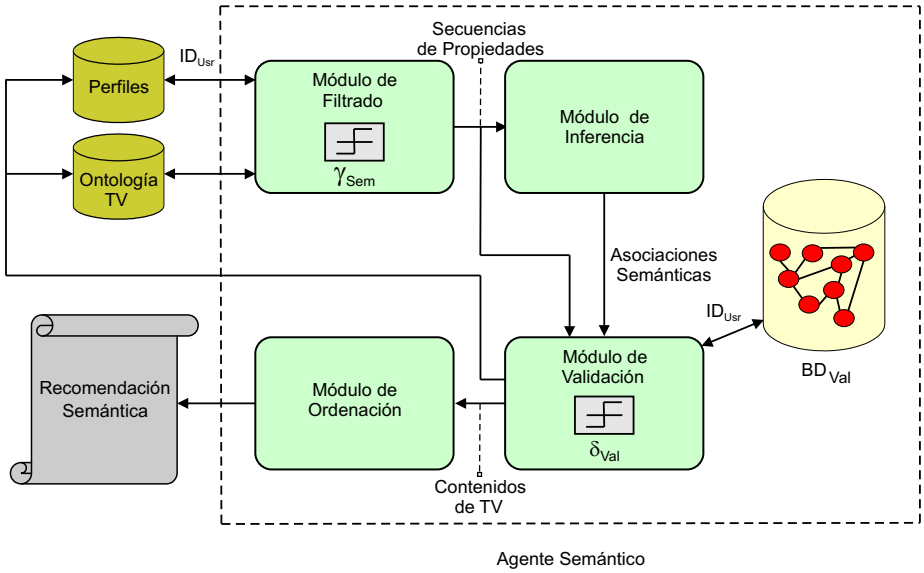

Figura 7.1: Estructura del agente semántico representado en la Fig. 4.4

- *Módulo de filtrado*: Es el encargado de filtrar las secuencias de propiedades que involucren instancias irrelevantes para el usuario activo, esto es, no relacionadas semánticamente con sus preferencias personales. Para ello, este proceso accede a dos fuentes de conocimiento: la ontología de dominio de AVATAR y el perfil del usuario correspondiente.

- *Módulo de inferencia*: Tal como muestra la Fig. 7.1, este módulo trabaja sobre las secuencias de propiedades seleccionadas por el *módulo de filtrado*, y se encarga de inferir a partir de ellas las asociaciones semánticas correspondientes. Para tal propósito, simplemente comprueba que dichas secuencias de propiedades verifiquen las restricciones relativas a cada tipo de asociación semántica (ver Sección 3.5.3).

- *Módulo de validación*: Como su nombre indica, este módulo es el que implementa la metodología de validación semántica propiamente dicha. Para ello, accede a las secuencias de propiedades seleccionadas por el *módulo de filtrado* y a las asociaciones inferidas desde las mismas. A partir de este conocimiento, el módulo construye la red sobre la que aplica las técnicas de propagación de activación, para así poder seleccionar los contenidos que están relacionados, de forma significativa, con las prefererencias del usuario. En adelante, nos referiremos a dicha red, almacenada en una base de datos de validación, como red de propagación o red SA (*Spreading Activation*).

- *Base de datos de validación*: De forma análoga a la comentada en los perfiles de usuario, AVATAR dispone de una base de datos de validación en la que se almacenan las redes SA de todos los usuarios registrados. Tal como muestra la Fig. 7.1, el sistema indexa esta base de datos con el identificador único asignado a cada espectador durante la fase de registro (ver Fig. 4.4).

- *Módulo de ordenación*: Este módulo es el encargado de ordenar los contenidos de TV seleccionados durante la fase de validación semántica. El proceso de ordenación se basa en dos criterios: (i) la relevancia de las asociaciones semánticas inferidas entre las preferencias del usuario activo y los contenidos sugeridos, y (ii) los niveles de interés concretos definidos en su perfil. De esta forma, los contenidos que ocupan las posiciones preferentes de la recomendación serán aquellos que estén relacionados (desde un punto de vista semántico) con las preferencias que más interesan al espectador.

7.2 Filtrado de secuencias de propiedades

Tal como se aprecia en la Fig. 7.1, para poder inferir las asociaciones semánticas entre las preferencias de un usuario activo y los contenidos recomendados, es necesario que el *módulo de filtrado* seleccione, en primer lugar, un conjunto de secuencias de propiedades concretas a partir del conocimiento representado tanto en su perfil como en la ontología OWL.

Para asegurar el carácter personalizado de las recomendaciones, dichas secuencias deben incluir únicamente instancias que estén relacionadas con las preferencias actuales del usuario[1]. Para asegurar el descubrimiento de relaciones semánticas significativas, nuestra metodología propone un algoritmo que, para cada usuario activo, explora las secuencias de propiedades definidas en la ontología cuyos orígenes son programas incluidos en su perfil (y pertenecientes a la ventana temporal actual), a la vez que filtra aquellas instancias de las mismas que no son relevantes para dicho espectador. Este proceso supone que la intensidad de la relación semántica establecida entre el origen de una secuencia y las restantes instancias de la misma, decrece con su longitud; como consecuencia, el objetivo del algoritmo propuesto es identificar el término de cada una de las secuencias analizadas, es decir, la última instancia que aún puede ser considerada relevante para el usuario.

Notación

Antes de abundar en los detalles algorítmicos, presentamos la notación utilizada en nuestra propuesta. En lo que sigue, suponemos que:

- U es el usuario activo del sistema AVATAR y P_U es su perfil personal. De acuerdo a lo explicado en la Sección 6.2.2, P_U registra los CRIDs de los programas que definen las preferencias positivas y negativas de U y, mediante ellos, referencia las instancias que identifican tales programas y sus características semánticas en la base de conocimiento del sistema.

[1] De acuerdo a lo explicado en el capítulo 6, tales preferencias son las instancias referenciadas desde el perfil cuyos índices de olvido pertenecen a la ventana temporal actualmente considerada por el sistema.

- En concreto, son N_I las instancias de la ontología OWL que identifican las preferencias de U, representadas en adelante como c_t, con $1 \leq t \leq N_I$.
- De las N_I instancias identificadas, sólo N_P representan contenidos concretos que interesaron a U en el pasado, denotados como x_i, con $1 \leq i \leq N_P$. El resto de instancias se refieren, por tanto, a programas que no han interesado al usuario y a las características semánticas asociadas tanto a éstos como a los contenidos x_i.
- Cada uno de los programas x_i es origen de R_i secuencias de propiedades en la base de conocimiento OWL del sistema.
- Cada una de dichas secuencias, denotadas como $ps_k^{x_i}$, con $1 \leq k \leq R_i$, contiene $N_{k,i}$ instancias; de ahí que, rescatando las funciones definidas en SemDis (ver Sección 3.5.1), se cumpla:

$$ps_k^{x_i}.PSNodesSequence() = \{x_i, v_1, \ldots, v_{N^{k,i}-1}\}$$

Algoritmo de filtrado propuesto

Tal como muestra el pseudocódigo representado en Alg. 1, para cada uno de los contenidos x_i que interesaron al usuario U (cuyo perfil es P_U), el algoritmo explora las R_i secuencias de propiedades donde dicha instancia figura como origen[2]. A medida que recorre estas secuencias $ps_k^{x_i}$, el algoritmo de filtrado analiza cada uno de sus nodos v_j y calcula el valor de un parámetro denominado *intensidad semántica* (representado como $\lambda_{Sem}(v_j)$).

Alg. 1 Algoritmo de filtrado de secuencias de propiedades

```
*****************************************************************
    * Entrada:    contenidos x_i del perfil P_U en la ventana temporal actual
    * Salida:     ps_U → conjunto de todas las secuencias inferidas para usuario U
*****************************************************************
ps_U  Filtrado-Secuencias  (P_U) {
  Para  i = 1:N_P
   Para  k = 1:R_i
      p̃s_k^{x_i}.añadir (x_i)
      Para  j = 1:N^{k,i}-1
         p̃s_k^{x_i}.añadir(v_j)
         Si  ((v_j ∉ P_U)  Y  (λ_Sem(v_j) < γ_Sem_U))  Entonces break
      ps_U.añadir (p̃s_k^{x_i})
}
```

Con el fin de determinar si el nodo v_j es suficientemente relevante para el usuario U, la metodología propuesta requiere un umbral con el que comparar el valor de intensidad semántica obtenido para el mismo. Para este propósito, nuestro enfoque define

[2]En el pseudocódigo mostrado en Alg. 1, $\widetilde{ps}_k^{x_i}$ representa la secuencia de propiedades seleccionada por el algoritmo de filtrado tras eliminar de la secuencia $ps_k^{x_i}$ formalizada en la ontología, todas las instancias que no son relevantes para el usuario U. El resto de instancias se añaden a $\widetilde{ps}_k^{x_i}$ mediante la función *añadir*, utilizada también para almacenar todas las secuencias seleccionadas en el conjunto ps_U que devuelve el algoritmo.

(para el usuario U) un umbral semántico γ_{Sem_U}, cuyo rango de posibles valores será revisado en detalle en la Sección 7.4.

- Si la intensidad de v_j supera dicho umbral, o bien si este nodo identifica una de las preferencias registradas en P_U, la instancia se incluye en la secuencia inferida y el algoritmo repite el proceso para el siguiente nodo de la misma.

- En caso contrario, la instancia v_j es el término de la secuencia seleccionada y el algoritmo explora la siguiente secuencia de propiedades iniciada en x_i. En este caso, la intensidad de v_j puede tomar valores positivos poco significativos (de ahí que sea término de la secuencia) o, incluso, valores negativos. En el primer caso, el algoritmo habría detectado relaciones débiles entre v_j y las preferencias positivas de U; en el segundo supuesto, las relaciones inferidas se establecerían entre v_j y las preferencias negativas del usuario, de ahí los niveles negativos de intensidad semántica que alcanza este nodo. Esta es la razón por la que el *módulo de filtrado* incluye la instancia v_l en la secuencia considerada (aun cuando no supera el umbral semántico), ya que, al establecer un vínculo con las preferencias negativas de U, aporta conocimiento útil de cara al proceso de personalización; en concreto, permite que la fase de validación identifique contenidos que, por estar relacionados con dicha instancia, no deben ser sugeridos al usuario U.

De forma intuitiva, la intensidad semántica de una instancia v_j en la secuencia $ps_k^{x_i}$ debe ser entendida como una medida de la fuerza de la relación semántica existente entre ese nodo y las preferencias del usuario U; de hecho, cuanto mayor sea su valor, más significativa será dicha relación. En este sentido, conviene destacar que un mismo nodo puede estar incluido en varias de las secuencias seleccionadas por el módulo de filtrado. Este hecho no debe pasar desapercibido en la metodología de razonamiento propuesta, que, en tal caso, debe incrementar el valor de intensidad semántica correspondiente a dicho nodo, por estar relacionado con varias de las preferencias registradas en el perfil de U.

7.3 Intensidad semántica

Para calcular la intensidad semántica de cada una de las instancias incluidas en las secuencias analizadas, es necesario emplear algunos parámetros que permitan detectar relaciones simples entre la instancia considerada y las preferencias del usuario activo. Este tipo de parámetros se describen en la siguiente sección, antes de ilustrar la metodología de inferencia propuesta en la tesis. Para este último propósito, emplearemos la secuencia de propiedades mostrada en la Fig. 7.2, cuyo origen es uno de los contenidos definidos en el perfil P_U. En concreto, calcularemos el valor de intensidad semántica correspondiente a la instancia v_l incluida en dicha secuencia.

7.3.1 Conceptos fundamentales

Sean x, y y z tres instancias específicas pertenecientes a las clases C_x, C_y y C_z, respectivamente, organizadas en una jerarquía $\mathcal{H}$.

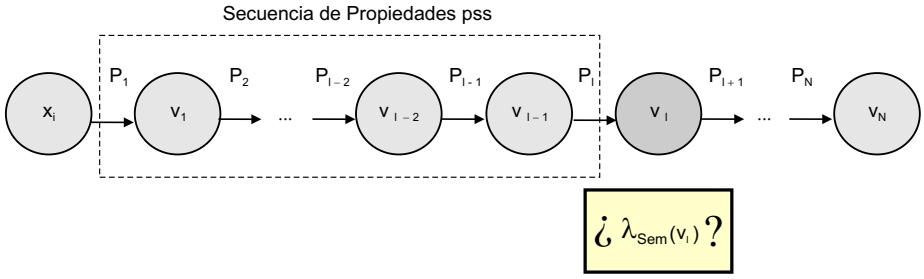

Figura 7.2: Secuencia de propiedades $ps_k^{x_i}$ definida en la ontología OWL de TV

Similitud semántica jerárquica

El valor de la similitud jerárquica entre x e y depende de la posición que sus respectivas clases C_x y C_y ocupan en $\mathcal{H}$. Para poder definir una expresión analítica para esta métrica, es necesario recurrir a los conceptos de *profundidad* y *antecesor común más cercano* (LCA) presentados en la Sección 3.3.1. Reuniendo sus respectivas contribuciones resulta la Ec. 7.1, desde la que es posible extraer algunas conclusiones interesantes acerca de los valores cuantificados por la métrica definida.

$$SemSim_{Hie}(x, y) = \frac{depth(LCA_{x,y})}{\max(depth(x), depth(y))} \tag{7.1}$$

- La similitud jerárquica $SemSim_{Hie}(x, y)$ se anula cuando el único antecesor común que comparten las clases C_x y C_y es el nodo raíz de la jerarquía $\mathcal{H}$, al que corresponde un valor de profundidad nula. Tal premisa es razonable, ya que en este escenario, las clases a las que pertenecen las instancias comparadas no comparten información en la estructura jerárquica considerada.

- En caso de que exista un LCA diferente de dicha clase raíz, la similitud entre x e y será más significativa cuanto mayor sea la profundidad del antecesor común. Esto se debe a que al ser más profundo, el LCA también es más específico y, como consecuencia, proporciona más conocimiento sobre la relación existente entre las dos instancias comparadas.

- Por último, la Ec. 7.1 conduce a valores de similitud más elevados cuanto más próximo esté el antecesor común $LCA_{x,y}$ de las clases a las que pertenecen dichas instancias en la jerarquía $\mathcal{H}$, debido a que, en este caso, la relación entre x e y es más significativa.

Tal como describimos en el capítulo 5, las instancias definidas en nuestra base de conocimiento OWL pueden pertenecer a varias de las clases formalizadas en la ontología. Así, en el supuesto de que la clasificación múltiple afecte a las instancias x e y comparadas, nuestro enfoque aplica la Ec. 7.1 sobre cada par de clases a las que éstas pertenecen y toma el valor máximo de todas las contribuciones calculadas.

Secuencia de propiedades geodésicas

De acuerdo a la noción de *camino geodésico* (o camino más corto) empleada en teoría de grafos [64], la metodología propuesta define las secuencias de propiedades geodésicas entre x e y (representadas como geo(x, y)) como aquellas de longitud mínima cuyo origen es x y cuyo término es y.

Intermediación

En teoría de grafos, la *intermediación* es un parámetro usado para valorar la *relevancia* de cada nodo en un grafo dado. De hecho, la *intermediación* de un nodo se define como el número de veces que, para unir cualquier par de nodos del grafo a través de caminos geodésicos, es necesario atravesar el nodo especificado.

En nuestra metodología, la *intermediación* de una instancia z con respecto a x y a y (representada como $b_z(x, y)$ en la Ec. 7.2) es la proporción de secuencias de propiedades geodésicas entre x e y en las que también está incluida z. Esta definición conduce a valores de $b_z(x, y)$ mayores cuanto más significativa sea la relación existente entre las tres instancias consideradas.

$$b_z(x, y) = \frac{\#\text{geo}(x, z, y)}{\#\text{geo}(x, y)} \tag{7.2}$$

7.3.2 Componentes de la intensidad semántica

A la hora de cuantificar el valor de la intensidad semántica de la instancia v_l mostrada en la Fig. 7.2, nuestra aproximación considera la contribución de componentes de muy diversa naturaleza que analizan tanto las clases como las intancias que identifican las preferencias del usuario U, con el claro propósito de inferir relaciones de interés entre dichas entidades y la instancia específica considerada. Cada una de las componentes incluidas en el cálculo de la intensidad semántica son revisadas a continuación.

I. $C_I(v_l, P_U)$: Nivel de interés de U en relación a x_i, origen de la secuencia analizada

En nuestra metodología, si la instancia v_l está relacionada con un contenido x_i que interesa mucho al usuario, su intensidad semántica debe ser muy significativa. Ésta es la premisa que inspira la componente $C_I(v_l, P_U)$, cuyo valor es directamente proporcional al nivel de interés del usuario U en relación al contenido x_i que inicia la secuencia en la que v_l está incluida.

$$C_I(v_l, P_U) = DOI_{P_U}(x_i) \tag{7.3}$$

Junto al índice DOI del programa x_i, la metodología debe considerar también la relevancia de las relaciones semánticas existentes entre la instancia v_l considerada y las preferencias de U. Ésta es la razón que nos lleva a incluir nuevas componentes en el proceso de cálculo de la intensidad semántica.

II. $C_L(v_l, P_U)$**: Longitud de la secuencia** pss **establecida entre** x_i **y** v_l

Uno de los parámetros que ayudan a valorar la relevancia de la relación establecida entre dos nodos es precisamente la longitud de la secuencia de propiedades que los une. Con esta idea en mente, la metodología propuesta introduce la componente $C_L(v_l, P_U)$, a la que, tal como se aprecia en la Ec. 7.4, le corresponderán valores mayores cuanto más cerca esté v_l del programa x_i en la secuencia establecida entre ellos, ya que, en este caso, la relación entre ambas instancias es más significativa.

$$C_L(v_l, P_U) = \frac{1}{\text{length}(pss)} \text{ donde } pss = [P_0, \ldots, P_l] \tag{7.4}$$

III. $C_H(v_l, P_U)$**: Similitud jerárquica entre** v_l **y las instancias incluidas en el perfil** P_U

Esta componente revisa las clases que identifican las preferencias de U y detecta relaciones de herencia entre dichas preferencias y el nodo v_l. Su valor depende de dos factores: (i) la similitud jerárquica cuantificada entre v_l y las Q instancias del perfil P_U que pertenecen a la misma jerarquía de clases[3] que v_l, y (ii) el nivel de interés de U en relación a estas últimas.

En primer lugar, el enfoque selecciona las instancias de P_U más similares (jerárquicamente) al nodo v_l considerado. Tal como muestra la Ec. 7.5, estas instancias —cuya similitud máxima con v_l toma un valor M— se incluyen en el conjunto H.

$$M = max\ \{SemSim_{Hie}(v_l, c_t)\} \ \ \forall t, 1 \leq t \leq Q$$

$$H = \{c_t \ / \ SemSim_{Hie}(v_l, c_t) = M\} \tag{7.5}$$

A continuación, el valor de la componente $C_H(v_l, P_U)$ se calcula como el promedio de los índices de interés de dichas instancias, ponderado por el valor (máximo) de similitud jerárquica entre éstas y v_l, tal como muestra la Ec. 7.6.

$$C_H(v_l, P_U) = \frac{M}{|H|} \cdot \sum_{t=1}^{|H|} DOI_{P_U}(c_t) \ \text{ con } \ c_t \in H \tag{7.6}$$

Con arreglo a la Ec. 7.6, la componente $C_H(v_l, P_U)$ será mayor cuantas más relaciones jerárquicas existan entre v_l y aquellas instancias que más interesan a U. Por el contrario, esta componente tomará valores negativos cuando el nodo considerado está relacionado con las preferencias negativas de este usuario activo.

IV. $C_R(v_l, P_U)$**: Relevancia semántica de** v_l **en el perfil** P_U

Además de las relaciones de inclusión establecidas en una estructura jerárquica, la metodología propuesta también considera las propiedades definidas en la ontología

[3]Por ejemplo, si v_l es un programa de TV, la métrica compara esta instancia frente al resto de programas del perfil P_U clasificados bajo la jerarquía de géneros de la ontología OWL. Si por el contrario, v_l identifica una alerta de contenido, el enfoque únicamente cuantifica la similitud entre esta instancia y el resto de alertas definidas en P_U (en la jerarquía correspondiente).

OWL para así poder inferir relaciones más complejas. De acuerdo a la noción de intensidad semántica definida en nuestra aproximación, cuanto más relevantes sean las relaciones descubiertas entre el nodo analizado y las preferencias del usuario U, más significativo será el valor que tome este parámetro.

Para poder cuantificar tal relevancia, la metodología de razonamiento propuesta introduce la componente $C_R(v_l, P_U)$. Para medir su valor concreto, este enfoque recurre al concepto de *intermediación* presentado en la Sección 7.3.1. Este factor es una métrica adecuada para tal propósito, ya que permite descubrir relaciones significativas entre las instancias involucradas en su cálculo. Así, por ejemplo, un valor elevado de $b_z(x, y)$ se traduce en una relación relevante entre las instancias x, y y z, ya que en muchas de las secuencias más cortas que unen las dos primeras está también incluida la tercera.

Teniendo en cuenta que nuestro enfoque debe cuantificar la relevancia de las relaciones establecidas entre las preferencias del usuario y el nodo v_l analizado, el cálculo de los factores de *intermediación* debe considerar, por una parte, dicho nodo y, por otra, el conjunto de instancias incluidas en su perfil. En concreto, la componente $C_R(v_l, P_U)$ se calcula sumando la *intermediación* de cada instancia c_t definida en P_U, calculada con respecto a v_l y a cada una de las $N_I - 1$ instancias restantes definidas en el mismo, tal como muestra la Ec. 7.7. De acuerdo a esta expresión, a partir de valores de *intermediación* elevados, la metodología infiere que el nodo v_l está muy relacionado con instancias que identifican las preferencias personales de este usuario, de ahí que su componente C_R contribuya a incrementar el valor de intensidad semántica $\lambda_{Sem}(v_l)$.

$$C_R(v_l, P_U) = \frac{1}{K} \cdot \sum_{f,g=1;f\neq g}^{N_I} b_{c_f}(c_g, v_l) \cdot \frac{DOI_{P_U}(c_f) + DOI_{P_U}(c_g)}{2} \tag{7.7}$$

donde $K = N_I \cdot (N_I - 1)$ es una constante de normalización para asegurar que el máximo valor de $C_R(v_l, P_U)$ es 1.

Dado que las instancias c_t almacenadas en el perfil de U pueden identificar tanto preferencias positivas como negativas, la componente de la relevancia semántica también debe considerar el nivel de interés del usuario en relación a cada instancia. Esta es la razón por la que incluimos los índices DOI de las preferencias de U en la Ec. 7.7. Así, si la instancia v_l está relacionada con las preferencias negativas del usuario, su relevancia semántica será poco significativa y, como consecuencia, su intensidad muy reducida; por ello, dicha instancia puede ser filtrada de la secuencia de propiedades analizada. Por el contrario, la Ec. 7.7 conduce a valores mayores cuánto más relevantes sean las relaciones establecidas entre v_l y aquellas instancias que más interesan a U, identificadas mediante niveles de interés positivos y elevados.

Métricas adoptadas en otros enfoques

Tal como describimos en el capítulo 3, la cuantificación de la relevancia de una determinada relación ya ha sido materia de estudio en otros enfoques [6, 87]. Para acometer tal tarea, estos trabajos proponen diferentes métricas. Algunas de ellas coinciden con las componentes identificadas en esta sección (longitud de las secuencias, interés del usuario y relaciones jerárquicas); sin embargo, también introducen otros parámetros como la especificidad, la rareza, la popularidad y la entropía de las clases, instancias y propiedades formalizadas en una ontología. Estas métricas tienen una base

común, relacionada con la frecuencia de aparición de dichas entidades en la conceptualización. Así, una clase es más específica si existen pocas instancias de la misma en la base de conocimiento, y una propiedad es común en ésta cuando está asociada a muchas de sus instancias. Tal como se explicó en la Sección 3.7, la entropía comparte esta misma idea. Así, basándose en esta métrica, el enfoque [6] cuantifica la relevancia de una propiedad en función de la información que aporta al usuario (que será máxima cuanto más específica sea dicha propiedad en la base de conocimiento RDF(S)) y del tipo de consulta formulada por el mismo. En las consultas basadas en descubrimiento de conocimiento, el usuario está interesado en relaciones indirectas, no usuales en la base de conocimiento, de ahí que las asociaciones más relevantes para él sean aquellas a las que les corresponden mayores valores de entropía.

Este tipo de métricas no tienen cabida en nuestro enfoque, en el que la relevancia de una instancia v_l depende de la existencia de relaciones semánticas significativas entre dicho nodo y las preferencias del usuario activo. Por esta razón, nuestro trabajo emplea la *intermediación* —para así poder detectar relaciones útiles en el proceso de recomendación—, y descarta factores como la especificidad, la popularidad o la entropía de una instancia (o propiedad), que no aseguran los requisitos de personalización perseguidos en nuestra metodología de razonamiento.

Expresión final de la intensidad semántica

Tal como adelantamos previamente, el nodo v_l puede estar incluido en varias secuencias de propiedades seleccionadas por el *módulo de filtrado*. Desde esto último, la metodología propuesta infiere que v_l está relacionado con muchas de las preferencias del usuario activo U, razón por la que su valor de intensidad semántica $\lambda_{Sem}(v_l)$ debe ser reforzado considerando la contribución de cada una de dichas secuencias. De acuerdo a lo explicado en la Sección 7.3.2, sólo dos de las componentes utilizadas en el cálculo de la intensidad semántica de un nodo dependen de la secuencia de propiedades concreta en la que éste está incluido. En concreto, las componentes C_I y C_L son función de la instancia origen de la secuencia analizada, mientras que la relevancia (C_R) y la componente jerárquica (C_H) se obtienen a partir de las preferencias registradas en el perfil del usuario activo. Por esta razón, tal como muestra la Ec. 7.8, cuando el módulo de filtrado analiza un nodo ya incluido en una secuencia previamente seleccionada, incrementa su intensidad considerando únicamente las componentes C_I y C_L de dicho nodo en la secuencia actual, ya que el valor de las dos componentes restantes permanece invariable.

$$\lambda_{Sem}(v_l) = \frac{w_1}{T} \cdot \sum_{i=1}^{T} C_I(v_l, P_U)\Big|_{ps_i} \cdot C_L(v_l, P_U)\Big|_{ps_i} + w_2 \cdot C_H(v_l, P_U) + w_3 \cdot C_R(v_l, P_U) \tag{7.8}$$

donde ps_i es la i-ésima de las T secuencias de propiedades (que han sido seleccionadas por el *módulo de filtrado*) en las que está incluido el nodo v_l, y $C_I(v_l, p_U)\Big|_{ps_i}$ y $C_L(v_l, p_U)\Big|_{ps_i}$ los valores que toman las componentes C_I y C_L de v_l en ps_i, respectivamente. Además, debe cumplirse $\sum_{z=1}^{3} w_z = 1$, para así garantizar que el máximo valor permitido para la intensidad semántica sea 1.

Estos pesos modulan la contribución de cada una de las componentes de la inten-

sidad semántica durante el proceso de inferencia. La metodología tiende a repartir de forma uniforme los valores de dichos pesos, destacando levemente la contribución de la relevancia semántica precisamente por ser un elemento clave a la hora de descubrir conocimiento útil para el proceso de personalización[4].

7.4 Criterio de filtrado: rango del umbral semántico

Una vez formalizada la expresión matemática de la intensidad semántica, es necesario establecer los posibles valores del umbral semántico γ_{Sem_U} empleado en el algoritmo de filtrado de secuencias de propiedades (ver Alg. 1). Con el ánimo de mejorar la personalización ofrecida por nuestra herramienta, dicho valor es ajustado dinámicamente para cada uno de los usuarios activos. Los parámetros que modulan este proceso son tanto la precisión de las recomendaciones ofrecidas por el sistema a cada espectador en el pasado, como el tipo de sugerencias que le interesan. En este sentido, nuestro enfoque considera tres tipos de recomendaciones: *especializadas*, en las que se incluyen contenidos relacionados de forma muy directa con las preferencias del usuario; *variadas*, en las que se identifican programas asociados con sus intereses de una forma indirecta, pero relevante para el mismo; y *mixtas*, en las que, en virtud de su nombre, se incorporan contenidos pertenecientes a las dos categorías anteriores.

A la vista del proceso de filtrado descrito en Alg. 1, el rango de operación del umbral γ_{Sem_U} se establece, *a priori*, entre los valores 0 y 1. Por una parte, γ_{Sem_U} debe ser siempre positivo, ya que el objetivo del filtrado de secuencias es identificar instancias relevantes para el usuario; de ahí que se premien aquellas cuya intensidad sea no sólo positiva sino elevada. Es precisamente dicha intensidad, cuyo valor máximo es 1, la que establece la cota superior de γ_{Sem_U}. Sin embargo, esta primera aproximación del rango de γ_{Sem_U} requiere un ajuste más fino, que garantice un equilibrio entre la *sobreespecialización* y la excesiva diversidad de las recomendaciones a las que conducirían los valores $\gamma_{Sem_U} = 1$ y $\gamma_{Sem_U} = 0$, respectivamente.

- En el supuesto de utilizar un umbral de valor 1, el *módulo de filtrado* únicamente podría seleccionar instancias que, por estar relacionadas con las preferencias del usuario de forma muy directa, hubiesen alcanzado valores de intensidad semántica también máximos, de ahí el carácter *sobreespecializado* de las sugerencias ofrecidas.

- Por el contrario, utilizando un umbral $\gamma_{Sem_U} = 0$, nuestro enfoque omitiría los requisitos de personalización del sistema, pudiendo sugerir incluso contenidos que, por tener asociada una intensidad semántica nula, son totalmente irrelevantes para el usuario.

Para evitar ambas situaciones, nuestro enfoque establece que los valores del umbral semántico oscilen entre ω_1 y ω_2, con $0 < \omega_1, \omega_2 < 1$.

- El primero de ellos es un parámetro crítico en la estrategia, dado que determina la cantidad de conocimiento que infiere el sistema durante la fase de filtrado (recordar el caso $\gamma_{Sem_U} = 0$). El valor de ω_1 dependerá, por tanto, del siste-

[4]En la herramienta implementada (ver Apéndice B) se han usado los pesos $w_1 = w_2 = 0{,}3$ y $w_3 = 0{,}4$.

ma recomendador en el que se aplique la metodología de razonamiento y, más concretamente, de la ontología de dominio utilizada en el mismo[5].

- Por su propia naturaleza, la cota superior del umbral semántico puede ser fijada de forma prácticamente arbitraria, siempre que toma valores próximos a 1 (recordar el caso $\gamma_{Sem_U} = 1$)[6].

Una vez delimitado el rango de operación de γ_{Sem_U}, es necesario definir el proceso de actualización del mismo. Son precisamente los tres tipos de sugerencias elaboradas en nuestra metodología —especializadas, variadas y mixtas— los que justifican que este parámetro no tome el mismo valor para todos los usuarios del sistema, sino que se adapte, de forma dinámica y personalizada, a cada uno de los mismos. Para este propósito, la metodología propuesta identifica, en primer lugar, la clase de recomendación que ha interesado al usuario U en el pasado y, a continuación, actualiza el valor de su umbral γ_{Sem_U} para que las sugerencias futuras pertenezcan a este mismo tipo.

- Empecemos suponiendo que este usuario se ha decantado por las sugerencias especializadas frente a las variadas en su última interacción con el sistema. Para poder seguir incluyendo programas muy similares a sus preferencias en las siguientes recomendaciones, el sistema debe establecer un valor elevado para su umbral semántico, de modo que la metodología filtre todos aquellos contenidos con intensidades semánticas poco significativas.

- Consideremos ahora que el espectador rechaza las sugerencias especializadas y muestra interés por contenidos relacionados de forma indirecta con sus preferencias personales. Con el propósito de seguir conservando la diversificación en las siguientes recomendaciones, la metodología propuesta debe reducir el valor del umbral semántico. De esta forma, aquellos programas cuyas intensidades semánticas no son muy elevadas también pueden ser seleccionados por el *módulo de filtrado* y recomendados al usuario tras el proceso de validación.

- Finalmente, para poder seleccionar recomendaciones mixtas para el usuario, el enfoque propuesto debe establecer un umbral semántico de valor intermedio, alcanzando así un equilibrio entre las dos situaciones descritas anteriormente.

Reuniendo los tres escenarios que acabamos de describir resulta la Ec. 7.9, mediante la cual se actualiza el valor del umbral semántico correspondiente al usuario U.

$$\gamma_{Sem_U}(t+1) = \begin{cases} \gamma_{Sem_U}(t) & \text{si} \quad P(t) = 1 \\ \omega_1 & \text{si} \quad P(t) = 0 \\ \omega_1 + (\omega_2 - \omega_1) \cdot \left(\frac{1-\omega_1}{1+\omega_1} \cdot \frac{\sum_{i=1}^{A} \lambda_{Sem}(a_i)}{\sum_{j=1}^{R} \lambda_{Sem}(r_j)}\right) & \text{en otro caso} \end{cases} \tag{7.9}$$

En esta expresión, consideramos que $t+1$ es el instante temporal en el que se presenta una sugerencia al usuario activo U, y t el correspondiente a la recomendación

[5]En el caso concreto de AVATAR, el valor mínimo que hemos utilizado para este parámetro se establece en torno a 0,66. Al considerar valores inferiores, nuestra validación experimental revela que la fase de filtrado no garantiza la relevancia de las instancias seleccionadas, debido a que éstas pueden estar relacionadas de forma muy débil con las preferencias del usuario.

[6]En nuestros experimentos, hemos utilizado valores de ω_2 en torno a 0.9.

anterior. Además, $\gamma_{Sem_U}(t)$ y $\gamma_{Sem_U}(t+1)$ son los valores de los umbrales utilizados en ambas recomendaciones, y $P(t)$ la precisión de la recomendación elaborada con $\gamma_{Sem_U}(t)$, definida como la proporción de programas sugeridos por el sistema en t y aceptados por U. Asimismo, a_i es el i-ésimo contenido aceptado por el usuario en la recomendación ofrecida en t, r_j es el j-ésimo programa rechazado por este espectador en dicha sugerencia, A y R son el número total de contenidos aceptados y rechazados, respectivamente, y $\lambda_{Sem}(a_i)$ y $\lambda_{Sem}(r_j)$ las intensidades semánticas correspondientes a tales programas.

Tal como muestra la Ec. 7.9, si la precisión del recomendador ha sido máxima en el instante t, el umbral conserva este valor en $t+1$. Por otra parte, si el usuario U no ha aceptado ninguno de los contenidos sugeridos por el sistema, la metodología fija el umbral semántico al valor ω_1, permitiendo que, en la siguiente sugerencia, el módulo de filtrado seleccione tanto contenidos similares a sus preferencias como otros relacionados con ellas de forma menos directa. En otro caso, el valor de $\gamma_{Sem_U}(t+1)$ se calcula a partir de las intensidades semánticas de los contenidos aceptados y rechazados en la sugerencia previa. Así, tal como explicamos anteriormente, la Ec. 7.9 conduce a umbrales elevados cuando U ha rechazado contenidos con intensidades moderadas y aceptado aquellos cuyos valores son muy significativos; si las selecciones y rechazos del usuario se invierten, entonces $\gamma_{Sem_U}(t+1)$ toma un valor reducido, ya que, en este caso, el usuario habrá mostrado interés por las sugerencias variadas frente a las especializadas.

A modo de conclusión, destacar que, inicialmente, la metodología puede asumir cualquier umbral γ_{Sem_U} en el rango permitido para este parámetro (entre $\omega_1 = 0{,}66$ y $\omega_2 = 0{,}9$ en nuestros experimentos), de forma que el recomendador semántico propuesto es capaz de adaptarse, de forma flexible, a las preferencias de cada tipo de usuario, diversificando, si fuese necesario, los contenidos sugeridos al mismo.

7.5 Asociaciones semánticas inferidas

Tal como se aprecia en la Fig. 7.1, una vez seleccionadas las secuencias de propiedades relevantes para el usuario, el *módulo de inferencia* debe descubrir asociaciones semánticas entre aquellas instancias incluidas en las mismas que identifican contenidos de TV concretos. Dichas asociaciones deben asegurar el carácter personalizado de las sugerencias ofrecidas a los espectadores. Para cumplir tal requisito, nuestra metodología descarta, de entre todas las asociaciones propuestas en [5], aquellas que por ser demasiado indirectas o débiles, desvirtuarían el carácter personalizado de las recomendaciones elaboradas.

En concreto, tal como se describe en [27], nuestra aproximación se basa en las asociaciones *ρ-path* y *ρ-join*, cuya definición se ha adaptado al dominio de aplicación del sistema AVATAR partiendo de los conceptos presentados en la Sección 3.5.3.

7.5.1 Asociaciones ρ-path

En nuestro enfoque, dos contenidos z_1 y z_2 están asociados de la forma *ρ-path*(z_1,z_2) si existe una secuencia de propiedades ps en la base de conocimiento OWL cuyo origen es z_1 y cuyo término es z_2, o bien, cuyo término es z_1 y cuyo origen es z_2.

La inferencia de este tipo de asociaciones a partir de las secuencias de propiedades

seleccionadas por el *módulo de filtrado* es inmediata. Además, considerando que cada una de estas secuencias se descompone a su vez en varias subsecuencias, es posible establecer una asociación *ρ-path* entre cada par de contenidos incluidas en las mismas.

7.5.2 Asociaciones ρ-join

De acuerdo a la definición introducida en la Sección 3.5.3, en nuestro enfoque decimos que z_1 y z_2 están asociados de la forma *ρ-join*(z_1,z_2) si es posibe encontrar dos secuencias PS_1 y PS_2 unidas mediante una clase C, y dos instancias de las mismas ps_1 y ps_2 en las que se cumple:

- x es el origen de ps_1 e y es el origen de ps_2, o bien
- x es el término de ps_1 e y es el término de ps_2.

Dicho de otra forma, la asociación *ρ-join*(z_1,z_2) se cumple siempre que las secuencias ps_1 y ps_2 antes mencionadas incluyan instancias —iguales o diferentes— pertenecientes a una misma clase, llamada *clase de unión*. Con arreglo a la definición de asociación *ρ-path* presentada antes, si dichas instancias son iguales es posible inferir dos tipos de relaciones entre ambos contenidos: *ρ-path*(z_1, z_2) y *ρ-join*(z_1, z_2). Por esta razón, las asociaciones *ρ-join* mencionadas a lo largo de la tesis se refieren únicamente a aquellas establecidas a través de instancias diferentes de una misma clase de unión; en caso de igualdad, la metodología infiere asociaciones *ρ-path* entre los contenidos considerados.

A fin de asegurar relaciones relevantes para el proceso de personalización, nuestro enfoque impone una restricción sobre dicha clase de unión. En concreto, C sólo puede identificar ciertas características semánticas de los contenidos formalizados en la ontología OWL. Dichas características se refieren al formato de producción o estructura de los programas, sus alertas de contenido, su temática (en el caso de que los programas tengan un propósito comercial), el tipo de audiencia a la que van dirigidos, su intención o su origen técnico. De esta forma, nuestra metodología únicamente asocia las preferencias positivas del usuario (origen de las secuencias inferidas) con otros contenidos a través de un conjunto de atributos previamente establecidos, impidiendo que el proceso de razonamiento semántico propuesto descubra conocimiento irrelevante para el mismo. Por ejemplo, en nuestro escenario de aplicación, es útil descubrir que dos programas están asociados semánticamente porque tratan una misma temática violenta, aun cuando los metadatos TV-Anytime describan dos tipos de alertas diferentes (por ejemplo, *violencia psicológica* (6.3.3) y *asesinatos* (6.3.4) en la Tabla 5.3). Formulemos ahora un razonamiento similar considerando la clase de unión *Actors*. En este caso, estaríamos asociando dos programas porque en ellos intervienen dos actores distintos. Es evidente que este tipo de relación sería completamente inservible de cara al proceso de personalización, de ahí la necesidad de limitar el conjunto de las posibles clases de unión consideradas en el enfoque.

La inferencia de las asociaciones *ρ-join* también es inmediata a partir de las secuencias seleccionadas por el *módulo de filtrado*. Para ello, basta analizar estas secuencias dos a dos, comprobando si sus respectivas instancias comparten alguna clase de unión. En caso afirmativo, el *módulo de inferencia* asocia los contenidos que figuran como origen de dichas secuencias (y también aquellos que figuren como término).

A fin de garantizar los requisitos de personalización impuestos en AVATAR, la estrategia desarrollada debe adaptarse a medida que evolucionan las preferencias de los usuarios. Este proceso involucra a todos los módulos que participan en el proceso de razonamiento semántico. Así, cuando el sistema descubra nuevas preferencias del usuario, el *módulo de filtrado* debe seleccionar las secuencias de propiedades originadas en dichos contenidos, y calcular los valores de intensidad semántica de sus respectivas instancias. Por su parte, el *módulo de inferencia* debe descubrir las asociaciones semánticas que originan dichas secuencias y, finalmente, es el *módulo de validación* quien, basándose en el conocimiento inferido, seleccione los contenidos recomendados a los espectadores.

7.6 Validación semántica

Tal como muestra la Fig. 7.1, el *módulo de validación semántica* debe procesar las secuencias de propiedades y asociaciones descubiertas por los módulos de filtrado e inferencia, respectivamente, con el fin de extraer desde las mismas aquellos contenidos de TV que son más adecuados para cada usuario activo.

Conviene destacar que las instancias seleccionadas mediante la metodología de filtrado propuesta en Alg. 1, estarán incluidas, típicamente, en varias de las secuencias de propiedades inferidas. De hecho, basta considerar la componente de la relevancia semántica de un nodo (ver Ec. 7.7), para advertir que nuestra metodología premia a aquellas instancias que están relacionadas a la vez con varias de las preferencias del usuario considerado. Precisamente por esta razón, existen relaciones entre las secuencias seleccionadas durante la fase de filtrado, cuyo conocimiento debe ser explorado en el proceso de razonamiento semántico y, en consecuencia, aplicado en la elaboración de las recomendaciones personalizadas.

A tal fin, el proceso de validación representa las secuencias de propiedades mediante una red genérica de nodos y enlaces, en la que los primeros se corresponden con las instancias incluidas en dichas secuencias, mientras que los segundos identifican las relaciones establecidas entre los mismos. Esta red debe ser procesada mediante algún tipo de mecanismo inteligente que permita explorar, de forma eficiente, el espacio de conceptos y relaciones formalizados en la misma, así como inferir conocimiento útil desde éstos.

Tal como se explicó en la Sección 3.9, las técnicas de propagación de activación soportan las dos capacidades requeridas en nuestra metodología, de ahí que este tipo de método de procesamiento semántico sea un candidato excelente para el proceso de validación. En concreto, dichas técnicas permiten identificar, a partir de un conjunto de conceptos inicialmente activados, aquellos que están relacionados con los primeros de una forma significativa. A diferencia de los enfoques existentes en la literatura, nuestra metodología combina el uso de las técnicas de propagación de activación con los requisitos de personalización de los sistemas recomendadores. Por esta razón, nuestra metodología de razonamiento debe adaptar tanto el proceso de ponderación de los enlaces utilizado en los enfoques SA actuales, como el tipo de relaciones definidas en la red sobre la que se aplican dichas técnicas.

- Por una parte, a fin de satisfacer los requisitos de personalización del sistema recomendador, en el proceso de ponderación de los enlaces no sólo deben participar los nodos conectados, sino también las preferencias personales del usuario

activo.

- Por otro lado, las relaciones conceptuales utilizadas durante el proceso de propagación también difieren de las consideradas en las propuestas existentes, pues incluyen tanto las propiedades explícitas definidas en las secuencias seleccionadas por el *módulo de filtrado*, como las asociaciones semánticas complejas descubiertas por el *módulo de inferencia*.

Dado que la estrategia propuesta persigue descubrir contenidos asociados semánticamente con las preferencias del usuario, el *módulo de validación* activa inicialmente aquellas intancias de la red que identifican los programas definidos en su perfil. De esta forma, tras el proceso de propagación, nuestra metodología debe valorar la relevancia de cada nodo de la red, y elegir, en base a la misma, los contenidos finalmente sugeridos al espectador. Para ello, este módulo selecciona aquellas instancias que satisfacen dos condiciones:

- Por una parte, deben pertenecer a la clase *TVContents* representada en la Fig. 5.1, para así garantizar que identifican programas concretos.
- Por otro lado, el nivel de activación de dichas instancias debe ser suficientemente elevado, asegurando con ello la relevancia de las relaciones inferidas entre los contenidos sugeridos y las preferencias del usuario (inicialmente activadas). Para verificar esta última condición, el enfoque introduce (para cada usuario activo U) un umbral de validación δ_{Val_U}, cuyo proceso de inicialización y rango de posibles valores se describirán en la Sección 7.6.6.

De acuerdo al criterio de recomendación descrito, aquellos contenidos seleccionados por el *módulo de validación* cuyo nivel de activación no sea suficientemente significativo, no serán sugeridos al usuario activo. La ventaja de este enfoque es que permite que la red de propagación se adapte, de forma flexible y dinámica, a las preferencias del usuario. Así, cuando su perfil registre nuevos intereses, éstos se incluirán en las secuencias inferidas por el *módulo de filtrado* y se añadirán a su red de propagación. De esta forma, es posible que los contenidos que antes no podían ser sugeridos, ahora vean incrementado su nivel y sean finalmente ofrecidos al espectador; dicho de otra manera, el proceso de validación debe gestionar el conocimiento aprendido por los *módulos de filtrado e inferencia*, reconociendo en qué momento un determinado contenido es relevante para el usuario (de acuerdo a sus preferencias actuales).

Con el fin de aplicar las técnicas de propagación de activación para este propósito, nuestra metodología debe justificar la elección tanto de un modelo de propagación concreto (técnicas puras o técnicas restringidas), como del algoritmo utilizado para su implementación (BNB o Hopfield Net).

7.6.1 Selección del modelo de propagación

Tal como describimos en la Sección 3.9, en la literatura se han propuesto tanto técnicas de propagación puras como modelos más elaborados (las llamadas técnicas CSA), basados en imponer ciertas restricciones durante el procesado de la red SA. A diferencia del modelo puro, en el que únicamente se computan los niveles de activación de los nodos, las técnicas CSA deben comprobar si cada nodo procesado satisface

las restricciones impuestas. En caso afirmativo, el proceso de propagación alcanza los vecinos de dicho nodo, mientras que, en caso contrario, se detiene en el mismo.

Para poder seleccionar el tipo de técnica de propagación más adecuada en nuestra metodología de validación, es necesario revisar las ventajas e inconvenientes de cada modelo. Por su simplicidad, las técnicas puras son computacionalmente menos costosas que el modelo CSA, ya que además del cómputo de los niveles de activación, es necesaria una segunda fase de verificación de restricciones. Sin embargo, tal como describimos en la Sección 3.9.2, la implantación de las técnicas puras en diferentes enfoques reveló dos limitaciones importantes, que motivaron la necesidad de proponer un modelo más complejo. Dichas limitaciones están relacionadas con la posibilidad de que el proceso de activación se propague sin control activando todos los nodos de la red y, por otra parte, con la dificultad de aplicar procesos inferenciales semánticos durante dicha propagación.

A pesar de tales inconvenientes, nuestro trabajo basa el proceso de validación semántica en las técnicas SA puras, aprovechando así las ventajas computaciones ofrecidas por este modelo. Para justificar esta elección basta comprobar que las limitaciones mencionadas no tienen efectos negativos en nuestra metodología:

- Por una parte, la propagación de activación por toda la red no es perjudicial durante el proceso de validación. Si ésta se produjera, todos los contenidos involucrados en las secuencias de propiedades inferidas serían suficientemente significativos para ser sugeridos al espectador, ya que habrían alcanzado niveles de activación capaces de superar el umbral de validación considerado en el enfoque.

- Por otra parte, hemos dicho que las técnicas SA puras dificultan la aplicación de procesos de inferencia sobre los nodos de la red, debido a que prescinden por completo de su semántica. Nuestra metodología no omite este tipo de razonamientos semánticos, sino que los considera durante las fases de filtrado e inferencia, previas al proceso de validación en el que utilizamos las técnicas SA. Además, como veremos en la siguiente sección, la filosolofía de las restricciones manejadas por el modelo CSA como solución a este problema, está implícitamente considerada en los pesos de los enlaces de la red de propación de cada usuario; por esta razón, decimos que la validación semántica propuesta en nuestra metodología está basada en *técnicas puras enriquecidas*[7].

7.6.2 Selección del algoritmo de propagación

Tal como describimos en el capítulo 3, en la literatura se han propuesto dos algoritmos para implementar el proceso de propagación en el que se basan las técnicas SA, aplicables tanto en el modelo puro como en el restringido. Éstos son el algoritmo BNB y el algoritmo Hopfield Net.

De acuerdo a la Tabla 3.1, ambos enfoques difieren fundamentalmente en el tipo de red en el que están basados, en el proceso de activación que llevan a cabo y en la condición de parada utilizada para detener la propagación de la activación. Revisemos estos tres parámetros para poder justificar, de forma razonada, cuál de los dos algoritmos es más adecuado para el proceso de validación semántica.

[7]No empleamos el término técnicas CSA porque nuestro enfoque no utiliza exactamente las mismas restricciones, y tampoco incluye la comprobación de las mismas en el proceso de propagación de los niveles de activación.

Tipo de red: Si bien el algoritmo BNB está basado en redes semánticas y Hopfield Net en redes neuronales, cualquiera de ellos podría ser aplicado, en principio, en nuestro enfoque, debido a que las técnicas de propagación de activación son válidas para cualquier red genérica de nodos interconectados. Así, en el caso de BNB, los conceptos y relaciones de la red semántica serían las instancias y propiedades, respectivamente, incluidas en las secuencias seleccionadas por el *módulo de filtrado*; por el contrario, en Hopfield Net, dichas instancias estarían identificadas mediante neuronas y las propiedades mediante conexiones sinápticas entre éstas.

Proceso de activación: Tal como se describió en la Sección 7.6, nuestro proceso de validación semántica debe calcular el nivel de activación de *todos* los nodos de la red construida desde las secuencias de propiedades inferidas, y seleccionar aquellos que tengan un valor superior al umbral δ_{Val_U} considerado. En este escenario, la capacidad de activación en paralelo del algoritmo Hopfield es claramente más eficiente que la llevada a cabo en BNB.

De acuerdo a lo explicado en la Sección 3.9.3, este último algoritmo sólo calcula en cada iteración los niveles de activación correspondientes a los vecinos del nodo extraído de la cola de prioridad. Como consecuencia, para poder computar los niveles de la red completa, nuestra metodología debería iterar hasta que dicha cola estuviese vacía, siendo esta última la condición de parada del algoritmo. Por el contrario, el algoritmo Hopfield permite calcular en paralelo los niveles de todos los nodos de la red en una sola iteración, de ahí que sea más adecuado en la metodología de validación propuesta.

El proceso de activación de cada algoritmo está estrechamente vinculado con el tipo de relaciones que pueden ser inferidas a partir del mismo, de ahí que este parámetro también deba ser considerado a la hora de decantarnos por un algoritmo específico.

- El objetivo en BNB es encontrar nodos relacionados de una forma muy significativa con aquellos que son más relevantes en la red, activados inicialmente en la misma. Precisamente para ello, este algoritmo extrae de la cola de prioridad el nodo con mayor nivel de activación (el más relevante) y explora únicamente sus vecinos, ya que el resto de nodos estarían relacionados con el primero de una forma más indirecta. De hecho, este proceso de activacion *en serie* confiere al algoritmo la propiedad de extraer, de forma eficiente, el camino más corto entre el nodo con mayor nivel durante la activación inicial, y aquel con mayor nivel tras el proceso de propagación.

  Si bien nuestra metodología de filtrado premia las instancias contenidas en estos caminos (elevando el valor de sus intensidades semánticas, tal como se describe en la Sección 7.3.2), las capacidades de razonamiento semántico de nuestro sistema requieren explorar también caminos más largos, favoreciendo así procesos inferenciales de mayor complejidad (que descubren más conocimiento).

- Por el contrario, al explorar la red completa en paralelo, Hopfield favorece la inferencia de relaciones más indirectas, desde las que es posible inferir conocimiento útil para el proceso de personalización.

Condición de parada: La condición empleada en Hopfield, basada en la convergencia de los niveles de activación de todos los nodos de la red, también resulta especialmente útil en nuestro enfoque de validación. Esto es debido a que garantiza la estabilidad de

los niveles utilizados tanto para seleccionar los programas recomendados al usuario, como para clasificarlos en orden adecuado en la sugerencia final presentada al mismo.

Precisamente la capacidad de convergencia de la red Hopfield, así como su propiedad de activación en paralelo, han sido los parámetros decisivos a la hora de seleccionar el algoritmo Hopfield Net como parte del proceso de validación semántica. Este algoritmo únicamente define el mecanismo concreto para calcular y propagar los niveles de activación de los nodos durante el proceso de validación[8]. Sin embargo, la metodología de razonamiento propuesta debe diseñar la red sobre la que se aplica este proceso de propagación, y garantizar que ésta representa todo el conocimiento que el agente semántico ha aprendido sobre las preferencias del usuario. Dicho conocimiento —que será empleado durante la elaboración de las recomendaciones personalizadas— está codificado, por una parte, en las secuencias de propiedades seleccionadas por el *módulo de filtrado*, y por otra, en las asociaciones semánticas inferidas durante el proceso de razonamiento. Veamos cómo nuestra metodología de inferencia incorpora ambos elementos en la red empleada en el algoritmo Hopfield Net.

7.6.3 La red de propagación de activación

La aplicación de Hopfield Net como algoritmo basado en técnicas de propagación de activación, impone una serie de restricciones sobre la red utilizada en nuestra metodología. En concreto, dichas restricciones están relacionadas con los enlaces (o conexiones sinápticas) establecidos entre los nodos (neuronas) de la misma. Tal como describimos en la Sección 2.2.7, el modelo propuesto por Hopfield consta de una red de nodos interconectados a través de enlaces bidireccionales ponderados de forma simétrica. Para cumplir este requisito, el *módulo de validación* incorpora en la red de propagación tanto las propiedades de las secuencias seleccionadas por el algoritmo de filtrado como sus inversas, y les asigna pesos idénticos.

Tal como adelantamos en la Sección 7.6, junto a las propiedades explícitas formalizadas en la ontología OWL implementada en el sistema AVATAR, la metodología propuesta debe codificar en la red de propagación de cada usuario las asociaciones semánticas seleccionadas por el *módulo de inferencia*, para así poder mejorar las recomendaciones elaboradas mediante procesos de razonamiento complejos. Para este propósito, el enfoque descrito en esta tesis define dos tipos diferentes de enlaces en las redes utilizadas durante la validación semántica:

- *Enlaces físicos*: Identifican las propiedades representadas en la ontología de TV, incluidas en las secuencias de propiedades seleccionadas por el *módulo de filtrado*.
- *Enlaces virtuales*: A diferencia de las propiedades explícitas antes mencionadas (consideradas en los enfoques convencionales), este tipo de enlaces identifican relaciones inferidas —mediante razonamiento semántico— a partir del conocimiento formalizado en la conceptualización OWL. En este conjunto, podemos identificar tanto relaciones jerárquicas simples como las asociaciones semánticas complejas que el *módulo de inferencia* descubre entre contenidos específicos. En función de la naturaleza de estas relaciones, nuestro enfoque define dos nuevos tipos de enlaces virtuales:

[8] De acuerdo a lo explicado en la Sección 3.9.4, el nivel de un nodo se computa combinando las contribuciones de todos sus vecinos mediante una función de activación no lineal.

- *Enlaces virtuales asociativos*: Como su nombre indica, este tipo de enlaces unen aquellos contenidos que están relacionados mediante las asociaciones *ρ-path* y *ρ-join* descubiertas por el *módulo de inferencia*.
- *Enlaces virtuales jerárquicos*: Para cada una de las asociaciones *ρ-join* descubiertas, el *módulo de validación* establece un enlace de este tipo entre las dos instancias de la clase de unión que origina tal relación semántica.

  A partir de una asociación de este tipo entre dos contenidos específicos, la metodología empleada define lo que, en adelante, denominamos *camino virtual*. Tal estructura permite ir de uno de los contenidos al otro, atravesando un número mínimo de enlaces físicos y el enlace jerárquico que origina la asociación *ρ-join* entre ambos. Al igual que la longitud de una secuencia de propiedades entre dos instancias es el número de propiedades incluidas en la misma, la longitud de un camino virtual se define como el número de enlaces físicos que contiene.

Ejemplo: Supongamos las secuencias $ps_1 = [p_0, p_1, p_2, p_3, p_4]$ y $ps_2 = [p_5, p_6]$ representadas en la parte superior de la Fig. 7.3. Consideremos además que las características semánticas C_1 y C_4 pertenecen a una misma clase de unión C. En este escenario, la inferencia de las asociaciones *ρ-path* y *ρ-join* mostradas en la figura son inmediatas. Las primeras son inferidas entre los programas que figuran tanto en las dos secuencias mencionadas como en las subsecuencias contenidas en las mismas; de ahí que el *módulo de inferencia* sea capaz de asociar entre sí tanto los contenidos P_1 y P_3, como P_1 y P_2. Este mismo razonamiento sigue siendo válido para las asociaciones *ρ-join*. Así, la metodología propuesta asocia tanto los contenidos P_1 y P_4, orígenes de las secuencias ps_1 y ps_2 que comparten la clase de unión C, como los programas P_2 y P_5, términos de las subsecuencias $[p_0, p_1]$ y $[p_5, p_6]$, respectivamente.
A partir de ambas secuencias y de dichas asociaciones, nuestro enfoque construye la red de propagación representada en la parte inferior de la Fig. 7.3, en la que junto a los enlaces virtuales y físicos podemos identificar tres caminos virtuales. El primero de ellos (de longitud 4) se establece entre los contenidos P_3 y P_5 e incluye los enlaces físicos P_3-C_2, C_2-P_2, P_2-C_1 y C_4-P_5, así como el enlace jerárquico entre C_1 y C_4. Este último enlace también está incluido en el camino virtual entre los programas P_1 y P_4 (de longitud 2), integrado además por los enlaces físicos P_1-C_1 y P_4-C_4. De forma análoga, es posible identificar un tercer camino virtual (también de longitud 2) entre los contenidos P_2 y P_5.

7.6.4 Ponderación de los enlaces de la red

En las propuestas existentes basadas en propagación de activación, el peso de un enlace cuantifica típicamente la *fuerza* o relevancia de la relación establecida entre los dos nodos que éste conecta [209, 235]. Además, debido a la naturaleza estática de las relaciones consideradas, dichos pesos suelen permanecer invariables a lo largo del tiempo. Así, por ejemplo, algunos enfoques en el campo de la recuperación de información aplican técnicas de propagación de activación con el fin de encontrar documentos de interés para los usuarios. En este dominio de aplicación, tal como se describió en la Sección 3.10.1, los términos (o conceptos en las herramientas semánticas) presentes en dichos documentos se modelan como nodos, mientras que los enlaces entre éstos identifican relaciones de ocurrencia en un mismo documento. Así, el peso asignado a

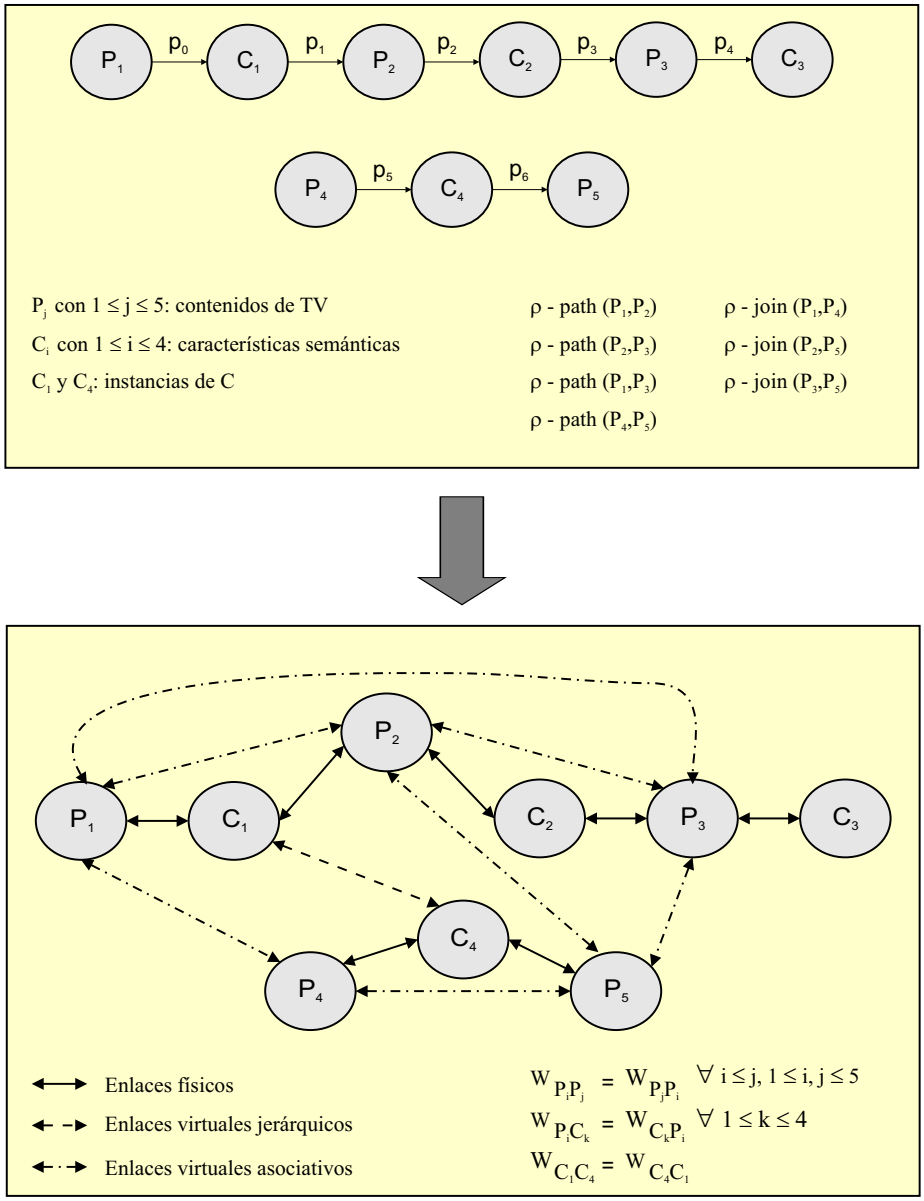

Figura 7.3: Construcción de una red de propagación a partir de secuencias de propiedades y asociaciones semánticas

cada enlace entre dos nodos depende únicamente del número de documentos en los que figuran ambos términos (conceptos). Como consecuencia de esto último, este tipo de enfoques consideran que dos términos están estrechamente relacionados cuando aparecen a la vez en muchos documentos, de ahí que asignen un peso muy elevado al enlace que los une en la red de propagación correspondiente. En este escenario, es evidente que los pesos de dicha red permanecerán invariables hasta que se modifique el conjunto de documentos considerados en el sistema de recuperación de información.

Tal como adelantábamos en la Sección 7.6, una de las características distintivas de la metodología propuesta en esta tesis, es la incorporación de los requisitos de personalización de un sistema recomendador en las técnicas de propagación de activación clásicas. Esta restricción exige que los pesos de los enlaces establecidos en la red SA considerada cumplan dos propiedades:

- Por una parte, el peso de un enlace no puede depender únicamente de los nodos conectados, sino de las posibles relaciones semánticas existentes entre éstos y las preferencias del usuario activo. Como consecuencia, es posible que el peso entre dos nodos dados sea muy elevado para un usuario y, sin embargo, muy poco significativo para otro espectador cuyas preferencias difieren de las del primero. Es evidente que este enfoque descarta la naturaleza estática de los pesos de los enlaces utilizados tradicionalmente en las técnicas de propagación de activación, tal como comentábamos al comienzo de la sección.

- Por otro lado, dichos pesos deben ajustarse dinámicamente ante los posibles cambios ocurridos en las preferencias de los usuarios. De esta forma, el sistema puede ofrecer recomendaciones basándose en información permanentemente actualizada, incrementando con ello la confianza del usuario.

En la metodología propuesta, el peso de un enlace —ya sea físico o virtual— se calcula a partir de las contribuciones de los dos nodos que une en la red de propagación del usuario activo U considerado. En concreto, la contribución de cada nodo se obtiene a partir de su llamada *función asociada*, cuyo valor es más elevado cuanto más relevante es dicho nodo para el usuario activo. Nuestro enfoque emplea dos parámetros a la hora de cuantificar tal relevancia, ambos incluidos en la expresión mostrada en la Ec. 7.10, en la que se computa el valor de la función asociada a un nodo i, unido a un segundo nodo j mediante el enlace ponderado en el enfoque aquí propuesto.

$$f_j(i) = \begin{cases} DOI_{P_U}(i) & \text{si } i \text{ está contenido en el perfil } P_U \\ \lambda_{Sem}(i) & \text{en otro caso} \end{cases} \tag{7.10}$$

- Si i está incluido entre las preferencias registradas en su perfil personal, su función asociada $f_j(i)$ se iguala al nivel de interés concreto del usuario en relación a dicho nodo, ya que éste es el parámetro más adecuado a la hora de valorar la relevancia del mismo para el espectador.

- En caso contrario, el valor de dicha función coincide con la intensidad semántica del nodo i, ya que, con arreglo a lo explicado en la Sección 7.3.2, cuanto mayor es la intensidad de un nodo, más significativa es la relación existente entre éste y las preferencias personales del usuario y, en consecuencia, más relevante es dicho nodo para el mismo.

Hasta este punto no hemos establecido ninguna diferencia entre el proceso de ponderación de un enlace físico y de un enlace virtual. Por su propia naturaleza, la metodología propuesta asume que el peso asignado a un enlace virtual debe ser menor que el que le correspondería si se tratase de un enlace físico, ya que en este último caso, la relación establecida entre los dos nodos unidos estaría explícitamente representada en la ontología OWL, mientras que en el primero, sería inferida mediante el proceso de razonamiento semántico. Con el fin de considerar tal premisa en el enfoque propuesto, introducimos el factor de atenuación $\mu_{ij} \in (0, 1]$ mostrado en la Ec. 7.11, donde se calcula el peso w_{ij} del enlace bidireccional establecido entre los nodos i y j de la red de propagación del usuario U.

$$w_{ij} = w_{ji} = \begin{cases} 0{,}5 \cdot (f_i(j) + f_j(i)) & \text{si enlace físico} \\ 0{,}5 \cdot \mu_{ij} \cdot (f_i(j) + f_j(i)) & \text{si enlace virtual} \end{cases} \tag{7.11}$$

El valor del factor de atenuación μ_{ij} depende del tipo de enlace virtual establecido entre los nodos i y j de la red de propagación considerada. En concreto, si dicho enlace es jerárquico, su peso simplemente se atenúa mediante una factor $0{,}85$, cuyo valor impide que la contribución de este enlace sufra un decremento excesivo. Por el contrario, en los enlaces asociativos, el valor del factor de atenuación depende de la relevancia de la asociación semántica inferida entre los dos contenidos conectados, de forma que cuanto más significativa sea esta última, mayor será dicho factor. En la metodología desarrollada, una asociación entre dos programas es más relevante cuanto más cerca están éstos en la red de propagación considerada. La métrica de distancia empleada en el enfoque depende del tipo de asociación inferida entre los dos contenidos unidos por el enlace asociativo:

- Si dicha asociación es *ρ-path*, nuestro trabajo emplea la longitud de la secuencia de propiedades establecida entre ambos programas para valorar la relevancia de esta relación semántica. Obviamente, cuanto mayor sea dicha longitud, menos significativa será esta relación y, como consecuencia, mayor la atenuación sufrida por el peso del enlace asociativo que une ambos contenidos.
- Por el contrario, si la asociación es *ρ-join*, nuestro enfoque utiliza como métrica la longitud del camino virtual existente entre los dos programas, tal como muestra la Ec. 7.12.

$$\mu_{ij} = \begin{cases} 0{,}85 & \text{si enlace virtual jerárquico} \\ \frac{1}{\widehat{\mu}_{ij}} & \text{si enlace virtual asociativo} \end{cases}$$

$$\widehat{\mu}_{ij} = \begin{cases} length\ (ps) & \text{si } \rho\text{-}path(i,j) \\ length\ (vpath) & \text{si } \rho\text{-}join(i,j) \end{cases} \tag{7.12}$$

donde *length* (ps) es la longitud de la secuencia de propiedades entre i y j a partir de la que se infiere la asociación *ρ-path*(i, j); y *length* $(vpath)$ es la longitud del camino virtual establecido entre ambos nodos en caso de existir una asociación *ρ-join* entre ellos.

De lo explicado hasta ahora se desprende que, por su propia naturaleza, el peso de un enlace asociativo entre dos contenidos depende tanto de la relevancia de ambos programas para el usuario activo considerado (igual que en el caso de los enlaces físicos y

jerárquicos), como del tipo de asociación semántica inferida entre éstos. Ambas contribuciones son las que diferencian la red de propagación empleada en nuestro enfoque, de las consideradas en otras propuestas existentes en el estado del arte también basadas en técnicas SA.

Ejemplo: Retomemos la red de propagación representada en la Fig. 7.3. En la Tabla 7.1, se muestran los factores de atenuación correspondientes a los enlaces virtuales identificados en la misma. En este escenario, al enlace asociativo entre los contenidos P_1 y P_2 le corresponde un factor de atenuación de valor $0{,}5$, ya que la asociación semántica establecida entre ambos es inferida a partir de una secuencia de propiedades de longitud 2. De forma análoga, el factor $\mu_{P_3P_5}$ toma un valor $0{,}25$, dado que el camino virtual que une ambos contenidos (a través del enlace jerárquico C_1-C_4) tiene una longitud 4. De dichos valores se desprende que el peso asignado al enlace asociativo entre P_1 y P_2 es más significativo que aquel establecido entre los programas P_3 y P_5. Esta premisa es razonable, dado que, tal como muestra la Fig. 7.3, la asociación semántica entre los dos primeros contenidos es más significativa que la inferida entre los dos últimos: *ρ-path*(P_1, P_2) se establece a través de dos propiedades explícitas en la ontología OWL; *ρ-join*(P_3, P_5), además de ser descubierta gracias a dos instancias diferentes de una misma clase de unión, involucra un número elevado de instancias (ver Tabla 7.1), que contribuyen a separar los dos programas considerados, debilitando con ello su relación semántica.

Nodos i, j	Asociación semántica	Nodos en PS	Nodos en $vpath$	μ_{ij}
C_1, C_4	–	–	–	0.85
P_1, P_2	*ρ-path* (P_1,P_2)	$P_1 - C_1 - P_2$	–	$\frac{1}{2}$
P_2, P_3	*ρ-path* (P_2,P_3)	$P_2 - C_2 - P_3$	–	$\frac{1}{2}$
P_1, P_3	*ρ-path* (P_1,P_3)	$P_1 - C_1 - P_2 - C_2 - P_3$	–	$\frac{1}{4}$
P_4, P_5	*ρ-path* (P_4,P_5)	$P_4 - C_4 - P_5$	–	$\frac{1}{2}$
P_1, P_4	*ρ-join* (P_1,P_4)	–	$P_1 - C_1 - C_4 - P_4$	$\frac{1}{2}$
P_2, P_5	*ρ-join* (P_2,P_5)	–	$P_2 - C_1 - C_4 - P_5$	$\frac{1}{2}$
P_3, P_5	*ρ-join* (P_3,P_5)	–	$P_3 - C_2 - P_2 - C_1 - C_4 - P_5$	$\frac{1}{4}$

Tabla 7.1: Factores de atenuación correspondientes a los enlaces virtuales representados en la Fig. 7.3

7.6.5 El algoritmo Hopfield Net

Una vez presentada la estructura de la red SA (nodos, enlaces y pesos asignados a estos últimos), nos centramos en los detalles del proceso de propagación llevado a cabo sobre la misma en el algoritmo Hopfield Net [102].

Tal como se explicó en la Sección 2.2.7, Hopfield propuso dos modelos para su red neuronal —discreto y continuo— acotando así los posibles valores que pueden tomar

las entradas y salidas de la misma; nuestra metodología recurre al modelo continuo. De esta forma, aquellas instancias que identifican contenidos que interesan al usuario (es decir, instancias con índices DOI positivos e índices de olvido gradual incluidos en la ventana temporal actual) son inicializadas con un nivel de activación igual a dicho nivel de interés, permitiendo así adaptar el algoritmo a los requisitos de personalización exigidos en el recomendador AVATAR. Al resto de los nodos de la red les corresponde un nivel de activación nulo durante la inicialización del algoritmo.

Dado que el algoritmo Hopfield Net está basado en una red neuronal, es necesario considerar una fase inicial en la que ésta debe aprender el conocimiento que utilizará durante el proceso de propagación de activación. Para proporcionar dicho conocimiento, es necesario identificar los nodos y enlaces definidos en la red, y ponderar estos últimos de forma conveniente. Tal como se describió en la sección previa, los pesos asignados a los enlaces son una medida de la relevancia de las relaciones (semánticas) existentes entre los nodos que interconectan y las preferencias personales del usuario. Basándose en dichos enlaces (y en sus pesos), la red neuronal es capaz de identificar aquellos nodos que están relacionados con los inicialmente activados de una forma significativa. En nuestro dominio de aplicación, esto último supone que la red neuronal es capaz de descubrir un conjunto de instancias relacionadas semánticamente con los contenidos que interesan al usuario activo, de ahí que los programas sugeridos al mismo deban ser extraídos de dicho conjunto.

La relevancia de las instancias seleccionadas por el algoritmo Hopfield Net se traduce en el algoritmo en niveles de activación elevados. En concreto, tal como se aprecia en el pseudocódigo mostrado en Alg. 2, el nivel de activación de cada nodo se calcula sumando la contribución de *todos* sus vecinos. Dado que esta contribución incluye tanto el nivel de activación del nodo vecino como el peso del enlace (físico o virtual) que los une, es fácil inferir que un nodo será más relevante cuanto más importantes sean sus vecinos (mayores niveles) y cuanto más significativa la relación establecida entre ellos y las preferencias del usuario (mayor peso), tal como ya adelantábamos al comienzo de la sección. Esta contribución se emplea como argumento de la función de activación sigmoide (ver Fig. 2.1). Por sus propiedades matemáticas, dados unos parámetros θ_1 y θ_2 concretos, esta función entrega niveles de activación mayores cuanto más elevada sea dicha contribución[9]. De acuerdo a la Ec. 7.10, el peso de un enlace en nuestra propuesta se corresponde con el nivel de interés del usuario en relación a los nodos que éste interconecta, o bien depende de sus respectivas intensidades semánticas. Como consecuencia, la función sigmoide cuantifica los niveles más elevados para aquellos nodos que están conectados con instancias que interesan mucho al usuario (DOI elevado), así como con nodos relacionados con sus preferencias (intensidad elevada). Por el contrario, y de acuerdo a los principios en los que se basa la estrategia de recomendación descrita en este capítulo, la metodología cuantifica niveles poco significativos para aquellas instancias relacionadas con las preferencias negativas del usuario, evitando de esta forma que sean sugeridas al mismo.

Por último, destacar que en el pseudocódigo del algoritmo de validación, es posible identificar tanto la propiedad de búsqueda en paralelo de la red Hopfield como su capacidad de convergencia. La primera permite calcular los niveles de todos los nodos de la red en una sola iteración, mientras que la segunda posibilita la selección de los contenidos de TV que satisfacen el criterio de recomendación considerado en la metodología, esto es, aquellos cuyos niveles de activación toman valores estables capaces

[9]En nuestros experimentos, hemos adoptado los parámetros $\theta_1 = 10$, $\theta_2 = 0{,}8$ y $\xi = 0{,}08$ utilizados en [102].

de superar el umbral δ_{Val_U} establecido[10].

Alg. 2 Algoritmo Hopfield usado en la validación semántica

```
*************************************************************
    * Entrada: R_U → red de propagación del usuario U
             P_U → perfil de U
    * Salida:  REC → conjunto de contenidos recomendados
*************************************************************
REC  Validación-Hopfield  (R_U, P_U) {
    t = 0;
    Para cada  nodo j  en  R_U  {
     Si  ((j ∈ P_U)  Y  (DOI_{P_U}(j) > 0))  Entonces  A_j(t) = DOI_{P_U}(j);
     En otro caso  A_j(t) = 0;
    }

    Para cada  nodo j  en  R_U  {
     net_j = Σ_{i=0}^{n-1} A_i(t) · w_ij ;

     A_j(t+1) = 1 / (1 + exp^{(θ_1 - net_j)/θ_2});
    }

    Mientras  (Σ_{j=0}^{n-1} |A_j(t+1) - A_j(t)| > ξ)  Hacer  {
     t++;
     Para cada  nodo j  en  R_U  {
      net_j = Σ_{i=0}^{n-1} A_i(t) · w_ij ;

      A_j(t+1) = 1 / (1 + exp^{(θ_1 - net_j)/θ_2});
     }
    }

    REC = Comprobar_Niveles_Activación (clase, δ_{Val_U});
}
```

7.6.6 Criterio de recomendación: rango del umbral de validación

De acuerdo a lo explicado en las secciones previas, el grado de personalización de las recomendaciones ofrecidas por la estrategia basada en contenido a un usuario U, se modula, principalmente, mediante los umbrales γ_{Sem_U} y δ_{Val_U}. Recordemos que cuanto mayor sea el primero, más significativa debe ser la relación existente entre las instancias contenidas en las secuencias de propiedades inferidas y las preferencias del usuario activo. Por su parte, cuanto mayor sea δ_{Val_U}, más elevado debe ser el nivel de activación de los contenidos involucrados en dichas secuencias para así poder ser

[10]Destacar que en Alg. 2, el argumento *clase* en la función *Comprobar_Niveles_Activación* puede ser la clase *TVContents* o cualquiera de sus subclases en la jerarquía de la Fig. 5.1, permitiendo así precisar el género que debe recomendar el módulo validador (contenidos de cualquier género, sólo programas deportivos, sólo películas de acción, etc.).

sugeridos al espectador. Esta premisa conduce a una reflexión interesante acerca de los valores permitidos para ambos parámetros.

En este sentido, es lógico permitir que el umbral γ_{Sem_U} tome valores menores que δ_{Val_U}, de forma que el *módulo de filtrado* infiera gran cantidad de conocimiento relacionado, tanto directa como indirectamente, con las preferencias del usuario; por su parte, el módulo validador gestionará dicho conocimiento, recomendando en cada momento únicamente aquellos contenidos que son relevantes para el espectador. En otras palabras, el *módulo de filtrado* llevará a cabo una primera búsqueda *extensa*, aunque siempre controlada (recordar las reflexiones hechas en la Sección 7.4 para ω_1), sobre el perfil del usuario y la ontología de TV, mientras que la validación se encargará de refinar este proceso y asegurar así la personalización.

De esto último se desprende que los valores del umbral δ_{Val_U} deben ser más restrictivos que los correspondientes a γ_{Sem_U} ($\delta_{Val_U} > \gamma_{Sem_U}$). En concreto, el umbral de validación oscila en el rango $[\omega_3, \omega_2]$, siendo ω_3 un valor elevado (en torno a 0.78 en nuestros experimentos) y ω_2 la misma cota inferior utilizada para el umbral γ_{Sem_U} (de valor aproximado 0.9).

De forma análoga a la descrita en el umbral semántico, δ_{Val_U} no permanece invariable a lo largo del tiempo, sino que nuestro enfoque modifica el valor correspondiente a cada uno de los usuarios a medida que éstos interaccionan con el sistema. En concreto, el parámetro que modula este proceso es la precisión que el usuario ha obtenido en las últimas sugerencias ofrecidas por el recomendador. Nuestra metodología propone decrementar el umbral de validación (hasta alcanzar el mínimo ω_3) cuando se advierte un incremento en la precisión del sistema y, por el contrario, aumentarlo (hasta ω_2) cuando esta última se reduce. En caso de que la precisión de la herramienta permanezca invariable entre dos recomendaciones consecutivas, el umbral también conserva su valor. La idea que se oculta tras este criterio es la siguiente: una vez que el sistema goza de la confianza del usuario —gracias a la precisión de las recomendaciones ofrecidas en el pasado—, la metodología decrementa el umbral δ_{Val_U} para así poder incluir en sus sugerencias contenidos con niveles de activación que, si bien no alcanzan el umbral utilizado en ese momento, están próximos al mismo. Si estos programas le interesan, nuestro enfoque habría descubierto contenidos relevantes para el usuario que otros enfoques omitirían; en caso contrario, el umbral de validación sería incrementado de nuevo para así seguir ofreciendo al usuario el tipo de sugerencias que le interesaron previamente.

Sean $t+1$ el instante temporal en el que se presenta una sugerencia al usuario activo U, y t el correspondiente a la recomendación anterior. Sean $P(t)$ y $P(t+1)$ los valores de precisión correspondientes a ambas recomendaciones, y $\delta_{Val_U}(t)$ y $\delta_{Val_U}(t+1)$ los valores de los umbrales utilizados en éstas. En este escenario, el valor del umbral de validación se actualiza en $t+1$ de acuerdo a la expresión mostrada en la Ec. 7.13.

$$\delta_{Val_U}(t+1) = \begin{cases} \omega_3 & \text{si} \quad \delta_{Val_U}(t) - \Delta_P(t+1,t) < \omega_3 \\ \omega_2 & \text{si} \quad \delta_{Val_U}(t) - \Delta_P(t+1,t) > \omega_2 \\ \delta_{Val_U}(t) - \Delta_P(t+1,t) & \text{en otro caso} \end{cases} \tag{7.13}$$

donde $\Delta_P(t+1,t) = P(t+1) - P(t)$.

Al igual que en el caso del umbral γ_{Sem_U}, el sistema recomendador puede establecer inicialmente cualquier umbral de validación en el rango permitido. Tal como

se ha descrito en esta sección, es la metodología propuesta la que modifica este valor a lo largo del tiempo, adaptándose de forma dinámica y flexible a las necesidades de personalización de cada usuario.

7.6.7 Etapas del proceso de validación semántica

Una vez presentados los detalles algorítmicos de la metodología, mostramos un diagrama de flujo en el que se resume el comportamiento del *módulo validador*.

Tal como muestra la Fig. 7.4, dicho módulo indexa la base de datos BD_{Val} con el identificador (único) del usuario activo, para así comprobar si ya existe una red de propagación para este espectador. En caso de que sea la primera vez que se ejecuta el proceso de validación para éste, se crea la red extrayendo los nodos y enlaces a partir de las secuencias de propiedades descubiertas por el *módulo de filtrado*. Por el contrario, si ya existe una red para este usuario, el módulo simplemente la actualiza incorporando nuevos nodos y enlaces a esta estructura, o incluso, eliminando aquellos que ya no son relevantes para el usuario. Una vez establecido el valor del umbral δ_{Val_U}, es necesario ponderar los enlaces de la red de propagación, en los que se codifica el conocimiento que el sistema ha aprendido sobre el usuario activo correspondiente. Finalmente, se ejecuta el algoritmo Validación-Hopfield mostrado en Alg. 2 y se extrae el conjunto de contenidos de TV que el *módulo de ordenación* representado en la Fig. 7.1 debe clasificar antes de presentar la sugerencia final al usuario.

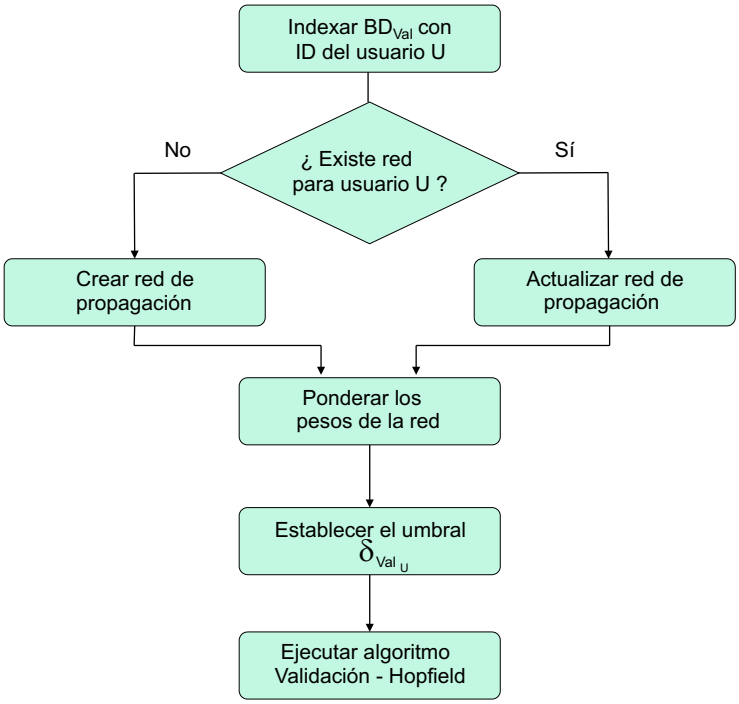

Figura 7.4: El proceso de validación semántica

El proceso descrito en el diagrama de la Fig. 7.4 debe adaptarse de forma dinámica a las preferencias del usuario activo. Así, a medida que el sistema dispone de conoci-

miento adicional sobre sus intereses, los *módulos de filtrado* e *inferencia* seleccionan nuevas secuencias de propiedades y asociaciones semánticas, respectivamente, y actualizan con ellas la red de propagación del espectador. En concreto, junto a los nodos y enlaces ya representados en la misma, el *módulo validador* incorpora las instancias y propiedades que el enfoque ha descubierto —mediante inferencia semántica— a partir de las nuevas preferencias del espectador. Como consecuencia, la metodología propuesta permite que el sistema reutilice el conocimiento que ha aprendido en el pasado (formalizado inicialmente en la red SA) y, a la vez, adapte el proceso de recomendación a las nuevas preferencias de los usuarios, asegurando con ello tanto la calidad como el carácter personalizado de las sugerencias ofrecidas.

7.7 Cálculo de la relevancia de los contenidos sugeridos

El *módulo de ordenación* mostrado en la Fig. 7.1 utiliza el nivel de activación de los contenidos seleccionados durante el proceso de validación semántica como único criterio de clasificación, de forma que son los programas con niveles más significativos los que ocupan los puestos preferentes de la sugerencia final.

Este sencillo mecanismo es, sin embargo, efectivo, ya que en el valor de los niveles de activación confluyen diferentes parámetros de innegable utilidad a la hora de valorar la importancia de un contenido sugerido. Entre ellos podemos destacar, tal como se aprecia en la Ec. 7.10, el nivel de interés del usuario (índice DOI) o la relevancia de las relaciones establecidas entre éste y sus preferencias (intensidad semántica).

Como consecuencia, el sistema puede reunir en una única sugerencia contenidos de naturaleza muy diversa: algunos relacionados con las preferencias positivas del usuario de forma directa, y otros asociados con ellas mediante relaciones menos inmediatas aunque útiles durante el proceso de recomendación. Tal diversidad, alcanzada en nuestra metodología gracias al razonamiento semántico, es precisamente una de las características que se echa en falta en los enfoques basados en contenido actuales. Los beneficios de este tipo de capacidades inferenciales se ilustran mediante un ejemplo en la siguiente sección.

7.8 Ejemplo de aplicación

Con el objetivo de describir el proceso de recomendación con suficiente claridad, el escenario de aplicación propuesto en este ejemplo asume ciertas simplificaciones razonables. Éstas no impedirán, sin embargo, mostrar los beneficios que aporta el razonamiento semántico a los sistemas recomendadores actuales.

- Por un lado, el perfil del usuario activo considerado contendrá un número reducido de programas y características semánticas. En este escenario, el ejemplo mostrará que, a pesar de disponer de un perfil inmaduro, el razonamiento semántico es capaz de descubrir contenidos relevantes para el usuario[11]. Además, al reducir la cantidad de características semánticas consideradas, es posible incluir extractos gráficos del perfil, claros e intuitivos, que facilitarán la comprensión del lector. En este sentido, destacar que los razonamientos aplicados en este ejemplo

[11] Obviamente, la calidad de estas sugerencias no es comparable a las elaboradas por el sistema cuando el conocimiento sobre las preferencias del espectador es más detallado.

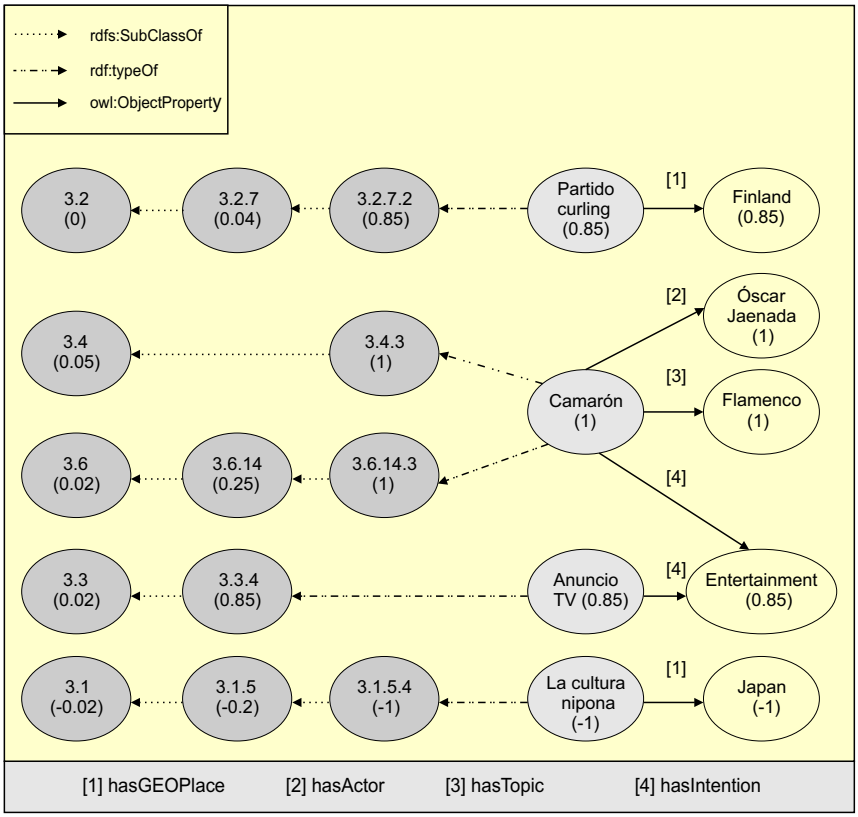

Figura 7.5: Perfil P_U

son fácilmente generalizables a escenarios en los que se incluyen descripciones más precisas de los contenidos de TV.

Las restricciones mencionadas aparecen reflejadas en el perfil mostrado en la Fig. 7.5, correspondiente al usuario activo U considerado en nuestro escenario de aplicación. Se trata de un espectador de nacionalidad finlandesa que ha visto los siguientes contenidos: (i) la película *Camarón*, en la que *Óscar Jaenada* da vida a este célebre intérprete del flamenco, (ii) un anuncio de publicidad sobre la agencia de viajes *Finnish Tours*, (iii) un partido de *curling* y (iv) *La cultura nipona*, un documental sobre las tradiciones culturales japonesas. La Fig. 7.5 representa estos contenidos, los géneros a los que pertenecen, algunas de sus características semánticas y los índices DOI asociados a los mismos (entre paréntesis). Estos índices permiten clasificar los tres primeros programas como preferencias positivas de U y, por otra parte, identificar su desinterés absoluto en relación al documental *La cultura nipona*.

- La segunda de las simplificaciones está relacionada con la base de conocimiento sobre la que se aplica el razonamiento semántico. Ésta se representa en la Fig. 7.6, donde se representan algunos programas de TV junto a sus características semánticas (parte superior de la figura) y a los géneros a los éstos pertenecen en la jerarquía de contenidos de la ontología (parte inferior). El escenario mostra-

do en la Fig. 7.6 también revela los beneficios de los razonamientos geográficos llevados a cabo en el sistema que, en este caso concreto, permiten relacionar el documental *Descubre Andalucía* con el país España.

Este ejemplo ilustrará las tres fases de la estrategia basada en contenido que hemos desarrollado. Por ello, en primer lugar, mostraremos las secuencias de propiedades que selecciona el *módulo de filtrado* a partir de las preferencias del usuario U (incluidas en la Fig. 7.5); a continuación, se presentarán las asociaciones semánticas que descubre el *módulo de inferencia* a partir de dichas secuencias y, finalmente, será el *módulo de validación* quien seleccione los contenidos recomendados a U.

7.8.1 Filtrado de secuencias de propiedades

Tras localizar las preferencias positivas de U en la ontología OWL, el *módulo de filtrado* evalúa las secuencias de propiedades iniciadas en ellas y calcula las intensidades semánticas de sus instancias, filtrando aquellas que no son relevantes para el usuario. Este ejemplo ilustra únicamente el proceso correspondiente a uno de los contenidos que interesaron a U (en concreto, la película *Camarón*), dado que el procedimiento para el resto de programas es análogo al aquí explicado.

Con arreglo al conocimiento formalizado en la Fig. 7.6 (parte superior), el *módulo de filtrado* debe avanzar desde *Camarón* a través de instancias como *Flamenco*, *Carmen*, *La esencia del sur*, *Spain*, etc., y decidir, en función de sus intensidades semánticas, si éstas son relevantes para U.

- *Carmen* y *La esencia del Sur*: Las intensidades semánticas de ambas instancias son suficientes para superar el umbral γ_{Sem_U} gracias a las contribuciones de sus respectivas componentes (ver Sección 7.3.2).
    - Las componentes C_I y C_L de ambas películas toman valores significativos porque: (i) están relacionadas con la película *Camarón* que interesa mucho a U y (ii) esta relación es muy directa, ya que los tres contenidos tratan una misma temática (el flamenco). En concreto, la Ec. 7.3 conduce a una componente $C_I = 1$ y la Ec. 7.4 establece $C_L = \frac{1}{2}$ para ambas películas.
    - La componente C_H también permite detectar que *Carmen* y *La esencia del Sur* son relevantes para U, ya que pertenecen a géneros que interesan al espectador. Tal como muestra la parte inferior de la Fig. 7.6, ambas películas están clasificadas bajo los mismos géneros que *Camarón* (3.4.3 y 3.6.14.3), de ahí que la Ec. 7.6 conduzca a un valor $C_H = 1$ para los dos contenidos ($SemSim_{Hie}$(*Carmen*, *Camarón*) = 1 y DOI_{P_U}(*Camarón*) = 1; ídem para *La esencia del Sur*).
    - Por último, la relevancia semántica (C_R) también toma un valor significativo[12], debido a las relaciones que descubre esta componente entre algunas de las preferencias de U y las dos películas cuya intensidad estamos considerando. En concreto, ambos contenidos están relacionados con *Camarón*

[12]Tal como muestra la Ec. 7.7, la componente C_R se calcula considerando todas las preferencias del usuario y explorando el conocimiento relacionado con ellas en la ontología. Debido a que la base de conocimiento de la Fig. 7.6 omite gran parte de las descripciones de estas preferencias (de hecho, sólo muestra un subconjunto de las relativas a *Camarón*), el ejemplo no recoge los valores concretos de la componente C_R; en su lugar se presentan valoraciones cualitativas.

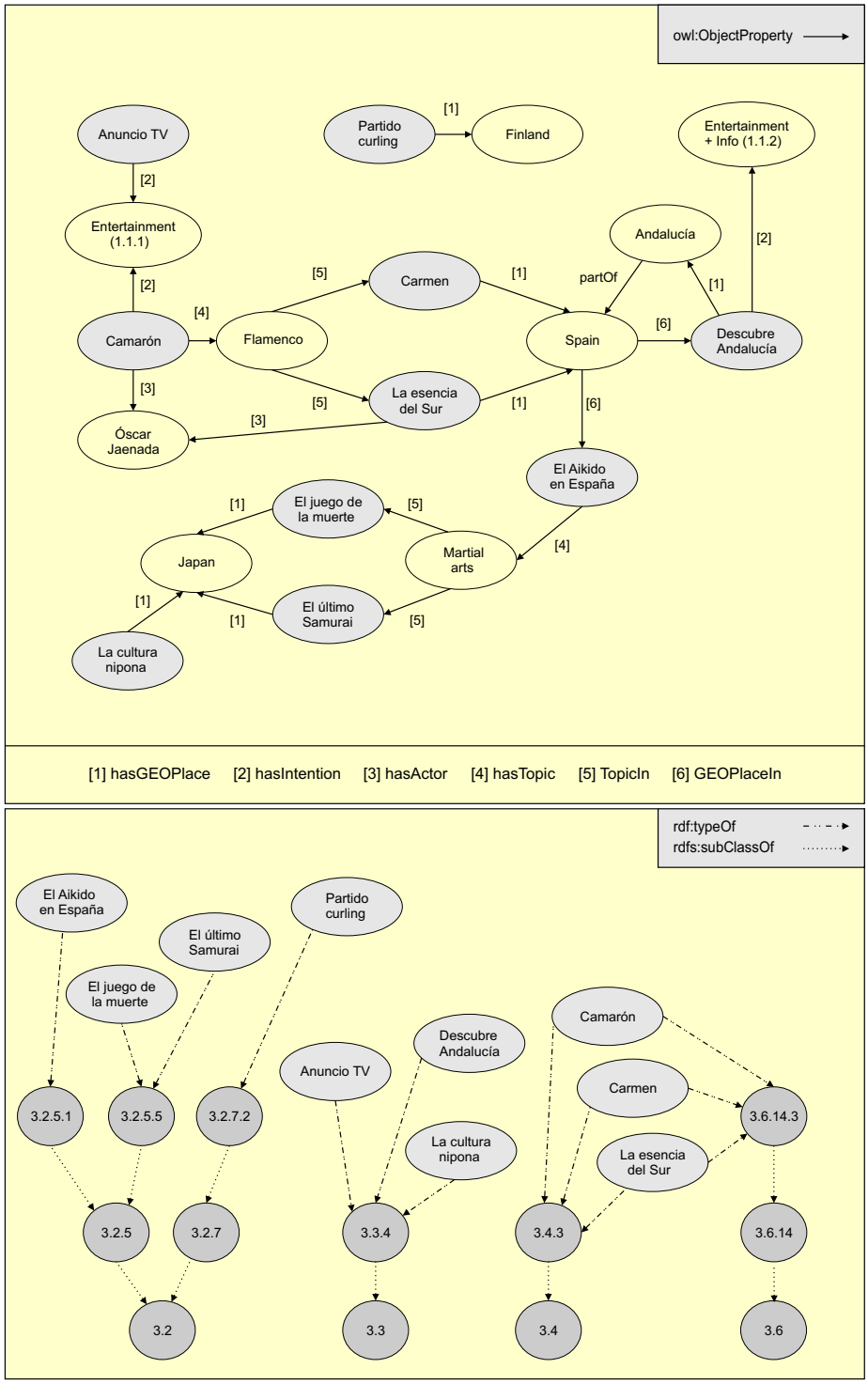

Figura 7.6: Extracto de la ontología de TV OWL

ya que comparten tanto alguno de sus créditos (el actor *Óscar Jaenada*) como la temática del arte flamenco[13].

Reuniendo las cuatro componentes en la Ec. 7.8, resultan sendos valores de intensidad semántica para las películas *Carmen* y *La esencia del Sur* capaces de superar el umbral semántico γ_{Sem_U} establecido en el sistema; de ahí que el *módulo de filtrado* seleccione ambos contenidos y continúe explorando la siguiente instancia en la secuencia representada en la Fig. 7.6.

- *Spain*: Si bien las componentes C_I y C_L permiten incrementar la intensidad de esta instancia, es la componente C_R la que impulsa en mayor medida este valor, provocando con ello su inclusión en la secuencia seleccionada por el *módulo de filtrado*. Así, aunque España no figura entre las características registradas en el perfil de U (ver Fig. 7.5), la relevancia semántica descubre que es un país de interés para este espectador. En concreto, esta componente infiere una relación (implícita) entre dicho país y el arte flamenco que interesa a U, debido a que son varios los contenidos de la ontología OWL que, a la vez que describen este género musical, están relacionados con España[14]. Esta relación contribuye a incrementar la componente C_R de la instancia *Spain* y, con ello, su intensidad semántica, resultando un valor mayor que el umbral γ_{Sem_U}; por esta razón, el *módulo de filtrado* selecciona esta instancia y sigue explorando los nodos de las secuencias mostradas en la Fig. 7.6: *Descubre Andalucía* y *El Aikido en España*.

- *Descubre Andalucía*: Gracias al razonamiento geográfico apuntado al comienzo de la sección, el *módulo de filtrado* alcanza directamente este documental turístico desde la instancia *Spain*. En esta ocasión, es la componente C_H la que contribuye a incrementar la intensidad semántica del contenido. Tal como muestra la Fig. 7.6 (parte inferior), este documental pertenece al mismo género (3.3.4) que el anuncio de agencias de viajes que interesó a U. En este caso, la Ec. 7.6 conduce a una componente C_H de valor 0.85, dado que $M = 1$, $H = \{$*AnuncioTV*$\}$ y DOI_{P_U}(*AnuncioTV*) = 0.85.

  Reuniendo esta componente junto a las restantes (omitidas en la explicación por ser análogas a las descritas anteriormente), resulta una intensidad semántica capaz de superar el umbral γ_{Sem_U}; de ahí que el *módulo de filtrado* explore la siguiente instancia de la secuencia en la Fig. 7.6.

- *Entertainment+Info. (1.1.2)*: El valor de intensidad semántica de esta instancia no es muy significativo, razón por la cual el *módulo de filtrado* considera que es el término en la secuencia seleccionada. Por una parte, la relación con *Camarón* es muy indirecta (dada la secuencia de longitud 5 que une ambas instancias en la Fig. 7.6) y, por otra, la centralidad semántica no permite detectar relaciones significativas entre esta instancia y las preferencias de U. Por el contrario, la componente C_H toma un valor elevado dada la similitud jerárquica entre este

[13]En términos de intermediación (ver Ec. 7.2), tales relaciones se traducen en la existencia de secuencias geodésicas que permiten alcanzar las dos películas consideradas (*Carmen* y *La esencia del Sur*) desde algunas de las preferencias de U (e.g. *Camarón*), a través de otras instancias que también interesan al usuario (e.g. *flamenco* y *Óscar Jaenada*).

[14]En términos de intermediación, en la Fig. 7.6 (parte superior) hay tres secuencias geodésicas entre *Spain* y *Camarón* en las que están incluidas otras preferencias positivas de U, como *Óscar Jaenada* y *flamenco*. Destacar que este ejemplo sólo es ilustrativo; en un escenario real, el sistema haría esta clase de razonamientos sólo cuando son muchas las secuencias que establecen una correspondencia entre el flamenco y nuestro país, asegurando así la fiabilidad de las relaciones inferidas.

nodo y el atributo *Entertainment (1.1.1)* asociado a las preferencias positivas de este espectador.

Tras decidir que el término de una de las secuencias seleccionadas es la instancia *1.1.2*, la metodología de filtrado sigue explorando nodos desde *Spain*. En concreto, alcanza un contenido deportivo en el que se recogen testimonios de jóvenes españoles aficionados al Aikido, un arte marcial de origen japonés.

- *El Aikido en España*: Gracias a su vínculo con España, las componentes C_I y C_L permiten relacionar (aunque de forma bastante indirecta) este programa con la película *Camarón* definida en el perfil P_U. Por su parte, C_H también es significativa debido al interés que despierta el género deportivo en este espectador[15].

  Finalmente, la relevancia semántica toma un valor muy reducido ya que, por una parte, detecta relaciones entre este programa y las preferencias positivas de U y, por otro lado, también asocia este programa con contenidos que no han interesado a este usuario (como *La cultura nipona* en Fig. 7.5). Concretamente, la Fig. 7.6 muestra que la instancia *Spain* permite trazar secuencias geodésicas entre *Camarón* y *El Aikido en España*, en las que están incluidas otras preferencias positivas de U (como *flamenco* y *Óscar Jaenada*). De forma análoga, también la instancia *Japan* permite establecer secuencias de longitud mínima entre el documental deportivo y *La cultura nipona* a través de las preferencias negativas de U (DOI_{P_U} (*Japan*) = -1 en Fig. 7.5).

  Tras reunir todas las componentes en la Ec. 7.8, asumimos que resulta un valor de intensidad semántica moderado, aunque suficiente para superar el umbral de filtrado, de ahí que el módulo analice la siguiente instancia de la secuencia (*martial arts*).

- *Martial arts*: Mediante un razonamiento similar al formulado en la instancia *Spain*, la metodología infiere relaciones semánticas entre el nodo *martial arts* y el contenido *La cultura nipona*. En concreto, el enfoque detecta que las artes marciales no son relevantes para U por ser una tradición propia de Japón, país que, a tenor del nivel de interés -1 que registra en P_U, no interesa al espectador.

  Este razonamiento se refleja en el nivel negativo de la componente C_R asociada a la instancia *martial arts*. Por una parte, las intermediaciones incluidas en la Ec. 7.7 toman valores elevados, ya que, tal como muestra la Fig. 7.6, en las dos secuencias geodésicas existentes entre esta instancia y *La cultura nipona*, también está incluido *Japan*. Sin embargo, al incorporar en dicha expresión los índices de interés de ambas preferencias (DOI_{P_U}(*La cultura nipona*) = DOI_{P_U}(*Japan*) = -1), resultan los valores negativos de C_R mencionados antes. En este ejemplo concreto, el resto de componentes de la intensidad (C_I, C_L y C_H) no compensan dichos valores negativos, razón por la que la intensidad de *martial arts* no es suficiente para superar el umbral semántico establecido. Así, este nodo es el término de la secuencia analizada por el módulo de filtrado, de ahí que no sea necesario seguir explorando las instancias alcanzadas desde éste en la base de conocimiento (*El último Samurai* y *El juego de la muerte* en la Fig. 7.6).

[15] La Fig. 7.6 muestra que el antecesor común entre el documental considerado y el partido de *curling* que interesó al usuario, es el género 3.2, de ahí que su similitud jerárquica se reduzca a 0.5. Combinando este valor con DOI_{P_U}(*Partido curling*) = 0.85 en la Ec. 7.6, resulta una componente $C_H = 0{,}425$.

Una vez concluido el proceso de filtrado, es necesario inferir asociaciones semánticas concretas entre los contenidos incluidos en las secuencias seleccionadas (representadas en la Fig. 7.7).

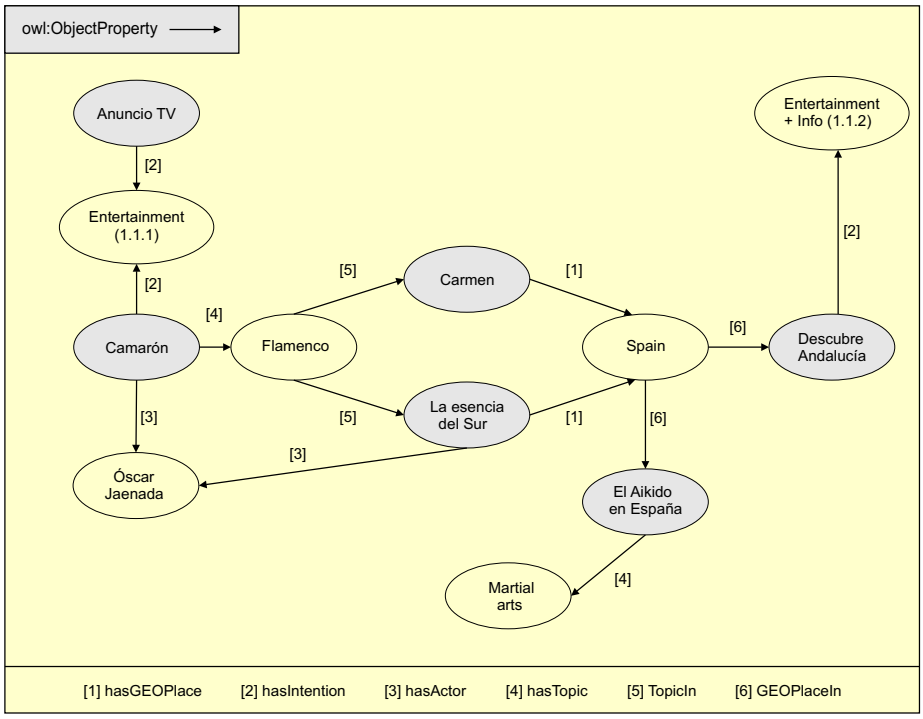

Figura 7.7: Secuencias de propiedades seleccionadas por el módulo de filtrado

7.8.2 Inferencia de asociaciones semánticas

Tal como muestra la Tabla 7.2, el *módulo de inferencia* no sólo descubre asociaciones entre las preferencias de U y otros contenidos de TV, sino también entre los diferentes programas incluidos en las secuencias seleccionadas.

Son estas asociaciones las que debe analizar el *módulo de validación* para poder seleccionar los contenidos recomendados a U. Para ello, este módulo representa el conocimiento aprendido en las dos fases previas, resultando la red de propagación representada en la Fig. 7.8, en la que, por claridad, se representan únicamente algunos de los enlaces físicos y virtuales descubiertos por los *módulos de filtrado* e *inferencia*.

7.8.3 Validación semántica

Sobre esta red SA, el enfoque debe calcular los niveles de activación de los contenidos de TV concretos incluidos en la misma y, en función de éstos, seleccionar los programas finalmente sugeridos.

Tal como describimos en la Sección 7.6.5, dichos niveles serán mayores para aquellos contenidos relacionados —ya sea explícitamente a través de enlaces físicos, o bien

Secuencias de propiedades	Asociaciones semánticas
Camarón - flamenco - Carmen - Spain - - Descubre Andalucía	*ρ-path* (*Camarón*, *Carmen*) *ρ-path* (*Carmen*, *Descubre Andalucía*) *ρ-path* (*Camarón*, *Descubre Andalucía*)
Camarón - flamenco - Carmen - Spain - - El Aikido en España	*ρ-path* (*Carmen*, *El Aikido en España*) *ρ-path* (*Camarón*, *El Aikido en España*)
Camarón - flamenco - La esencia del Sur - - Spain - Descubre Andalucía	*ρ-path* (*Camarón*, *Descubre Andalucía*) *ρ-path* (*Camarón*, *La esencia del Sur*) *ρ-path* (*La esencia...*, *Descubre...*)
Camarón - flamenco - La esencia del Sur - - Spain - El Aikido en España	*ρ-path* (*La esencia del Sur*, *El Aikido...*)
Camarón - Entretenimiento (1.1.1) *Descubre Andalucía - Entre.+Info. (1.1.2)*	*ρ-join* (*Camarón*, *Descubre Andalucía*)

Tabla 7.2: Secuencias de propiedades y asociaciones semánticas inferidas para U

implícitamente a través de enlaces virtuales— con otros programas relevantes para el usuario U. A continuación, revisamos los valores correspondientes a los contenidos incluidos en la red SA de U (ver Fig. 7.8).

- *Carmen* y *La esencia del sur*: De acuerdo a lo explicado en las secciones previas, ambas películas son atractivas para U ya que no sólo tratan la temática del *flamenco* (declarada explícitamente entre sus preferencias positivas), sino que están ambientadas en España, un país que el razonamiento semántico ha descubierto que interesa a este espectador.

  Tal como muestra la Fig. 7.8, los enlaces físicos y asociativos que alcanzan ambas películas contribuyen a incrementar sus respectivos niveles de activación. Los primeros representan relaciones ya formalizadas en la ontología a través de propiedades explícitas, mientras que los segundos permiten que el proceso de recomendación considere conocimiento que estaba oculto en la misma (como la asociación existente entre la película favorita del usuario —*Camarón*— y los dos contenidos considerados).

  Gracias a las contribuciones enumeradas, *Carmen* y *La esencia del Sur* alcanzarían niveles superiores al umbral δ_{Val_U} y, por tanto serían sugeridos a U. Sin embargo, tal como muestra la red SA de este usuario, ambas instancias también reciben enlaces asociativos procedentes de los documentales *Descubre Andalucía* y *El Aikido en España*, analizados a continuación.

- *Descubre Andalucía*: Tal como muestra la Fig. 7.8, este contenido recibe un gran número de contribuciones, debidas tanto a enlaces físicos representados explícitamente en la ontología del sistema, como a enlaces virtuales inferidos por la metodología de razonamiento semántico. Si bien todos ellos contribuyen en el cálculo del nivel de activación del programa, aquí nos centramos únicamente en aquellos cuya influencia es más significativa.

  Por las mismas razones argumentadas anteriormente, el nivel de este documental turístico se ve incrementado fuertemente gracias al enlace físico que lo relaciona con *Spain*, y a los enlaces asociativos procedentes de *Camarón*, *Carmen* y *La esencia del Sur*. El interés de U en la primera película es conocido por el siste-

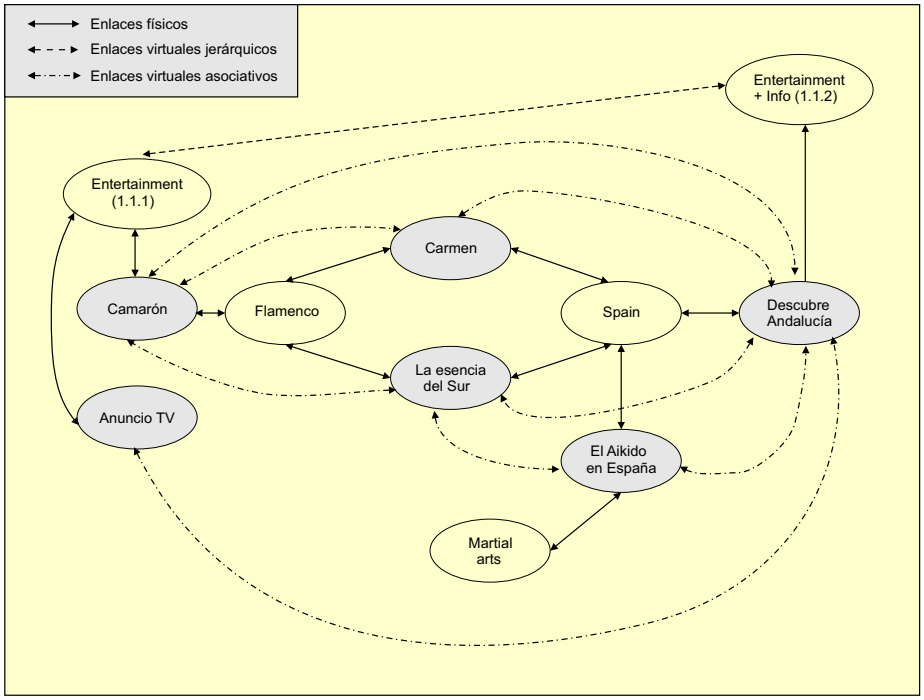

Figura 7.8: Extracto de la red SA del usuario U

ma (recordar que DOI_{P_U}(*Camarón*=1)); por su parte, la importancia de los dos contenidos restantes (inferida por la metodología de razonamiento) ya ha sido justificada anteriormente.

- *El Aikido en España*: Tal como describimos en la Sección 7.8.1, el *módulo de filtrado* ha inferido una relación semántica entre este documental deportivo y las preferencias negativas de U. Su origen son las artes marciales, una tradición muy arraigada en el país de Japón que, tal como muestra el perfil P_U, no interesa en absoluto a este espectador.

  Esta relación semántica contribuye a decrementar el nivel de activación de *El Aikido en España* y, en consecuencia, trata de inhibir la recomendación de este programa a U. Para que se dé esta circunstancia es necesario que el peso del enlace físico que une el documental y la instancia *martial arts* tome un peso negativo. Según las expresiones Ec. 7.10 y Ec. 7.11, dicho peso es el promedio de la intensidad semántica de los dos nodos que conecta. De acuerdo a lo explicado en la Sección 7.8.1, la intensidad del documental es moderada y la correspondiente a *martial arts* toma un valor negativo, de ahí que el peso del enlace físico atenúe el nivel de activación de *El Aikido en España*, tal como mencionábamos antes.

  Sin embargo, con arreglo a las asociaciones mostradas en la Tabla 7.2, este documental deportivo también está relacionado a través de enlaces asociativos (algunos de ellos omitidos en la Fig. 7.8) con otros programas relevantes para U. De forma análoga a la explicada previamente, dichos enlaces contribuyen a aumen-

tar el nivel del documental, contrarrestando con ello el efecto inhibidor de la instancia *martial arts*[16]. De ello se desprende que, para detectar que *El Aikido en España* no es interesante para U, el *módulo de filtrado* debería haber descubierto otras características semánticas que asocien este documental con las preferencias negativas del usuario[17].

7.8.4 El razonamiento semántico vs. los enfoques convencionales

Los enfoques basados en contenido convencionales se limitarían a sugerir a U las películas *Carmen* y *La esencia del Sur*, ya que comparten algunos de sus atributos semánticos: en concreto, la temática y uno de los créditos (*flamenco* y *Óscar Jaenada* en Fig. 7.6). Por el contrario, para diversificar las recomendaciones, AVATAR sugiere, junto a estos contenidos, el documental *Descubre Andalucía*, un programa interesante para U quien, a la vista de los anuncios de agencias de viajes y películas de flamenco que ha visto en el pasado, podría estar interesado en visitar nuestro país.

A diferencia de los enfoques existentes, que comparan los contenidos de TV considerando mecanismos sintácticos más o menos sofisticados (ver Sección 2.6.2), esta recomendación basada en razonamiento semántico explota tanto las relaciones jerárquicas explícitas en la ontología (por ejemplo, la existente entre el anuncio de agencias y el documental turístico *Descubre Andalucía* en la Fig. 7.6), como las asociaciones implícitas inferidas entre las preferencias positivas (y negativas) del usuario y los contenidos finalmente sugeridos.

7.9 Beneficios de la estrategia basada en contenido

Tal como acabamos de ilustrar en el ejemplo descrito en la sección previa, las capacidades de razonamiento semántico del enfoque presentado en esta tesis permiten superar algunas de las limitaciones identificadas en los enfoques de personalización existentes en el estado del arte. En concreto, nuestra aproximación combate, de forma efectiva, tanto los problemas de *sobreespecialización* como el *new user ramp-up* tradicionalmente vinculados a las estrategias basadas en contenido.

La solución más utilizada ante la falta de diversidad asociada a los recomendadores basados en contenido, ha sido recurrir a enfoques de filtrado colaborativo, en los que, en lugar de considerar únicamente las preferencias del usuario activo se tienen en cuenta también los intereses de otros usuarios del sistema. Sin embargo, esta aproximación obliga al recomendador a disponer de un número suficientemente elevado de usuarios, para con ello, poder empezar a ofrecer sugerencias de calidad. Por el contrario, la estrategia propuesta en nuestra metodología diversifica las recomendaciones ofrecidas a los usuarios sin necesidad de recurrir a métodos colaborativos, pudiendo así prescindir de este tipo de restricción y de las limitaciones asociadas a los mismos.

En concreto, nuestro trabajo se enfrenta a la ***sobreespecialización*** de las sugerencias empleando procesos de razonamiento que exploran la semántica de los contenidos

[16]Conviene destacar que la influencia del enlace físico es más significativa, ya que su valor no se ve atenuado por el factor μ_{ij} considerado en los enlaces virtuales (ver Ec. 7.11).

[17]Este es principio en el que se basan los razonamientos utilizados en un entorno de aplicación real, en el que, en lugar de considerar un extracto de la ontología, se tiene en cuenta todo el conocimiento formalizado en la misma.

de TV formalizados en la ontología OWL, e infieren asociaciones semánticas significativas entre las preferencias del usuario y los programas finalmente recomendados. Tales relaciones permiten descubrir que ciertos contenidos son interesantes para el usuario activo, aun cuando no compartan atributos semánticos idénticos a los definidos en su perfil; con ello, nuestra metodología elimina el carácter *sobreespecializado* de las recomendaciones (basadas en contenido) actuales.

El problema del ***new user ramp-up***, presente incluso en los enfoques de personalización más extendidos, ha sido combatido de forma eficaz por Middleton en su sistema recomendador QuickStep [149]. Tal como describimos en la Sección 3.10.3, su aproximación asume que el sistema dispone de una base de conocimiento en la que se representa información tanto sobre los usuarios de la herramienta (investigadores) como sobre los productos que ésta sugiere (artículos de investigación). Así, ante la llegada de un nuevo usuario al sistema, su enfoque accede a la ontología de conocimiento y completa su perfil, añadiendo las publicaciones en las que este autor ha participado, así como trabajos de otros colegas que han colaborado estrechamente con él en el pasado. De esta manera, el sistema dispone de conocimiento preciso sobre las preferencias del usuario recién llegado, facilitando así la elaboración de sugerencias de calidad.

Este enfoque, sin embargo, no es efectivo en nuestra metodología, debido precisamente a que nuestra ontología de conocimiento formaliza únicamente conocimiento sobre contenidos de TV específicos y sus descripciones semánticas, omitiendo información sobre los espectadores del sistema. Precisamente por esta razón, nuestra aproximación requiere un mecanismo alternativo que combata el *new user ramp-up* empleando únicamente la semántica de los programas de TV formalizada en la conceptualización OWL.

Para este fin, la estrategia desarrollada infiere gran cantidad de conocimiento sobre las preferencias de un usuario explotando las descripciones semánticas de los contenidos representados en la ontología. Imaginemos un caso extremo, en el que el perfil de un usuario recién llegado al sistema registra un único contenido de TV. En este escenario, el *módulo de filtrado* podría completar el conocimiento disponible en el perfil, extrayendo de la ontología sus principales características semánticas. Así, explorando las propiedades que permiten asociar dichas características al programa original, este módulo podría seleccionar un conjunto de secuencias de propiedades relevantes para el usuario, y establecer a la vez relaciones con otros contenidos de la ontología con los que completar su perfil inicial (por ejemplo, aquellos que comparten dichas características semánticas). Obviamente, tanto este proceso como el de inferencia y validación realizados posteriormente, serían menos precisos que aquellos en los que las preferencias del usuario fueran más detalladas. Sin embargo, aun en tales condiciones, este enfoque superaría las aproximaciones en las que no es posible inferir conocimiento adicional sobre las preferencias (inmaduras) de un nuevo usuario, por no disponer de una representación formal del conocimiento sobre el dominio de aplicación.

Por último, hacemos una reflexión sobre la naturaleza de las asociaciones semánticas que sustentan nuestro proceso de recomendación. Tal como describimos a lo largo de este capítulo, la estrategia propuesta se basa en recomendar los contenidos que están relacionados (asociados) semánticamente con los que interesan a un usuario dado. Este enfoque podría considerar un conjunto de reglas estáticas (por ejemplo, reglas de asociación [184] o reglas lógicas [67]), en las que se identificasen relaciones bien conocidas entre los programas de TV (utilizando, por ejemplo, un conjunto de entrenamiento inicial). La aplicación de dichas reglas en la estrategia de recomendación sería inmediata y, por tanto, computacionalmente menos costosa que la metodología de fil-

trado, inferencia y validación propuesta en esta tesis. La principal limitación de este tipo de reglas en un entorno de personalización como el nuestro es precisamente su **naturaleza estática**. Como consecuencia de la misma, el sistema emplearía las mismas reglas para todos los usuarios, sin tener en cuenta el conjunto global de preferencias que figuran en su perfil.

Por el contrario, con el claro objetivo de preservar los requisitos de personalización del sistema recomendador AVATAR, las asociaciones semánticas consideradas en esta tesis han sido concebidas como reglas dinámicas, inferidas para cada usuario y adaptadas a sus preferencias personales. Para ello, nuestro enfoque recurre a mecanismos inteligentes que permiten explorar todo el conocimiento disponible sobre un usuario y, en base al mismo, elaborar sus recomendaciones semánticas (esta influencia se manifiesta claramente en la componente de la relevancia semántica descrita en la Sección 7.3.2, en la que se calcula la *intermediación* de una instancia en relación a todas las demás incluidas en el perfil del usuario). Si bien este tipo de inferencias se podrían traducir a reglas como las mencionadas, este proceso conduciría a un número inmanejable de asociaciones estáticas, difíciles de predecir en ausencia de razonamiento semántico, hecho que justifica la utilidad de una metodología como la desarrollada en esta tesis.

Capítulo 8

Estrategia híbrida

Este capítulo describe una estrategia de recomendación híbrida que consta de dos fases: la primera de ellas utiliza el filtrado basado en contenido y sugiere productos similares a los que interesaron a cada usuario; la segunda, inspirada en los principios del filtrado colaborativo, recomienda a los usuarios aquellos productos que resultan atractivos a sus vecinos. La piedra angular sobre la que se asientan ambas fases es la métrica de similitud semántica definida en la estrategia. Gracias a las capacidades de razonamiento de nuestra metodología, esta métrica es capaz de cuantificar el parecido entre dos productos concretos, basándose en las relaciones semánticas (explícitas y ocultas) establecidas entre ellos en la ontología del sistema. Tras presentar las bases teóricas y los detalles algorítmicos de la estrategia, el capítulo ilustra su aplicación con un ejemplo concreto en el dominio del sistema AVATAR. Finalmente, se describen las ventajas principales de esta estrategia frente a los enfoques híbridos actuales.

8.1 Introducción

Tal como acabamos de describir en el capítulo previo, al combatir de forma efectiva el problema de la *sobreespecialización* de sus recomendaciones, nuestro trabajo rescata los métodos basados en contenido, una estrategia de personalización que había caído en desuso. Precisamente, tal limitación fue la causa del ya mencionado abandono de dichas métodos en favor de las técnicas colaborativas primero y de los enfoques híbridos más tarde.

Tal como describe Robin Burke en [42], la mayoría de los enfoques de personalización propuestos en la literatura han optado por combinar el filtrado basado en contenido y las estrategias colaborativas, beneficiándose con ello de las ventajas asociadas a ambas técnicas y mitigando sus respectivas limitaciones (individuales). Este escenario nos parece adecuado para proponer una segunda estrategia de personalización, que pondrá de manifiesto los beneficios que aportan los procesos de razonamiento semántico de nuestra metodología a los sistemas híbridos existentes en la literatura. Mediante tales razonamiento, nuestro trabajo define una métrica de *similitud semántica*, flexible y genérica, que permite comparar contenidos de TV concretos basándose en las asociaciones semánticas descubiertas entre ellos.

En virtud de su naturaleza híbrida, son dos las fases que podemos identificar en la estrategia desarrollada. La primera de ellas utiliza las técnicas basadas en contenido

para sugerir un conjunto de programas objetivo (previamente seleccionados) a aquellos usuarios que hayan mostrado interés por contenidos similares a éstos. Los usuarios que no cumplan este requisito se convierten en candidatos para la segunda etapa de la estrategia que, al basarse en filtrado colaborativo, sugiere los contenidos objetivo a todos aquellos usuarios candidatos cuyos vecinos estén interesados en estos programas.

La ejecución de esta técnica de personalización es una tarea encomendada al agente híbrido representado en la Fig. 4.4. Los módulos que integran dicho agente, asegurando el correcto funcionamiento de la estrategia, se muestran en la Fig. 8.1. En ésta podemos identificar tres bloques genéricos: el *módulo basado en contenido*, el *módulo basado en filtrado colaborativo* y el *módulo de ordenación*. Este último clasifica los contenidos sugeridos al usuario y decide el orden que éstos deben ocupar en la recomendación final, en función de su relevancia para el mismo. Los dos primeros se encargan de implementar cada una de las dos fases descritas previamente; tal como se representa en la Fig. 8.1, delegan gran parte de sus funciones en módulos independientes, cuya descripción se aborda en las siguientes secciones.

8.2 Fase de filtrado basado en contenido

Debido a la aplicación de los métodos basados en contenido, esta etapa considera únicamente las preferencias personales de cada usuario activo y, en función de las mismas, valora en qué medida un determinado contenido objetivo podría ser adecuado para éste. Nuestro enfoque cuantifica dicha medida de adecuación mediante el llamado nivel de *matching* semántico.

8.2.1 Nivel de *matching* semántico

Dados un contenido objetivo a y un usuario activo U que no conoce tal programa, esta etapa cuantifica un nivel de *matching* semántico entre a y las preferencias (actuales) de U registradas en su perfil P_U, representado como $matching(a, U)$. Para este propósito, el módulo *matching semántico* representado en la Fig. 8.1 debe recurrir a alguna métrica que estime la similitud entre el contenido objetivo y aquellos programas del perfil P_U que estén incluidos en la ventana temporal actualmente considerada para el usuario U (ver Sección 2.5.3).

Para ello, nuestra metodología define una métrica llamada *similitud semántica* que permite detectar que dos contenidos son parecidos sin necesidad de compartir los *mismos* atributos semánticos. Para conferir tal flexibilidad a la métrica, este trabajo aprovecha las capacidades de razonamiento semántico de nuestra metodología, que establecen que dos programas son similares cuando es posible inferir relaciones semánticas significativas entre ellos.

Una vez definida la noción de similitud semántica, nuestro enfoque asume que el nivel de *matching* semántico entre U y el contenido a debe ser mayor cuanto más se parezca éste a los programas que más interesan a este usuario. Ésta es precisamente la premisa que inspira la Ec. 8.1.

$$matching(a, U) = \frac{1}{\#N_U} \sum_{i=1}^{\#N_U} |SemSim(a, c_i)| \cdot DOI_{P_U}(c_i) \tag{8.1}$$

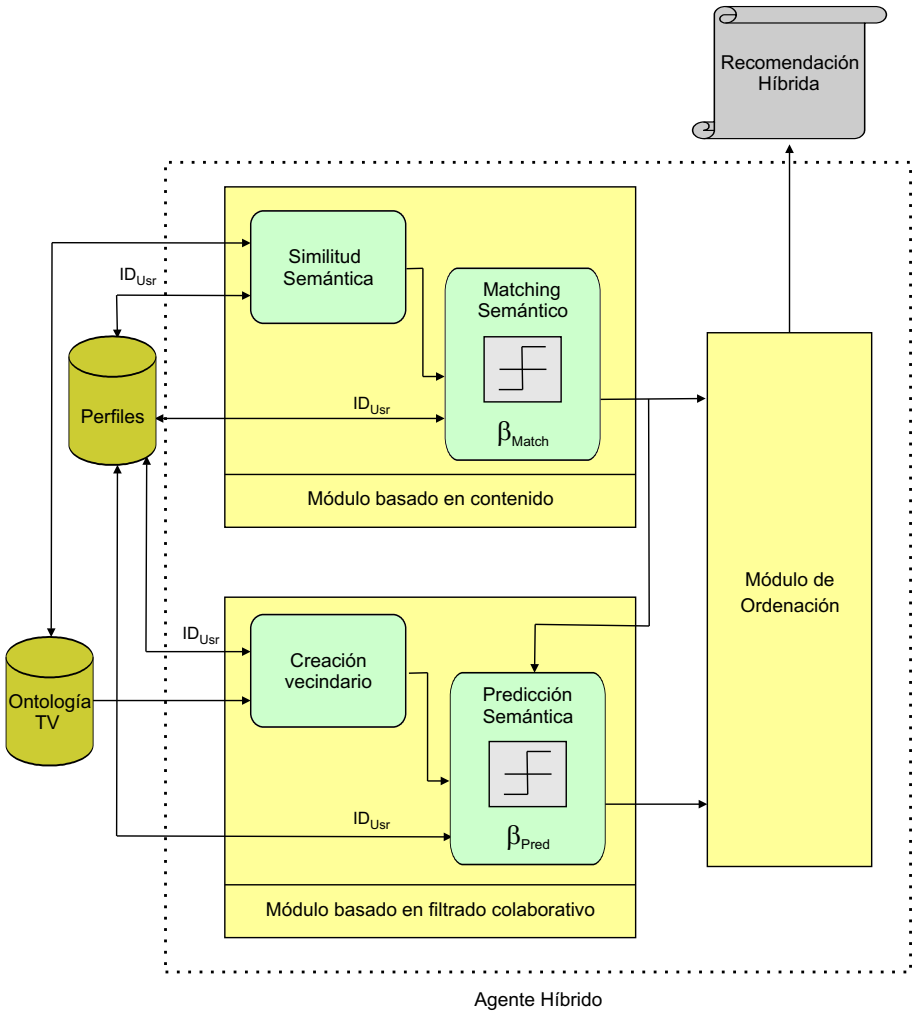

Figura 8.1: Estructura del agente híbrido representado en la Fig. 4.4

donde c_i, $DOI_{P_U}(c_i)$, $\#N_U$ y $SemSim(a, c_i)$, representan el contenido i-ésimo definido en el perfil P_U, el nivel de interés de U en relación a dicho programa, el número total de contenidos almacenados en su perfil y la similitud semántica entre c_i y el programa objetivo, respectivamente.

8.2.2 Similitud semántica

La cuantificación de la similitud semántica es una tarea encomendada al módulo homónimo en la Fig. 8.1. Este módulo debe acceder tanto a los perfiles de los usuarios como a la ontología de TV, para así poder entregar un valor de similitud concreto al bloque que cuantifica los niveles de *matching* semántico. Esto último es debido a que, tal como adelantamos en la sección previa, la similitud semántica entre dos contenidos específicos depende de las relaciones semánticas inferidas entre ambos en la ontología

del sistema.

Tal como describimos en la Sección 3.10.2, Rada [201], Ganesan [77], Resnik [206] y Lin [138], entre otros, propusieron distintas métricas de similitud semántica para comparar los conceptos de un grafo, basándose únicamente en las relaciones de herencia explícitas establecidas entre los mismos. Dado que nuestra metodología cuenta con una ontología OWL —en la que se definen tanto relaciones jerárquicas como propiedades entre las instancias—, la métrica de similitud semántica utilizada puede explotar este conocimiento e inferir relaciones más complejas (como las asociaciones semánticas). Esto último es especialmente importante en nuestro enfoque, dado que cuanto mayor sea el número de relaciones semánticas descubiertas entre los contenidos comparados, más precisos serán los valores de similitud semántica cuantificados y, por tanto, más fiable el nivel de *matching* calculado entre U y el contenido objetivo a.

En esta misma dirección, podemos destacar el enfoque propuesto en [141], también revisado en la Sección 3.10.2. En este trabajo, Maguitman *et al.* cuantifican la similitud semántica entre documentos Web clasificados bajo el directorio ODP. Junto a los enlaces jerárquicos, este enfoque contempla enlaces simbólicos y relacionales, de ahí que dé cabida en su métrica a relaciones de naturaleza más compleja que las jerárquicas mencionadas anteriormente. Sin embargo, esta propuesta tampoco es suficiente en nuestra metodología, ya que únicamente permite cuantificar la similitud entre instancias pertenecientes a una única clase en la jerarquía de conocimiento considerada. Por el contrario, tal como se describió en el capítulo 5, nuestra ontología de TV soporta la clasificación múltiple de las instancias, esto es, permite que éstas pertenezcan a varias de las clases definidas en la jerarquía de géneros del sistema (ver Fig. 5.1). Por otra parte, nuestra metodología de razonamiento no requiere un modelo de conocimiento como el descrito en [141], en el que las relaciones existentes entre los documentos (formalizadas mediante enlaces simbólicos y relacionales) deben ser identificadas antes de computar la similitud semántica; en nuestro trabajo, el descubrimiento de tales relaciones y su posterior aplicación en el proceso de cuantificación, es llevado a cabo automáticamente por el módulo *similitud semántica* mostrado en la Fig. 8.1.

Tal como se describió en el capítulo 3, tanto el enfoque de Maguitman *et al.* como los descritos en [160] y [86] incluyen una ontología de conocimiento en el proceso de cuantificación de la similitud semántica. Sin embargo, todos ellos comparten una misma limitación, debido a que consideran principalmente la estructura jerárquica de la ontología y solamente algunas de las propiedades explícitamente representadas en la misma, omitiendo la inferencia del tipo de relaciones perseguidas en nuestra metodología. Tales relaciones están ocultas en la base de conocimiento e involucran tanto las propiedades formalizadas en la conceptualización como las instancias entre las que éstas se establecen.

Habida cuenta de las limitaciones de las que adolecen los trabajos existentes en el estado del arte, esta tesis redefine la métrica de la similitud semántica, persiguiendo una aproximación, flexible y generalizable, basada en el conocimiento formalmente representado en la ontología OWL del sistema. Dicha métrica, descrita en [26], considera tanto el conocimiento explícito representado en la conceptualización OWL mediante relaciones jerárquicas, como el implícito que puede ser inferido desde la misma a partir de las propiedades establecidas entre sus instancias. La primera de las contribuciones puede ser cuantificada fácilmente mediante la *similitud semántica jerárquica* ($SemSim_{Hie}$), introducida en la Sección 7.3.1 y utilizada en la metodología de razonamiento empleada en la estrategia de recomendación descrita en el capítulo previo. Sin embargo, para poder soportar la inferencia de conocimiento implícito,

dicha métrica debe ser extendida. Para este propósito, nuestra aproximación introduce una nueva componente en la similitud semántica, denominada *similitud inferencial* ($SemSim_{Inf}$). La contribución de cada una de ellas al valor finalmente medido se modula mediante un parámetro de combinación α, tal como se muestra en la Ec. 8.2, en la que se cuantifica la similitud semántica entre los contenidos a y c_i.

$$SemSim(a, c_i) = \alpha \cdot SemSim_{Inf}(a, c_i) + (1 - \alpha) \cdot SemSim_{Hie}(a, c_i) \quad (8.2)$$

Las expresiones matemáticas de ambas componentes son presentadas a continuación e ilustradas mediante diferentes ejemplos.

Similitud semántica jerárquica

En la Ec. 8.3, particularizamos la métrica de similitud jerárquica presentada en la Sección 7.3.1 para el caso en el que las instancias comparadas son los programas a y c_i.

$$SemSim_{Hie}(a, c_i) = \frac{depth(LCA_{a,c_i})}{max(depth(a), depth(c_i))} \quad (8.3)$$

Esta expresión conduce a valores de similitud nulos cuando el único antecesor que comparten ambos programas en la jerarquía de géneros de la ontología es la clase raíz de la misma. En caso de que esta premisa no se cumpla, el valor es más significativo cuanto más específicos (profundos) son ambos programas y cuanto más cerca de ellos está su antecesor común en dicha jerarquía.

Similitud semántica inferencial

Tal como adelantamos en las secciones previas, la similitud inferencial está basada en descubrir relaciones implícitas entre los contenidos comparados, utilizando para ello las propiedades definidas en la ontología de TV. Es precisamente por esta razón por lo que el proceso de inferencia se basa en las asociaciones semánticas utilizadas en la estrategia descrita en el capítulo anterior. En concreto, las asociaciones consideradas en nuestra métrica de similitud inferencial son menos indirectas que las inferidas en la estrategia basada en contenido (ver Sección 7.5). Para justificar la naturaleza de tales asociaciones, basta recordar el propósito de la métrica desarrollada en nuestro trabajo.

De acuerdo a lo explicado en el capítulo 7, el objetivo de la estrategia basada en contenido es descubrir los programas que están asociados semánticamente con los que interesaron al usuario en el pasado, superando así el carácter *sobreespecializado* de las recomendaciones tradicionales. Por el contrario, en la estrategia híbrida es el enfoque colaborativo el que diversifica las sugerencias, de forma que la fase basada en contenido únicamente debe cuantificar el parecido entre los contenidos objetivo (que ya han sido previamente seleccionados) y las preferencias del espectador. De esta forma, exigiendo asociaciones directas entre los contenidos comparados, garantizamos la fiabilidad de los valores de similitud cuantificados por la métrica. Esta es la razón que nos lleva a imponer una serie de restricciones sobre las asociaciones semánticas definidas en la Sección 7.5, que no impiden, sin embargo, que la métrica propuesta explore tanto las relaciones jerárquicas consideradas en los enfoques tradicionales, como el conocimiento oculto en la ontología del sistema.

En concreto, la similitud inferencial descubre únicamente asociaciones *ρ-path* y *ρ-join* originadas a partir de secuencias de propiedades de longitud muy reducida, cuyos orígenes y términos son precisamente los contenidos comparados. Ambos tipos de asociaciones comparten una misma intuición, que establece que dos programas a y c_i están relacionados cuando comparten características semánticas (créditos, alertas de contenido, temática, intención o propósito, información geográfica y temporal, etc.), tal como se aprecia en las definiciones presentadas a continuación:

- **Asociación** ***ρ-path***(a,c_i): Este tipo de asociaciones se infieren a partir de una secuencia de propiedades de longitud dos como la representada en la Fig. 8.2, donde la instancia intermedia (denominada *instancia de unión*) identifica una característica semántica presente tanto en a como en c_i.

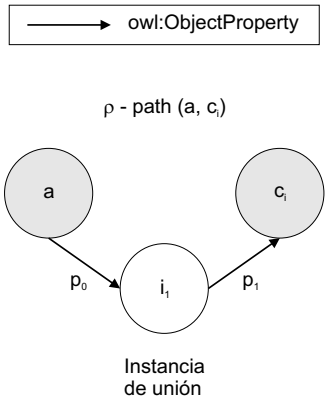

Figura 8.2: Instancia de unión entre los contenidos a y c_i

Ejemplo: Consideremos el extracto de la ontología de TV representado en la Fig. 8.3. En este escenario, es posible identificar una relación implícita ([A] en la figura) entre los documentales *España durante la Segunda Guerra Mundial* y *El Franquismo*. Tal como muestra la Fig. 8.3, la inferencia de esta asociación está basada en los procesos de razonamiento temporal aplicados en el enfoque, que descubren que la Segunda Guerra Mundial estalló durante el régimen franquista[1].
Por el contrario, los enfoques jerárquicos convencionales no descubrirían relación alguna entre los dos programas mencionados, debido a que no existe ninguna categoría (diferente del nodo raíz) que sea antecesor común de sus respectivas clases. Como consecuencia de ello, este tipo de propuestas cuantificarían una similitud nula entre los dos documentales considerados, a pesar de estar implícitamente relacionados a través de una instancia de unión, tal como infiere nuestra métrica.

- **Asociación** ***ρ-join***(a,c_i): Esta asociación es inferida a partir de dos secuencias de longitud uno como las mostradas en la Fig. 8.4. En este caso, los orígenes

[1]Conviene destacar que la similitud inferencial cuantificada en nuestra metodología considera un gran número de características semánticas, de forma que dos programas son similares únicamente cuando comparten un número significativo de atributos. En este ejemplo, proponemos un escenario muy simple para ilustrar el tipo de asociaciones inferidas en el enfoque propuesto.

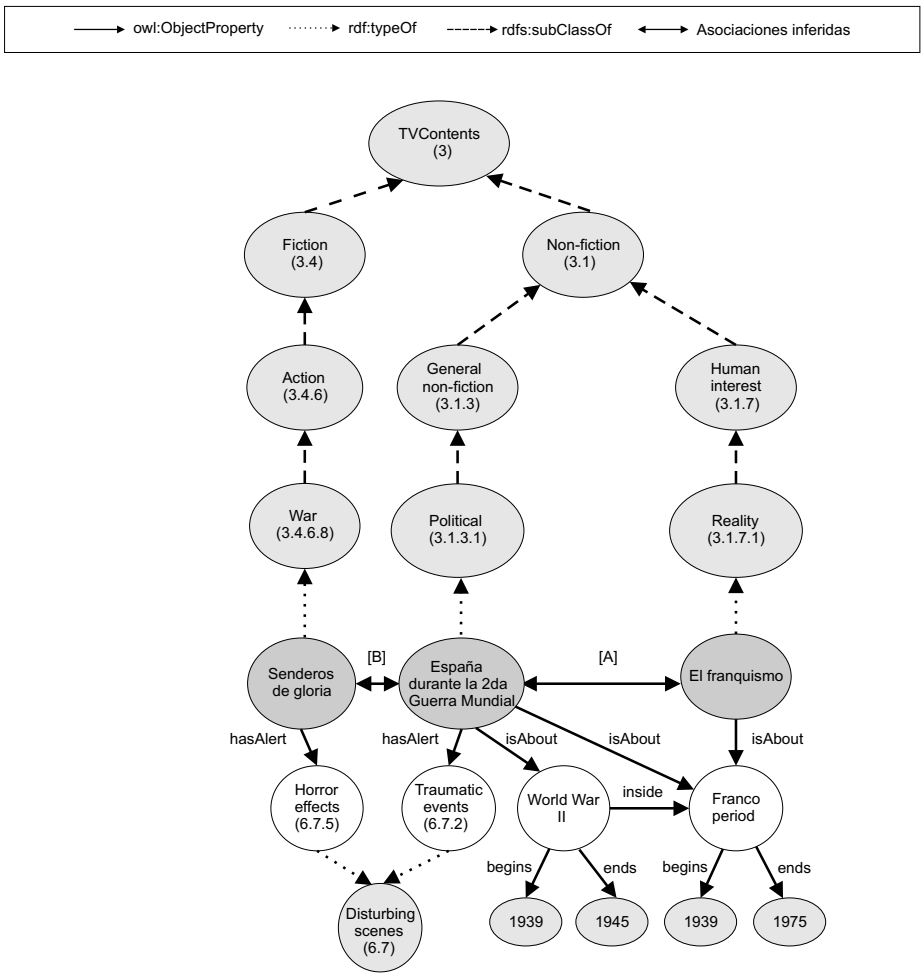

Figura 8.3: Algunas asociaciones semánticas utilizadas en el cálculo de la similitud semántica inferencial

de ambas secuencias son los programas comparados y los términos identifican sendas instancias de una misma *clase de unión* (de ahí la asociación *ρ-join*).

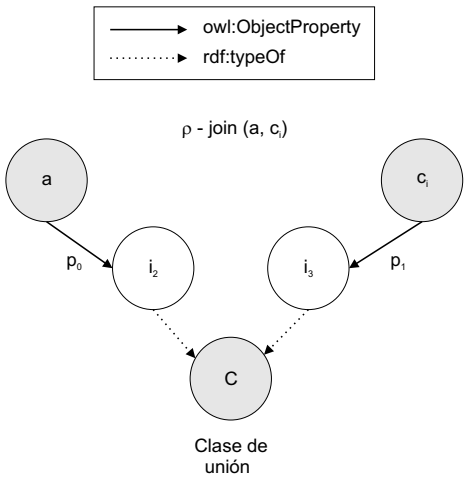

Figura 8.4: Clase de unión entre los contenidos a y c_i

Ejemplo: En la Fig. 8.3 es posible identificar una asociación implícita entre el documental *España durante la Segunda Guerra Mundial* y la película *Senderos de gloria* (representada como [B]). Ésta es debida a que ambos programas contienen escenas que pueden herir la sensibilidad del usuario. Gracias a esta alerta de contenido, y a pesar de que los metadatos TV-Anytime describen atributos diferentes para ambos programas (6.7.2 y 6.7.5, respectivamente), nuestro enfoque es capaz de relacionarlos semánticamente mediante la clase de unión 6.7.

De forma genérica, a las instancias que comparten los contenidos comparados las denominaremos, en lo que sigue, *instancias comunes*. Con arreglo a las asociaciones consideradas en nuestra métrica, dichas instancias pueden ser: (i) instancias de unión, como i_1 en la Fig. 8.2, o (ii) las instancias de la clase de unión que están definidas en el perfil del usuario activo, como i_3 en la Fig. 8.4 (recordar que, en esta figura, a es el contenido objetivo y c_i uno de los programas definidos en el perfil del usuario U).

Para poder establecer el máximo número de instancias comunes entre dos programas, rescatamos el concepto de *grado* de la teoría de grafos [64]; éste define el grado de un nodo como el número total de enlaces asociados al mismo. En nuestro dominio de aplicación, el grado de un contenido de TV corresponde al número de características semánticas asociadas al mismo en la ontología, sin considerar en el cómputo todas aquellas que hayan sido inferidas mediante los procesos de razonamiento temporal y geográfico descritos en el capítulo 5, tal como explicaremos más adelante[2]. De acuerdo a tal definición, el máximo número de instancias comunes entre dos contenidos viene dado por el mínimo de sus respectivos grados.

[2]Aunque dichas características no se incluyen en el grado de un contenido, participan en el proceso de inferencia de relaciones semánticas en el que se basa la cuantificación de la similitud semántica.

Son varias las premisas que hemos considerado para poder dar forma a la expresión matemática de la similitud inferencial:

1. En primer lugar, el valor de dicha métrica debería ser más elevado cuanto mayor sea el número de instancias comunes entre el contenido objetivo y aquellos programas que más interesan al usuario activo.
2. Por otra parte, cuanto más específicas sean las clases a las que pertenecen dichas instancias comunes, más elevado debería ser el valor de similitud (inferencial) cuantificado. De hecho, al ocupar dichas clases niveles muy profundos en la jerarquía a la que pertenecen, el enfoque dispone de conocimiento muy preciso para estimar la similitud, preservando con ello la fiabilidad de los valores obtenidos.
3. La última de las premisas consideradas establece que las asociaciones inferidas mediante instancias de unión son más significativas que las descubiertas mediante clases de unión. En el primer caso, los contenidos comparados comparten exactamente los mismos atributos, mientras que en el segundo escenario, estas características semánticas, si bien pertenecen a una misma clase, son diferentes.

Para poder incluir dichas premisas en el cálculo de la similitud inferencial, nuestro trabajo asocia dos umbrales a las instancias comunes consideradas durante el proceso de cuantificación.

- *Factor de atenuación* (f_{at}): De acuerdo a la segunda de las premisas, el valor del factor de atenuación de una instancia común dependerá de la profundidad de la clase a la que ésta pertenezca. Así, cuando esta clase ocupa el último nivel en su jerarquía (i.e. es clase hoja), su valor es 1; en caso contrario, el factor es proporcional a su profundidad.
- *Factor de penalización* (f_p): Con arreglo a la tercera premisa, el objetivo de este factor es penalizar los valores de similitud inferencial obtenidos a partir de las clases de unión existentes entre los programas comparados.

Reuniendo las ideas y los parámetros presentados hasta este punto, resulta la expresión mostrada en la Ec. 8.4, donde $\#CI(a,b)$ es el número de instancias comunes entre a y c_i, i_k es la k-ésima de estas instancias, $\#CI_{MAX}(a,b)$ es el numero máximo de posibles instancias comunes entre ambos programas, $f_p(i_k)$ y $f_{at}(i_k)$ son los factores de penalización y atenuación, respectivamente, asociados a la instancia común i_k, y C_{i_k} la clase a la que ésta pertenece.

$$SemSim_{Inf}(a,c_i) = \frac{1}{\#CI_{MAX}(a,c_i)} \sum_{k=1}^{\#CI(a,c_i)} f_{at}(i_k) \cdot f_p(i_k) \cdot DOI_{P_U}(i_k) \quad (8.4)$$

donde:

$$\#CI_{MAX}(a,c_i) = min(grado(a), grado(c_i))$$

$$f_{at}(i_k) = \begin{cases} 1 & \text{si } C_{i_k} \text{ es clase hoja} \\ 1 - \frac{1}{depth(C_{i_k})} & \text{en otro caso} \end{cases}$$

$$f_p(i_k) = \begin{cases} 1 & \text{si } i_k \text{ es instancia de unión} \\ 0{,}85 & \text{si } i_k \text{ es instancia de una clase de unión} \end{cases}$$

Influencia de los razonamientos geográficos y temporales en la similitud semántica

Mediante los procesos de razonamiento geográfico-temporales mencionados previamente, el agente híbrido es capaz de descubrir instancias de unión entre los contenidos comparados y, en consecuencia, incrementar su valor de similitud inferencial. Dichas instancias de unión identifican emplazamientos (geográficos o temporales) que comparten los contenidos; por ello, ayudan a establecer vínculos entre los programas que, aunque ocultos en la base de conocimiento, pueden ser útiles para el usuario.

Por ejemplo, supongamos que nuestro enfoque debe cuantificar la similitud inferencial entre los contenidos p_1 y p_2 mostrados en la Fig. 8.5, extraída del capítulo 5.

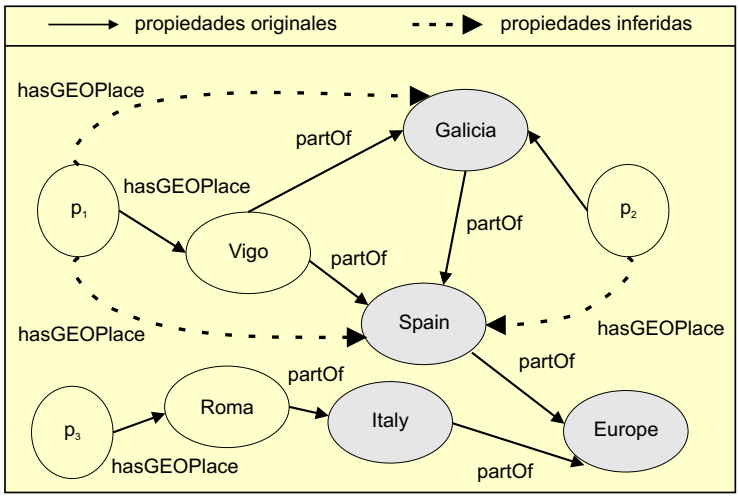

Figura 8.5: Ejemplo de razonamiento geográfico

En este caso, el grado de ambos programas es 1, ya que, de acuerdo a los metadatos TV-Anytime asociados a los mismos, ambos contenidos están únicamente relacionados con la ciudad *Vigo* y con la comunidad *Galicia*, respectivamente. Sin embargo, tras el proceso de razonamiento geográfico llevado a cabo por el agente de clasificación, el sistema infiere que ambos programas están relacionados con dicha comunidad autónoma, de ahí que el enfoque considere que *Galicia* es la instancia de unión entre ellos[3].

[3]Aunque en la Fig. 8.5 es posible identificar dos instancias de unión entre p_1 y p_2 (*Galicia* y *Spain*), la similitud inferencial únicamente considera la más específica de ellas, por ser la que permite inferir una asociación más significativa entre los contenidos comparados. La segunda instancia de unión también es útil en nuestra metodología de inferencia, y más concretamente en la primera de las estrategias incluidas en la misma (descrita en el capítulo previo), ya que habilita una secuencia de propiedades entre los contenidos p_1 y p_2 que puede ser explorada durante el proceso de razonamiento semántico.

Evaluación de la similitud semántica

La métrica mostrada en la Ec. 8.4 es similar a la propuesta por Rocha *et al.* en su herramienta de búsqueda semántica basada en técnicas de propagación de activación [209]. Tal como describimos en la Sección 3.10.1, dichos autores proponen una métrica de similitud para cuantificar el peso de cada enlace establecido entre dos nodos de la red, según la cual dos conceptos son similares cuando comparten muchas relaciones con los mismos nodos de la red (ver Ec. 3.44). Este enfoque se corresponde precisamente con la inferencia de relaciones *ρ-path* en nuestra metodología. Sin embargo, a diferencia de [209], nuestro trabajo contempla también el descubrimiento de asociaciones más complejas (como las *ρ-join*) durante el proceso de cuantificación de la similitud semántica.

Una vez presentadas las expresiones que permiten formalizar la métrica de similitud semántica propuesta, es conveniente evaluar algunas de sus propiedades.

En este sentido, destacamos tanto su simetría ($SemSim(a, c_i) = SemSim(c_i, a)$) como la capacidad de cuantificar un valor de similitud máximo entre dos contenidos específicos, aun cuando éstos no sean necesariamente iguales. Para clarificar esta última propiedad, es necesario analizar las dos componentes definidas en la métrica:

- Por una parte, la Ec. 8.3 conduce a valores de similitud jerárquica máxima entre a y c_i ($SemSim_{Hie}(a, c_i) = 1$) cuando ambos contenidos son instancias de un mismo género en la taxonomía representada en la Fig. 5.1. Por esta razón, no es necesario que los dos programas sean iguales (es decir, la misma instancia) para así cuantificar un parecido máximo entre ellos.

- Por otro lado, la evaluación de la Ec. 8.4 conduce a valores de similitud inferencial máximos entre a y c_i ($SemSim_{Inf}(a, c_i) = 1$) cuando: (i) ambos programas comparten todas sus posibles características semánticas ($\#CI_{MAX}(a, c_i) = \#CI(a, c_i)$ en la Ec. 8.4), (ii) las instancias comunes definidas en el perfil del usuario activo tienen un nivel de interés máximo ($DOI_{P_U}(i_k) = 1 \; \forall \; k$), y (iii) no influyen los factores de atenuación.

Como conclusión, destacar que las propiedades de nuestra métrica de similitud semántica se derivan fácilmente del proceso de razonamiento semántico aplicado en nuestra metodología: en lugar de centrarse en los contenidos comparados en sí mismos, este proceso explora sus descripciones semánticas, de ahí la posibilidad de detectar que dos programas son semánticamente idénticos aun siendo diferentes.

8.2.3 Criterio de recomendación: rango del umbral β_{Match_U}

Una vez incluida la contribución de la similitud semántica en la Ec. 8.1, la estrategia puede computar niveles de *matching* semántico concretos entre las preferencias de cada usuario activo y el contenido objetivo. Esta expresión conduce a un *matching* perfecto (de valor 1) cuando el contenido objetivo a es (semánticamente) idéntico a los programas del perfil P_U que, además, registran niveles de interés máximos.

A continuación, es necesario establecer algún criterio para decidir, en base a los niveles calculados, si el programa a es sugerido o no al espectador activo U. Para este propósito, nuestro trabajo recurre a un umbral β_{Match_U}, de forma que a sólo es recomendado a U cuando su nivel de *matching* supera el valor establecido para el

mismo. De forma completamente análoga a la descrita en la Sección 7.6.6 para el umbral δ_{Val_U}, dicho valor será ajustado dinámicamente para cada usuario U en función de la precisión que obtuvo del sistema en las últimas recomendaciones. Además, este proceso también adopta el rango de operación utilizado para el umbral de validación, imponiendo, por tanto, las cotas ω_3 y ω_2 mostradas en la Ec. 8.5.

$$\beta_{Match_U}(t{+}1) = \begin{cases} \omega_3 & \text{si} \quad \beta_{Match_U}(t) - \Delta_P(t+1,t) < \omega_3 \\ \omega_2 & \text{si} \quad \beta_{Match_U}(t) - \Delta_P(t+1,t) > \omega_2 \\ \beta_{Match_U}(t) - \Delta_P(t+1,t) & \text{en otro caso} \end{cases} \tag{8.5}$$

donde $\Delta_P(t+1,t) = P(t+1) - P(t)$.

En virtud de su naturaleza híbrida, la estrategia propuesta da una segunda oportunidad a todos aquellos usuarios cuyos niveles de *matching* no superen el umbral β_{Match_U} establecido. Éstos se convierten así en espectadores candidatos en la fase colaborativa incluida en el enfoque, siempre que el filtrado basado en contenido no detecte que el programa objetivo es similar a sus preferencias negativas. En este caso —en el que el nivel de *matching* cuantificado alcanza valores negativos (ver Ec. 8.1)—, la metodología de razonamiento semántico infiere el desinterés del usuario en relación al contenido objetivo, de ahí que omita la etapa colaborativa y no sugiera dicho programa.

8.3 Fase basada en filtrado colaborativo

En esta fase, la estrategia propuesta *predice* el nivel de interés de cada usuario candidato en relación a los contenidos objetivo. De acuerdo a los principios básicos del filtrado colaborativo, para obtener esta medida de *predicción semántica* es necesario analizar las preferencias del resto de usuarios del sistema recomendador.

En primer lugar, debemos identificar el tipo de enfoque colaborativo —basado en usuario o basado en *item*— que vamos a incluir en la estrategia. Tal como se describió en la Sección 2.6.3, el primero selecciona aquellos usuarios que comparten las preferencias de cada usuario candidato. Por el contrario, el segundo modelo extrae, de entre todos los contenidos que han visto los espectadores del sistema, aquellos que son más similares a los que han interesado al usuario candidato en el pasado.

Nuestro enfoque debe aprovechar los beneficios del razonamiento semántico para combatir las limitaciones de ambas aproximaciones. Los enfoques convencionales basados en *item* establecen que dos contenidos son similares si aparecen juntos (y con índices de interés similares) en un número significativo de perfiles de usuarios. Para superar los efectos del *sparsity problem* que se desprenden de esta aproximación, nuestro trabajo debería considerar las descripciones semánticas de los programas comparados, de forma que los valores de similitud no dependan únicamente de la frecuencia de aparición de los contenidos en los perfiles del sistema. De acuerdo a lo explicado en la Sección 8.2.2, en este contexto, nuestro enfoque calcularía la similitud semántica entre los contenidos del perfil activo y los definidos en el resto de los perfiles disponibles, lo que se traduce en un coste computacional difícil de asumir a medida que aumenten el número de usuarios del sistema y el número de programas definidos en sus perfiles.

Para aliviar tales costes, nuestro enfoque debería detectar similitud entre las preferencias de los usuarios, sin necesidad de comparar los contenidos específicos incluidos en sus perfiles. Éste es precisamente el planteamiento que explora la estrategia colabo-

rativa (basada en usuario) desarrollada en nuestro trabajo, que incorpora procesos de razonamiento semántico en las tres fases que caracterizan su funcionamiento:

- Selección de los vecinos de cada usuario candidato.
- Establecimiento del tamaño de dicho vecindario.
- Predicción semántica basada en los niveles de interés de los usuarios vecinos en relación a los contenidos objetivo.

8.3.1 Selección de usuarios con preferencias similares

Tal como se describió en la Sección 2.6.3, en la literatura se han propuesto tres métodos diferentes para comparar las preferencias de los usuarios. Éstos incluyen los enfoques basados en clasificadores automáticos, el *clustering* y las técnicas de selección de los vecinos más cercanos. La adopción de esta última técnica en nuestro trabajo responde a dos razones principales. La primera de ellas se basa en los resultados obtenidos por Breese *et al.* en [35], un trabajo que revela que, en presencia del *sparsity problem*, el *clustering* ofrece recomendaciones menos precisas que los métodos basados en la selección de los vecinos más cercanos. Por otro lado, las técnicas basadas en clasificadores no son útiles en nuestra metodología de razonamiento, ya que estos modelos omiten la semántica de los productos considerados en el sistema recomendador, basándose únicamente en la frecuencia de aparición de sus atributos en un conjunto de entrenamiento predefinido (ver Sección 2.2.7); por ello, los clasificadores tradicionales impedirían aprovechar la inferencia semántica en el proceso de personalización.

El proceso de selección de usuarios con preferencias similares a las de cada espectador candidato se desarrolla en el módulo denominado *creación del vecindario*, representado en la Fig. 8.1. Tal como se aprecia en la misma, este módulo debe acceder tanto a los perfiles de los usuarios como a la ontología OWL de TV. Esto último es debido a que nuestro enfoque adopta el principio "*collaboration via content*" propuesto por Pazzani [189] (ver Sección 2.6.4), de ahí que, a la hora de comparar las preferencias de dos usuarios, utilice tanto los niveles de interés definidos en sus perfiles como las descripciones semánticas de sus preferencias, formalizadas en la ontología.

En concreto, nuestro enfoque se basa en la aproximación propuesta por Cai Ziegler *et al.* en [259] y construye un *vector de interés* para cada usuario candidato. Tal como se detalló en la Sección 3.10.3, su sistema recomendador de libros adopta una jerarquía en la que se representan diferentes categorías de géneros literarios. Son precisamente los niveles de interés del usuario en relación a cada una de dichas categorías los que se incluyen como componentes de los vectores utilizados en su trabajo.

La aplicación inmediata de este enfoque en nuestra metodología tendría asociado un coste computacional muy elevado, ya que cada vector de interés incluiría los índices DOI asociados en el perfil del usuario a todas las clases representadas en la jerarquía de géneros de la Fig. 5.1. Con el claro propósito de combatir este coste, nuestra metodología reduce la dimensión de los vectores de interés considerando únicamente aquellas clases de la jerarquía de géneros que son más significativas para el proceso de personalización. El criterio para elegir estas clases coincide con el descrito en la Sección 6.4.2 (empleado allí para seleccionar las preferencias menos recientes del usuario que deben ser eliminadas de su perfil de forma permanente). En concreto, nuestro enfoque considera que las preferencias que aportan más conocimiento al recomendador son aquellas

que tienen asociados niveles de interés muy positivos o muy negativos, ya que éstas identifican tanto los géneros que resultan más atractivos para el usuario, como aquellos que no le interesan en absoluto. Para delimitar tales preferencias, la metodología definida en esta tesis emplea un umbral ψ. En concreto, nuestro enfoque exige que los niveles de interés de los géneros definidos en un perfil P_j excedan dicho umbral para así poder ser incluidos en su vector de interés v_j, tal como muestra la siguiente expresión:

$$\forall\, r \quad |v_j[r]| \geq \psi \tag{8.6}$$

Al igual que en el caso del umbral ϵ considerado en la expresión 6.5, nuestra metodología emplea el mismo valor de ψ para todos los usuarios del sistema, pero, a diferencia del primero, ψ permanece invariable a lo largo del tiempo[4]. De acuerdo a lo explicado en la Sección 6.4.2, el valor de ϵ puede variar en función del espacio de almacenamiento disponible en el servidor del sistema, de forma que cuando son pocos los usuarios registrados en el mismo, es posible decrementar el umbral y almacenar más información sobre sus respectivas preferencias. Por el contrario, decrementando ψ, nuestra metodología no sólo requeriría mayores capacidades de almacenamiento, sino que además incrementaría el coste computacional del enfoque, puesto que en este escenario los vectores de interés contendrían un número mayor de componentes. Esta es la razón por la cual el enfoque propuesto no contempla el ajuste dinámico del parámetro ψ.

Una vez construidos los vectores de interés de todos los usuarios candidatos, la estrategia aplica la expresión de la correlación Pearson-r para así detectar similitud entre sus intereses personales. Así, por ejemplo, a la hora de crear el vecindario del usuario U_j, nuestra metodología construye su vector de interés v_j seleccionando aquellos géneros que cumplen la condición impuesta en la expresión 8.6; a continuación, incluye estas mismas clases en los vectores del resto de espectadores del sistema y, finalmente, calcula la correlación entre sus respectivos niveles de interés. Esta expresión está formalizada en la Ec. 8.7, en la que se comparan las preferencias incluidas en los perfiles P_j y P_k utilizando sus respectivos vectores de interés v_j y v_k (cuyos valores medios son $\overline{v_j}$ y $\overline{v_k}$, respectivamente).

$$corr(P_j, P_k) = \frac{\sum_r (v_j[r] - \overline{v_j})(v_k[r] - \overline{v_k})}{\sqrt{\sum_r (v_j[r] - \overline{v_j})^2 \cdot \sum_r (v_k[r] - \overline{v_k})^2}} \tag{8.7}$$

La Ec. 8.7 permite cuantificar tanto niveles positivos como negativos de correlación entre las preferencias de los usuarios U_j y U_k. De hecho, los valores oscilan entre -1 y 1, obteniendo niveles positivos si U_k ha clasificado con niveles de interés similares a los de U_j contenidos pertenecientes a los mismos géneros definidos en el vector v_j, y negativos cuando las clasificaciones de ambos usuarios son muy diferentes.

Finalmente, destacar que el método de formación de vecindario empleado en nuestra estrategia permite reducir los efectos adversos del *sparsity problem*, tradicionalmente vinculado a los sistemas colaborativos. El objetivo es generar solapamientos entre los perfiles de los usuarios comparados aun cuando éstos no contengan exactamente los mismos contenidos. Esta es la razón por la que los vectores de interés incluyen los

[4]Por su naturaleza, este parámetro debe tomar un valor próximo a 1, debido a que ésta es la cota superior de los niveles de interés DOI utilizados en la metodología. En nuestro dominio de aplicación concreto, hemos utilizado valores de ψ entre $0{,}75$ y $0{,}85$.

índices DOI de los géneros a los que pertenecen los programas registrados en el perfil del usuario, en lugar de incluir las clasificaciones de estos últimos. De esta forma, para detectar que las preferencias de dos usuarios son similares, el enfoque no exige que los programas que contienen sus perfiles sean idénticos, sino únicamente que pertenezcan a las mismas categorías jerárquicas (ver Fig. 5.1).

8.3.2 Selección del tamaño del vecindario

Una vez calculada la correlación entre un usuario candidato y el resto de espectadores del sistema, es necesario delimitar el tamaño de vecindario considerado. Tal como se detalló en la Sección 2.6.3, la técnica más utilizada para este propósito se basa en seleccionar los M vecinos a los que corresponden los valores de correlación más elevados. De forma análoga a la comentada antes, en presencia del *sparsity problem*, esta técnica conduce a recomendaciones más precisas que las obtenidas mediante otros métodos, como el centroide o el basado en definir un umbral para limitar el nivel de correlación mínimo que debe existir entre un usuario y sus vecinos [93]. Por esta razón, nuestro enfoque incluye en el vecindario de cada usuario candidato U_c, los M espectadores[5] cuyas preferencias son más similares a las registradas en su perfil personal, empleando en esta selección los valores de correlación obtenidos mediante la Ec. 8.7.

8.3.3 Predicción basada en el vecindario del usuario

Una vez identificado el vecindario del usuario candidato U_c, la estrategia debe predecir el nivel de interés concreto de éste en relación al contenido objetivo a y, en función del mismo, decidir si debe o no recomendar este programa.

Dicho proceso de predicción incorpora una novedad en relación a los métodos colaborativos (basados en usuario) convencionales descritos en la Sección 2.6.3. A la hora de predecir el interés de un usuario en un programa concreto, los enfoques existentes promedian las clasificaciones que sus vecinos han asignado a dicho contenido, ponderando cada una de ellas mediante los valores de correlación existente entre sus respectivas preferencias (perfiles). Como consecuencia, este modelo sugiere a cada usuario aquellos programas que han despertado mayor interés entre los vecinos con preferencias más parecidas a las suyas.

De lo explicado se desprende que estos enfoques predicen el interés de un usuario considerando únicamente las contribuciones de los vecinos que ya conocen el contenido objetivo. El resto de espectadores que conforman el vecindario son obviados, ya que el enfoque no conoce sus niveles de interés concretos en relación al programa objetivo y, por tanto, no es posible incluir dichas clasificaciones en el proceso de predicción.

A diferencia de este tipo de aproximaciones, nuestra metodología considera el vecindario completo de cada usuario candidato U_c. Así, si un vecino ya conoce el contenido a, la predicción se basa en el índice DOI que figura en su perfil; en caso contrario, nuestra estrategia predice este nivel de interés a partir de la similitud semántica existente entre a y las preferencias actuales de dicho vecino (es decir, los contenidos y

[5] Si bien el valor del parámetro M depende de cada dominio de aplicación, en nuestra evaluación experimental hemos adoptado tamaños de vecindario en torno al 10-15 % del total de usuarios modelados en el sistema. Considerando valores más elevados, el enfoque incluía en el vecindario del usuario activo espectadores con preferencias poco similares a las suyas, debilitando con ello la calidad de las recomendaciones ofrecidas en la fase colaborativa.

atributos incluidos en la ventana temporal actual considerada para este usuario). En concreto, este índice de interés coincide con el nivel de *matching* semántico calculado durante la fase basada en contenido entre a y el perfil de este vecino. Esta misma idea aparece reflejada en la Fig. 8.1, donde se representa el módulo encargado de llevar a cabo la predicción semántica, utilizando tanto el perfil del usuario U_c y su vecindario, como los niveles de *matching* computados por el bloque correspondiente.

La expresión que permite cuantificar el nivel de interés de U_c con respecto a a se muestra en la Ec. 8.8, donde $\mathcal{N}_k$ es el k-ésimo vecino del usuario candidato U_c, $corr(U_c, \mathcal{N}_k)$ es la correlación entre ambos, M es el tamaño del vecindario y $\delta(\mathcal{N}_k)$ es un factor cuyo valor depende de si este vecino conoce o no el programa a. De acuerdo a lo explicado previamente, en el primer supuesto, su valor coincide con el nivel de interés de este usuario en relación a a; en caso contrario, se iguala al nivel de *matching* entre dicho usuario y el programa a.

$$Pred(a, U_c) = \frac{1}{M} \sum_{k=1}^{M} \delta(\mathcal{N}_k) \cdot corr(U_c, \mathcal{N}_k)$$

$$\delta(\mathcal{N}_k) = \begin{cases} DOI_{P_{\mathcal{N}_k}}(a) & \text{si } \; a \text{ está almacenado en el perfil de } \mathcal{N}_k \\ matching(a, \mathcal{N}_k) & \text{en otro caso} \end{cases} \tag{8.8}$$

Según la Ec. 8.8, a será un contenido más interesante para U_c cuanto mayor sea el nivel de interés de sus vecinos en relación a dicho contenido, y cuanto más se parezcan sus preferencias personales a las de estos últimos. Esta es precisamente la premisa en la que se fundamenta el filtrado colaborativo: si a ha interesado a muchos usuarios con preferencias muy parecidas a las de U_c, este programa también será relevante para él.

Al igual que en la fase basada en contenido, una vez cuantificados los niveles de predicción semántica $Pred(a, U_c)$, es necesario definir algún criterio para decidir si estos valores son suficientemente significativos para recomendar el contenido objetivo a dicho candidato.

8.3.4 Criterio de recomendación: rango del umbral $\beta_{Pred_{U_c}}$

El mecanismo empleado en la fase colaborativa a la hora de decidir si un programa debe o no ser sugerido al usuario es completamente análogo al descrito en la Sección 8.2.3 para la etapa basada en contenido. De esta forma, nuestra estrategia sugiere el contenido a a cada usuario candidato U_c cuyo nivel de predicción superen un umbral $\beta_{Pred_{U_c}}$ previamente establecido.

Tal como muestra la Ec. 8.9, los valores de dicho umbral —pertenecientes al rango $[\omega_3, \omega_2]$— también se ajustan dinámicamente para cada uno de los usuarios candidatos U_c, en función de la precisión de las recomendaciones que le ofreció el sistema en el pasado.

$$\beta_{Pred_{U_c}}(t{+}1) = \begin{cases} \omega_3 & \text{si } \; \beta_{Pred_{U_c}}(t) - \Delta_P(t+1, t) < \omega_3 \\ \omega_2 & \text{si } \; \beta_{Pred_{U_c}}(t) - \Delta_P(t+1, t) > \omega_2 \\ \beta_{Pred_{U_c}}(t) - \Delta_P(t+1, t) & \text{en otro caso} \end{cases} \tag{8.9}$$

donde $\Delta_P(t+1, t) = P(t+1) - P(t)$.

Una vez revisadas las dos fases incluidas en la estrategia de personalización propuesta, reunimos en Alg. 3 la contribución de cada una de ellas al algoritmo[6] de recomendación ejecutado por el agente híbrido (ver Fig. 4.4). Tal como muestra el pseudocódigo, nuestro enfoque sólo recurre a la fase colaborativa si el nivel de *matching* semántico no detecta similitud entre el contenido a y las preferencias negativas del usuario activo U.

Alg. 3 Algoritmo de recomendación utilizado en la estrategia híbrida

```
************************************************
    * Entrada:    contenido objetivo a y usuario activo U
    * Salida:     resultado de la recomendación
************************************************
Recomendar (a,U) {
 Si (matching(a,U) ≥ β_Match_U) Entonces devolver SI
 Si no, si ((matching(a,U) ≥ 0) Y (Pred(a,U) ≥ β_Pred_Uc)) Entonces devolver SI
 En otro caso devolver NO
}
```

8.4 Cálculo de la relevancia de los contenidos sugeridos

Una vez ejecutado el algoritmo antes descrito, el agente híbrido debe ordenar los contenidos objetivo sugeridos a cada usuario activo de acuerdo a su relevancia para el mismo. Esta tarea es la desempeñada por el *módulo de ordenación* representado en la Fig. 8.1.

El mecanismo de clasificación se basa únicamente en el tipo de estrategia de recomendación utilizada para descubrir cada contenido. Así, si un programa es sugerido mediante la fase basada en contenido, será más relevante cuando mayor sea su nivel de *matching* con respecto a las preferencias del usuario activo (Ec. 8.1). Por el contrario, si es la etapa colaborativa la que recomienda dicho programa al espectador, éste será más significativo cuanto mayor sea el nivel de predicción semántica calculado para el mismo (Ec. 8.8).

Por último, el módulo funde en una única recomendación todos aquellos contenidos sugeridos al usuario activo, y los dispone en una lista final en orden decreciente según su nivel de *matching* o de predicción semántica (según corresponda). De esta forma, los contenidos que ocupan las primeras posiciones son aquellos que, o bien son muy similares a las preferencias que más interesan al usuario activo, o por el contrario, son muy interesantes para otros espectadores que comparten gran parte de sus intereses televisivos.

8.5 Técnicas de hibridación empleadas

Tal como describimos en la Sección 2.6.4, Burke acuñó un conjunto de técnicas que permiten combinar estrategias de personalización de diferente naturaleza en un sistema

[6]En el pseudocódigo mostrado en Alg. 3, las funciones *matching* y *Pred* computan los niveles de *matching* y predicción semántica, respectivamente.

híbrido [42]. En esta sección, identificamos cuáles de estos métodos, cuyas principales características recordamos en la Tabla 8.1, se aplican en la estrategia descrita en este capítulo.

Métodos	Descripción	Estrategia híbrida
Ponderado	Se combinan las decisiones (ponderadas) de varias estrategias de personalización para elaborar una única recomendación.	—
Conmutación	El sistema conmuta entre las diferentes estrategias cuando se cumple alguna condición.	√
Mixto	Las recomendaciones elaboradas por diferentes estrategias se funden en una única lista.	√
Combinación de características	Se funden en único conjunto los datos que utilizan las diferentes estrategias de personalización, y con éste se ejecuta un solo algoritmo de recomendación.	—
Cascada	La recomendación elaborada por una de las estrategias es refinada por las restantes.	—
Incorporación de características	La salida de una de las estrategias se usa como parámetro de entrada en las siguientes.	√
Metanivel	El modelo completo aprendido por una de las estrategias es utilizado por las restantes.	—

Tabla 8.1: Métodos de combinación de estrategias de recomendación empleados por el agente híbrido (ver Fig. 8.1)

- El método de **conmutación** está presente en nuestro enfoque ya que, tal como se aprecia en el pseudocódigo mostrado en Alg. 3, el sistema empieza ejecutando la fase basada en contenido y cuando se verifica una determinada condición —en nuestro caso, que los niveles de *matching* semántico computados no superan el umbral β_{Match_U} establecido— conmuta a la etapa colaborativa.

- Por otra parte, tal como muestra la Ec. 8.8, cuando los vecinos de un usuario candidato no conocen el contenido objetivo, la fase colaborativa emplea los niveles de *matching* semántico calculados en la etapa basada en contenido para predecir sus niveles de interés en relación al mismo. Por esta razón, es posible identificar en esta fase el método de **incorporación de características** descrito en la Tabla 8.1.

- Por último, es fácil advertir la presencia del método **mixto** en nuestro enfoque híbrido sin más que considerar el proceso de clasificación descrito en la Sección 8.4. Esto último es debido a que los programas sugeridos mediante la fase basada en contenido y aquellos descubiertos en la etapa colaborativa pueden ser mostrados en una misma recomendación.

Antes de concluir este capítulo destacando los beneficios de la estrategia híbrida desarrollada en este trabajo, presentamos un ejemplo en el que se incluyen las dos fases del enfoque.

8.6 Ejemplo de aplicación

Con el objetivo de ilustrar la técnica de personalización con suficiente claridad, en este escenario asumimos las mismas simplificaciones utilizadas en el ejemplo de aplicación de la estrategia basada en contenido, descrito en la Sección 7.8. En concreto, retomamos como usuario activo el espectador U (cuyo perfil recordamos en la Fig. 8.6) y utilizamos como contenido objetivo el documental *España durante la Segunda Guerra Mundial*, representado en el extracto de la ontología OWL mostrado en la Fig. 8.3.

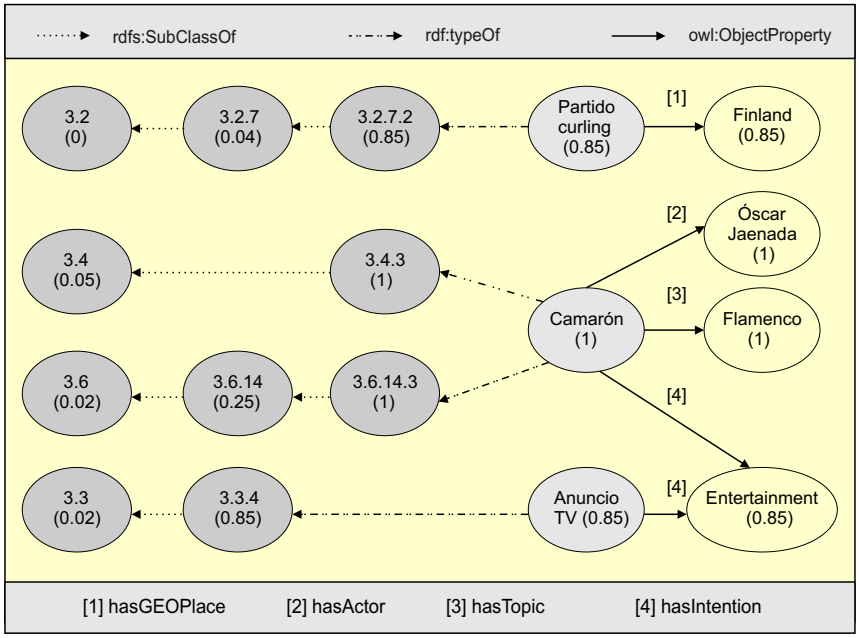

Figura 8.6: Perfil P_U

8.6.1 Fase basada en contenido

De acuerdo a lo explicado en la Sección 8.2, el agente híbrido debe empezar calculando el nivel de *matching* semántico entre el contenido objetivo y las preferencias

de U. Para ello, es necesario incorporar en la Ec. 8.1 los valores de similitud semántica (jerárquica e inferencial) entre dicho programa y aquellos registrados en el perfil P_U.

En este escenario, el enfoque basado en contenido infiere que el documental *España durante la Segunda Guerra Mundial* no es relevante para U, pues no es similar a ninguno de los programas que han interesado a este espectador en el pasado:

- Por una parte, la similitud jerárquica entre el documental y las preferencias de U se anula para todos los contenidos definidos en su perfil, pues el único LCA existente entre ellos es la clase raíz de la jerarquía de géneros incluida en la ontología OWL (el documental pertenece al género 3.1, mientras que el perfil P_U registra las clases 3.2, 3.3, 3.4 y 3.6).
- Por otro lado, el agente híbrido no infiere ninguna asociación implícita entre el documental *España durante la Segunda Guerra Mundial* y los contenidos definidos en el perfil de U. Tal como muestran las Fig. 8.6 y Fig. 8.3, estos programas no comparten ninguna instancia ni clase de unión, de ahí que la Ec. 8.4 conduzca a un valor de similitud inferencial nulo.

Dado que el enfoque detecta un nivel de *matching* nulo entre las preferencias de U y el contenido objetivo, el agente híbrido reconsidera su perfil durante la fase colaborativa.

8.6.2 Fase colaborativa

De acuerdo a lo explicado en la Sección 8.3, la etapa colaborativa debe: (i) seleccionar los vecinos de U, y (ii) predecir si el programa *España durante la Segunda Guerra Mundial* es relevante para este espectador basándose tanto en la correlación existente entre sus respectivas preferencias, como en el interés que despierta el contenido objetivo entre dichos vecinos.

Formación del vecindario

Siguiendo las directrices descritas en la Sección 8.3.1, el enfoque propuesto selecciona los vecinos de U recurriendo a los géneros que figuran en sus perfiles, en lugar de considerar los contenidos específicos definidos en los mismos. Así, uno de los usuarios que podría estar incluido en el vecindario de U es el espectador V_1, cuyo perfil se muestra en la Fig. 8.7.

Aunque V_1 no ha visto ninguno de los programas que interesan al usuario activo, nuestro enfoque infiere que sus preferencias son similares a las de este espectador porque en su perfil figuran géneros que también aparecen en P_U (y con índices DOI parecidos), como contenidos de entretenimiento sobre viajes (3.3.4), programas musicales sobre bailes (3.6.14) y contenidos sobre deportes de invierno (3.2.7). La efectividad de este mecanismo como medida para combatir el *sparsity problem* es evidente, pues no requiere un solapamiento directo entre los programas definidos en los perfiles de los usuarios, para poder detectar similitud entre sus preferencias.

Predicción semántica

A diferencia de las propuestas existentes, nuestro enfoque colaborativo incorpora en el proceso de predicción semántica la contribución de todos los vecinos de U. En

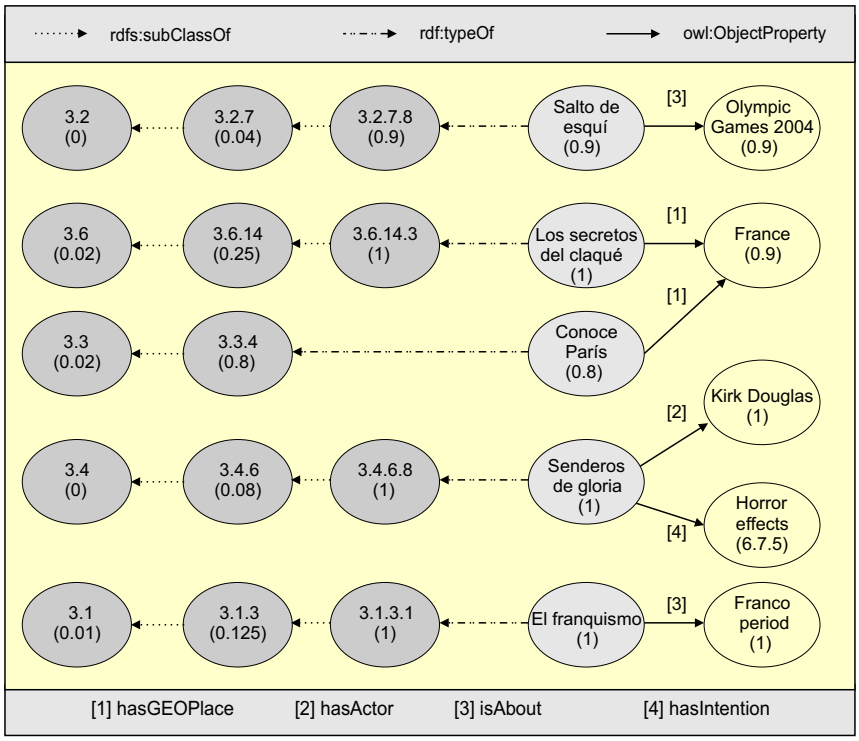

Figura 8.7: Perfil del usuario V_1, vecino de U

este caso, el enfoque infiere un índice DOI significativo para aquellos vecinos que están interesados en programas *similares* al documental objetivo. Ilustremos este razonamiento utilizando el perfil del vecino V_1 mostrado en la Fig. 8.7.

En este escenario, el enfoque predice un nivel de interés elevado para *España en la Segunda Guerra Mundial*, dado que este programa se parece a las preferencias positivas de V_1. Para detectar tal similitud, es necesario incorporar en la Ec. 8.2 las relaciones jerárquicas y asociaciones semánticas implícitas inferidas entre estos contenidos. Tal como muestra la Fig. 8.3, *España durante la Segunda Guerra Mundial* está relacionado con dos de los programas registrados en el perfil P_{V_1}: el documental *El franquismo* y la película *Senderos de Gloria*.

- Por una parte, la instancia de unión *Franco period* (identificada mediante razonamiento temporal) permite inferir que *España durante la Segunda Guerra Mundial* es un contenido relevante para V_1, pues describe un hecho histórico incluido en el régimen político que interesa al usuario[7]. Además, tal como muestra la Fig. 8.3, el programa *El franquismo* y el documental objetivo comparten un LCA en la jerarquía de géneros de nuestra ontología (el género 3.1), de ahí que

[7]El objetivo de este ejemplo es ilustrar algunos razonamientos sencillos utilizados en AVATAR, de ahí que consideremos que el documental objetivo es similar al programa *El franquismo* basándonos únicamente en una de sus características semánticas. En un escenario real, el sistema utiliza el conocimiento completo de la ontología OWL, siendo necesario que dos contenidos compartan un gran número de características para así establecer que son similares.

la Ec. 8.3 conduzca a una similitud jerárquica entre ambos programas de valor $\frac{2}{4} = 0{,}5$.

- Por otro lado, mediante la clase de unión *disturbing scenes (6.7)* vinculada a la película *Senderos de gloria* que ha visto V_1, el sistema descubre que, a pesar de contener imágenes que puedan dañar la sensibilidad del espectador, el documental bélico puede ser sugerido a este usuario.

Las relaciones jerárquicas y las asociaciones semánticas que infiere el agente híbrido contribuyen a incrementar el valor de similitud entre el documental objetivo y las preferencias del vecino V_1 y, con ello, el nivel de *matching* semántico incluido en la Ec. 8.1. Mediante dicho nivel, el enfoque predice un índice DOI_{V_1}(*España durante la Segunda Guerra Mundial*) muy significativo. Si este mismo patrón de comportamiento se repite en muchos de los vecinos de U que no conocen el documental bélico, la Ec. 8.8 conducirá a un nivel de predicción semántica $Pred$(*España durante la Segunda Guerra Mundial*, U) elevado, capaz de superar el umbral β_{Pred_U}; así, este programa pueda ser finalmente sugerido a U.

8.6.3 El razonamiento semántico vs. los enfoques convencionales

El ejemplo presentado pone de manifiesto las ventajas de nuestro enfoque híbrido frente a las propuestas actuales, carentes de la capacidad de razonamiento semántico incorporada en AVATAR.

- Por una parte, nuestra metodología permite una mayor flexibilidad a la hora de establecer la similitud entre contenidos de TV concretos, combatiendo las limitaciones derivadas de los enfoques (sintácticos) tradicionalmente adoptados en los métodos basados en contenido.

- Por otro lado, el razonamiento semántico también mejora las técnicas colaborativas actuales, pues permite descubrir conocimiento útil sobre las preferencias de los vecinos de un usuario dado, y aplicarlo durante la predicción de su interés en el contenido objetivo.

8.7 Beneficios de la estrategia híbrida

Para concluir este capítulo, recapitulamos las ventajas más destacables de nuestra estrategia híbrida describiendo los mecanismos con los que combate las limitaciones reconocidas en los enfoques actuales.

De forma análoga a la comentada en el capítulo previo, el razonamiento incorporado en nuestra métrica de similitud semántica contribuye a diversificar las sugerencias basadas en contenido —tradicionalmente ***sobreespecializadas***— sin comprometer su carácter personalizado. Para ello, esta métrica (y más concretamente su componente inferencial), descubre asociaciones semánticas de diversa naturaleza (ver Sección 8.2.2) entre los contenidos comparados, permitiendo detectar parecido entre programas que no comparten, necesariamente, los *mismos* atributos semánticos.

Esta métrica de similitud también permite seleccionar recomendaciones personalizadas para aquellos usuarios cuyas preferencias son muy diferentes a las del resto

de espectadores del sistema (el denominado ***gray sheep problem***). En este escenario, los enfoques colaborativos tradicionales no son capaces de crear el vecindario de tales usuarios, hecho que imposibilita la elaboración de sugerencias de calidad. Por el contrario, la similitud semántica permite —gracias al proceso de razonamiento— establecer vínculos interesantes entre las preferencias de estos usuarios y los contenidos de la ontología del sistema que están asociados con ellos, seleccionando, de esta forma, recomendaciones potencialmente interesantes.

La estrategia propuesta también reduce los efectos adversos del ***sparsity problem*** asociado a los enfoques colaborativos tradicionales. Esta limitación es debida a que este tipo de propuestas únicamente detectan que dos usuarios tienen preferencias similares cuando sus respectivos perfiles registran los mismos contenidos. De esta forma, cuando dicha premisa no se cumple, no es posible formar el vecindario de cada usuario y, en consecuencia, el sistema no puede ofrecer recomendaciones colaborativas. Tal como describimos en la Sección 2.6.3, en la literatura se han propuesto enfoques que tratan de combatir tal limitación gracias al descubrimiento de reglas de asociación entre los programas comparados durante el proceso de formación del vecindario. Los valores de confianza asociados a dichas reglas se emplean como métrica para detectar similitud entre contenidos diferentes, mitigando con ello los efectos adversos del problema comentado [184]. La principal limitación de este tipo de enfoques es que el descubrimiento de una regla entre dos programas dados exige que éstos estén almacenados a la vez en los perfiles de entrenamiento utilizados durante el proceso de detección de las mismas, de ahí que los efectos del *sparsity* sigan limitando la calidad de sus recomendaciones. Con el propósito de combatir este problema, nuestra aproximación utiliza la estructura jerárquica en la que se organizan las diferentes categorías de contenidos de TV en la ontología OWL. De esta forma, la Ec. 8.7 conduce a valores de correlación no nulos entre usuarios cuyos perfiles no deben contener necesariamente los mismos contenidos, sino que únicamente se exige que éstos pertenezcan a los mismos géneros jerárquicos.

Otra de las limitaciones vinculadas a los enfoques colaborativos son los **tiempos de latencia** que deben transcurrir desde que un nuevo contenido llega al sistema hasta que es sugerido a los usuarios, debido a que es necesario que éste sea clasificado por un número suficientemente elevado de espectadores antes de ser incluido en la recomendación (ver Sección 2.6.3). Este problema es especialmente crítico en el dominio de la DTV, ya que impide que los espectadores puedan recibir, sin retardos innecesarios, un gran número de contenidos nuevos que se distribuyen continuamente (y que, además, en algunos casos, sólo están disponibles durante un tiempo limitado). Nuestro enfoque elimina tales latencias, dado que la fase colaborativa de la estrategia puede recomendar a un usuario candidato contenidos que no hayan sido clasificados previamente por ninguno de sus vecinos. Para ello, nuestro enfoque predice los niveles de interés de estos vecinos a partir de la similitud semántica entre los programas considerados y sus preferencias.

En la Sección 2.6.4 ya adelantamos que los efectos sinergéticos de los métodos basados en contenido y el filtrado colaborativo también han sido aprovechados en los sistemas híbridos existentes, para aliviar tanto los tiempos de latencia como el *sparsity problem* antes mencionado. En concreto, si un sistema híbrido no puede sugerir un determinado contenido a un usuario mediante la técnica colaborativa —debido a los efectos del *sparsity problem*— puede recurrir al filtrado basado en contenido, ya que éste no está limitado por tal problema. De forma análoga, el sistema también puede utilizar este tipo de filtrado a la hora de recomendar un programa que acaba de lle-

gar al sistema, eliminando con ello los tiempos de latencia asociados a los enfoques colaborativos. De acuerdo al principio de funcionamiento de las técnicas basadas en contenido, en ambos casos, el programa en cuestión sólo será sugerido al usuario si es sucientemente similar a sus preferencias personales. Basta recordar las limitaciones de las métricas de similitud descritas en la literatura, y los problemas de *sobreespecialización* que de ellas se derivan, para advertir que dicho contenido sólo será recomendado si es muy parecido a las preferencias del usuario activo; en caso contrario, los efectos del *sparsity problem* y de los tiempos de latencia, seguirán comprometiendo la calidad de las recomendaciones ofrecidas por los sistemas híbridos actuales. Tal como adelantamos antes, nuestra métrica de similitud semántica (jerárquica e inferencial) combate este tipo de debilidades, para las que aún no se habían propuesto soluciones efectivas en la literatura.

Los problemas de **escalabilidad** también suponen una limitación muy importante en los enfoques colaborativos, ya que, a medida que aumenta el número de usuarios en el sistema y la cantidad de preferencias registradas en sus perfiles, se incrementa también el coste computacional del proceso de formación del vecindario de cada uno de ellos. Nuestra aproximación permite controlar tal complejidad gracias a que los vectores de interés únicamente contienen aquellas categorías de la jerarquía de géneros cuyos índices DOI son realmente significativos para el proceso de personalización (es decir, mayores que el umbral de interés para las preferencias positivas, y menores para las negativas). Por otra parte, conviene destacar que un incremento en el número de programas registrados en el perfil de un usuario no se traduce necesariamente, en nuestro modelo, en un aumento en la dimensión de los vectores de interés, ya que muchos de estos programas pueden pertenecer a los géneros ya definidos en dichos vectores. Los beneficios de esta técnica en términos de escalabilidad son evidentes.

Por último, nuestro enfoque también permite combatir de manera efectiva los efectos del denominado ***new user ramp-up***, causante de la escasa precisión de las recomendaciones ofrecidas por un recomendador a un usuario que se acaba de registrar en el sistema, cuyo perfil contendrá, típicamente, un conjunto muy reducido de preferencias. A diferencia de los sistemas que prescinden de representaciones del conocimiento de su dominio de aplicación concreto, nuestro enfoque puede emplear la ontología de TV para completar los perfiles inmaduros de tales usuarios. Así, aun cuando dichos perfiles contengan un único programa, la estrategia híbrida propuesta puede consultar la base de conocimiento OWL y descubrir contenidos relacionados con éstos, ya sea porque pertenecen a las mismas categorías de la jerarquía de programas o porque comparten características semánticas de interés para dicho usuario. Este enfoque es similar al propuesto por Middleton en el sistema QuickStep [149] descrito en la Sección 3.10.3, ya que este trabajo también completa las preferencias del usuario a partir del conocimiento disponible en una taxonomía. Sin embargo, durante este proceso, Middleton considera únicamente los enlaces jerárquicos establecidos en la ontología de dominio del sistema QuickStep, y por tanto, prescinde del conocimiento oculto que aportarían las asociaciones semánticas complejas consideradas en nuestra metodología de razonamiento.

Parte V

Conclusiones

Capítulo 9

Conclusiones y líneas futuras

Este capítulo resume las principales conclusiones extraídas del trabajo presentado en la tesis, haciendo especial hincapié en los beneficios que aporta el razonamiento semántico —tradicionalmente vinculado a la comunidad de la Web Semántica— a las estrategias de personalización adoptadas en los sistemas recomendadores actuales. Seguidamente, se presentan algunas líneas de investigación futura que extenderían, de manera natural, el trabajo desarrollado.

9.1 Conclusiones

El trabajo desarrollado en esta tesis arrancó con el objetivo de enriquecer las capacidades de personalización ofrecidas por los sistemas recomendadores actuales, recurriendo a técnicas de inferencia de conocimiento tradicionalmente vinculadas a la Web Semántica. La incorporación de dichas técnicas permite: (i) mejorar las sugerencias que selecciona el sistema recomendador, posibilitando el descubrimiento de productos que pasarían desapercibidos en los enfoques existentes, y (ii) combatir de forma efectiva las debilidades que limitan las estrategias de personalización utilizadas en los mismos, derivadas, principalmente, de su naturaleza sintáctica. Para lograr ambos objetivos, hemos desarrollado una metodología de razonamiento que infiere relaciones entre los productos considerados en el sistema basándose en sus descripciones semánticas, previamente formalizadas en una ontología de conocimiento. Gracias a este proceso de razonamiento semántico, los recomendadores disponen de conocimiento adicional para *relacionar* las preferencias de los usuarios con los productos disponibles en el sistema, favoreciendo con ello procesos de personalización más precisos.

Con el propósito de explotar el potencial de la semántica representada en la ontología, y con ello enriquecer las sugerencias ofrecidas a los usuarios, nuestro trabajo descubre relaciones ocultas en la base de conocimiento utilizando tanto los enlaces jerárquicos establecidos entre sus clases, como las propiedades explícitas definidas entre las instancias. Es precisamente la naturaleza implícita y compleja de estas relaciones la que nos obliga a desarrollar una metodología de inferencia propia, prescindiendo de los razonadores lógicos existentes en la Web Semántica y de los lenguajes de consulta de ontologías propuestos en la literatura:

- Los primeros aprovechan sus capacidades lógicas para comprobar la consistencia de una ontología y para clasificar de forma automática sus instancias bajo las

clases oportunas, omitiendo, por tanto, la inferencia de relaciones (complejas) como las que nos interesan.

- Por otra parte, los lenguajes existentes manejan únicamente relaciones explícitas en la ontología (propiedades), y además obligan al usuario a conocer en detalle su estructura para así poder formular correctamente sus consultas. Esta exigencia es abusiva en un entorno de personalización como el asumido en este trabajo, donde el usuario no tiene por qué conocer la base de conocimiento utilizada en el sistema.

Si bien la metodología desarrollada es suficientemente flexible y genérica para ser reutilizada en diferentes dominios con requisitos de personalización, en este trabajo hemos adoptado un contexto específico para poder ilustrar el enfoque con claridad.

En concreto, la metodología desarrollada se ha aplicado en un sistema recomendador para Televisión Digital, un medio en el que, gracias al incremento en la capacidad de transmisión habilitado por el proceso de digitalización, se difunde una cantidad ingente de contenidos audiovisuales y aplicaciones software interactivas. Tal sobrecarga de información justifica la utilidad de una herramienta de personalización en este dominio, cuya tarea es seleccionar automáticamente aquellos contenidos que puedan ser interesantes para el usuario, teniendo en cuenta las preferencias registradas en su perfil personal. Éste es precisamente el objetivo del sistema AVATAR descrito en nuestro trabajo, un recomendador de programas de televisión que combina las estrategias de personalización adoptadas en los enfoques actuales, con las capacidades de razonamiento semántico incorporadas en la metodología desarrollada. Tales capacidades permiten descubrir recomendaciones atractivas para los espectadores, empleando las relaciones semánticas (implícitas) establecidas entre los programas definidos en sus perfiles y los disponibles en la base de conocimiento del sistema.

Una vez delimitado el contexto de aplicación de la metodología de razonamiento desarrollada, revisamos a continuación los tres pilares identificados en la misma: (i) la ontología sobre el dominio de la DTV, (ii) los perfiles de los usuarios, en los que se modelan sus preferencias, y (iii) las estrategias de personalización utilizadas para seleccionar los contenidos de TV finalmente sugeridos.

9.1.1 Ontología de TV

Son dos los elementos requeridos para formalizar el dominio de la DTV mediante una ontología: (i) las descripciones semánticas de los contenidos de TV disponibles en el sistema recomendador, y (ii) un lenguaje con suficiente capacidad expresiva para representar los conceptos (clases e instancias) y relaciones (enlaces jerárquicos y propiedades) identificados en las primeras.

En nuestro enfoque, tales descripciones han sido extraídas de la especificación TV-Anytime. Esta norma, ampliamente aceptada en el dominio de la DTV, simplifica considerablemente el proceso de formalización semántica, ya que proporciona un conjunto de metadatos (aplicaciones XML) para describir detalladamente contenidos audiovisuales genéricos (identificados de forma unívoca mediante su CRID - *Content Reference IDentifier*).

A partir de estos metadatos, nuestro enfoque define de forma inmediata diferentes jerarquías de clases y propiedades en la ontología, así como instancias específicas de las mismas. Los programas de TV representados en esta base de conocimiento son

instancias pertenecientes a una jerarquía de géneros, organizada en diferentes niveles (*fiction*, *music*, *sports*, etc.). Junto al CRID de cada programa, nuestra ontología representa también sus atributos semánticos, identificados mediante instancias pertenecientes a jerarquías de muy diversa naturaleza, en las que se organiza información relacionada con alertas de contenido, intención con la que han sido creados los programas, tipo de audiencia a la que van dirigidos, etc. Cada uno de estos atributos se asocia al programa en cuestión mediante propiedades convenientemente etiquetadas (*hasAlert*, *hasIntention*, *hasIntendedAudience*, etc.).

Nuestro enfoque complementa estas descripciones semánticas gracias al conocimiento formalizado en otras ontologías. En concreto, hemos importado dos conceptualizaciones para poder asociar información geográfica y temporal a los contenidos de TV. Sobre el conocimiento representado en ambas ontologías, nuestro trabajo aplica razonamientos lógicos automáticos que permiten relacionar contenidos vinculados a mismo emplazamiento geográfico o temporal. A pesar de su simplicidad, este tipo de relaciones son un complemento perfecto en procesos inferenciales que, al ser más elaborados, permiten mejorar las recomendaciones ofrecidas en los enfoques tradicionales.

La capacidad expresiva de OWL —que permite formalizar conceptos y relaciones no soportados en otros formatos, como RDF y RDFS— ha sido decisiva para elegir este lenguaje de implementación en nuestro ontología. Más concretamente, de los tres sublenguajes incluidos en OWL —Lite, DL y Full— nos hemos decantado por OWL DL. Esta elección responde a dos razones: en primer lugar, OWL Lite carece de ciertas funcionalidades útiles en el proceso de formalización semántica (e.g. restricciones de cardinalidad genéricas, disyunción de clases, tipos enumerados, etc.), lo que limitaría el conocimiento representado en la ontología; por otra parte, actualmente, no existen razonadores lógicos decidibles para el lenguaje OWL Full, lo que impediría aprovechar estas herramientas a la hora de comprobar la consistencia de nuestra ontología de TV.

9.1.2 Perfiles de usuario

Aprovechando la capacidad de reutilización de conocimiento asociada a las ontologías, nuestro enfoque modela las preferencias de cada espectador a partir de la conceptualización OWL descrita en la sección previa. Así, al disponer de una representación formal de los perfiles de los usuarios, la metodología desarrollada puede utilizar el razonamiento semántico para comparar, de forma más efectiva, las preferencias del espectador frente al resto de contenidos disponibles en la ontología de TV.

En concreto, nuestro trabajo propone un modelo semántico para cada usuario en el que se almacena información tanto sobre los programas que le interesaron o no en el pasado (preferencias positivas y negativas, respectivamente), como sobre sus características semánticas y los géneros a los que éstos pertenecen. Una de las ventajas de este modelo es que, a la vez que proporciona una descripción muy detallada de las preferencias del usuario, no requiere que las instancias, clases y propiedades que las identifican se definan en cada perfil, lo que se traduce en un ahorro significativo en cuanto a la capacidad de almacenamiento requerida en el sistema. En su lugar, nuestro enfoque emplea un único repositorio de conocimiento (la ontología de TV) y únicamente registra en los perfiles los CRIDs de los programas que identifican las preferencias positivas y negativas del usuario. Dado que los atributos de estos programas y los géneros a los que pertenecen están ya formalizados en la ontología, no es necesario incluir esta in-

formación en el perfil, sino simplemente consultarla en la base de conocimiento OWL (utilizando el CRID como identificador).

Dados los requisitos de personalización de nuestro sistema, la técnica de modelado propuesta debe identificar también los niveles de interés concretos asociados tanto a los programas como a sus descripciones semánticas. En nuestro enfoque, dichos niveles de interés (llamados índices DOI y pertenecientes al rango [-1,1]) pueden ser declarados explícitamente por el usuario, o bien inferidos por el sistema a partir de su comportamiento de visionado (programas aceptados y rechazados en una recomendación, proporción vista de un programa sugerido por el sistema, etc.). Una vez descubierto el índice DOI correspondiente a un programa concreto, nuestra metodología incorpora un mecanismo para: (i) inferir los niveles de interés de sus características semánticas, y (ii) propagar dichos niveles por toda la jerarquía de géneros a los que el contenido pertenece.

Junto a estos niveles de interés, nuestro enfoque incluye un segundo tipo de índices (los llamados *índices de olvido*) que posibilitan la adaptación de la técnica de modelado a medida que evolucionan las preferencias del usuario. Como su nombre indica, dichos índices permiten que el sistema identifique las preferencias actuales del usuario y olvide aquéllas que se han quedado obsoletas.

En definitiva, el modelo de usuario propuesto en este trabajo es adecuado para un sistema basado en razonamiento semántico como AVATAR, pues aporta conocimiento muy detallado sobre sus preferencias, utilizado, de forma efectiva, en las estrategias de personalización incluidas en la metodología.

9.1.3 Estrategias de recomendación

Tal como adelantamos en la Sección 9.1, la metodología desarrollada en esta tesis parte de las técnicas de personalización adoptadas en los enfoques existentes e incorpora en las mismas capacidades de razonamiento semántico. En concreto, este trabajo se centra en las tres estrategias más extendidas en el dominio de los sistemas recomendadores: los métodos basados en contenido —que sugieren productos similares a los que interesaron al usuario en el pasado—, el filtrado colaborativo —que le recomienda productos interesantes para otros usuarios con preferencias similares a las suyas—, y los enfoques híbridos, que combinan las dos técnicas anteriores.

En nuestro enfoque hemos explorado los beneficios y las limitaciones de estas técnicas, resultando dos estrategias de personalización diferentes: la primera de ellas está inspirada en los métodos basados en contenido, mientras que la segunda combina dichos métodos con el filtrado colaborativo.

Estrategia basada en contenido

En esta primera estrategia, el objetivo es combatir el problema de la *sobreespecialización* asociado a los métodos basados en contenido, que provoca que las sugerencias ofrecidas incluyan únicamente contenidos muy parecidos a los que el usuario ya conoce. Para superar esta limitación —origen del abandono de los métodos basados en contenido en favor del filtrado colaborativo y los enfoques híbridos—, nuestro trabajo propone sugerir programas *relacionados semánticamente* con aquéllos que interesaron al usuario en el pasado, prescindiendo con ello de la noción de similitud (sintáctica) tradicionalmente adoptada en los enfoques existentes.

El razonamiento semántico llevado a cabo por la metodología permite descubrir relaciones semánticas ocultas en la base de conocimiento del sistema, inferidas a partir de los enlaces jerárquicos y propiedades formalizadas en la misma. Hasta donde sabemos, este tipo de relaciones sólo han sido estudiadas en el contexto del proyecto de investigación SemDis, donde reciben el nombre de *asociaciones semánticas complejas*.

Si bien nuestro trabajo adopta la clasificación de asociaciones propuesta en SemDis, la ausencia de requisitos de personalización en dicho proyecto imposibilita la reutilización de la metodología de inferencia allí definida. En concreto, el objetivo perseguido en SemDis es encontrar todas las posibles asociaciones establecidas en una ontología RDFS entre dos instancias especificadas por un usuario, mientras que nuestro enfoque debe descubrir aquellos contenidos que están asociados (semánticamente) con las preferencias del usuario, y asegurar que tales asociaciones son significativas, preservando con ello el carácter personalizado de las sugerencias ofrecidas.

La metodología desarrollada para satisfacer este requisito consta de dos etapas: una primera *fase de filtrado*, en la que se seleccionan en la ontología OWL un conjunto de instancias y propiedades relevantes para el usuario, desde las que se infieren las asociaciones semánticas propuestas en SemDis; y una segunda *fase de validación*, en la que se realiza una búsqueda más exhaustiva sobre dichas entidades, con el objetivo de extraer de las mismas los contenidos finalmente sugeridos al usuario activo.

1. *Fase de filtrado*: En este etapa, el enfoque identifica en la ontología los programas que interesaron al usuario (registrados en su perfil con un nivel de interés positivo) y avanza, sucesivamente, a través de las propiedades que lo relacionan con otras instancias.

   A fin de garantizar la viabilidad computacional, este proceso se implementa mediante un mecanismo de búsqueda controlada. Cada vez que este mecanismo alcanza un conjunto de instancias desde una dada: (i) cuantifica la relevancia de las mismas, (ii) filtra todas aquellas que no son significativas para el usuario, y (iii) sigue avanzando únicamente a través de las propiedades que permiten alcanzar nuevas instancias desde aquéllas que son relevantes para el espectador. Como consecuencia, el enfoque únicamente explora entidades de interés para el usuario, omitiendo aquéllas que no aportan conocimiento útil para el proceso de recomendación.

   A la hora de cuantificar la relevancia de una instancia, nuestro enfoque considera que su peso es mayor cuanto más significativa es la relación existente entre ésta y las preferencias del usuario. Para cuantificar la fuerza de tal relación, el mecanismo propuesto emplea diferentes criterios como el número de propiedades explícitas existentes entre la instancia considerada y aquéllas referidas a contenidos que interesaron al usuario, el nivel de interés de estos últimos, la existencia de relaciones jerárquicas entre dicha instancia y las preferencias del espectador, e incluso, algunos parámetros utilizados en teoría de grafos que permiten detectar relaciones implícitas significativas entre un conjunto de nodos (instancias) del mismo.

   Una vez seleccionadas las instancias relacionadas con las preferencias del usuario y las propiedades que las unen entre sí, el enfoque debe inferir las asociaciones semánticas propuestas en SemDis[1] entre aquellas instancias que identifiquen

[1] Este proceso es inmediato y se basa en comprobar que las instancias y propiedades seleccionadas durante la fase de filtrado cumplen determinadas condiciones [5].

contenidos de TV concretos.

En este punto, la metodología dispone de una red genérica de nodos (instancias) y enlaces (propiedades y asociaciones semánticas), cuyo conocimiento debe ser explorado en la fase de validación.

2. *Fase de validación*: Para poder seleccionar los contenidos recomendados al usuario a partir de los nodos y relaciones formalizados en dicha red, nuestra metodología emplea las técnicas de propagación de activación (técnicas SA).

   Tal como se describió en la Sección 3.9.1, estas técnicas se basan en asignar un nivel de activación a cada nodo de la red y ponderar adecuadamente los enlaces establecidos en la misma. Tras activar un conjunto de nodos y aplicar un proceso de propagación —en el que los niveles de cada nodo se propagan a sus vecinos a través de los enlaces de la red—, las técnicas SA son capaces de identificar aquellos nodos que están relacionados de forma significativa con los inicialmente activados.

   La aplicación de las técnicas SA en el trabajo propuesto es inmediata: inicialmente, nuestra metodología activa aquellos nodos de la red que identifican los programas que interesaron al usuario en el pasado; una vez concluido el proceso de propagación, sugiere al mismo los contenidos con mayores niveles de activación.

   Para ello, es necesario ponderar convenientemente los enlaces formalizados en la red, en los que se representa tanto el conocimiento explícito representado en la ontología OWL (propiedades), como el conocimiento oculto inferido desde la misma durante la fase de filtrado (asociaciones semánticas). En nuestro enfoque, el peso de un enlace entre dos instancias se calcula a partir de los índices de relevancia asignados a las mismas en esa fase, mientras que la propagación de activación se realiza mediante el algoritmo Hopfield Net. Este algoritmo calcula el nivel de activación de cada nodo a partir de los niveles de sus vecinos en la red y de los pesos asignados a los enlaces que los unen, finalmente incluidos como argumento en una función sigmoide. Como consecuencia, una vez finalizado el proceso de propagación, los mayores niveles de activación corresponden a contenidos de TV cuyos vecinos son también relevantes en la red (de ahí que registren niveles elevados) y que, además, están relacionados con las preferencias del usuario (de ahí los elevados pesos de sus enlaces).

   Por su naturaleza, el mecanismo de propagación empleado en la metodología tiene dos ventajas especialmente importantes:

   - Por una parte, los procesos de razonamiento semántico consideran tanto las preferencias positivas como negativas del espectador. Mientras que las primeras ayudan a identificar contenidos interesantes para el usuario, las segundas decrementan los niveles de activación de los nodos de la red con los que están relacionados (ya sea explícitamente a través de propiedades, o implícitamente mediante asociaciones semánticas). De esta forma, la metodología desarrollada inhibe la recomendación de programas asociados con las preferencias negativas del espectador.
   - Por otro lado, nuestro enfoque no sólo favorece la reutilización de conocimiento, sino que permite además que la red de propagación se adapte, de forma flexible y dinámica, a las preferencias del usuario. A medida que

éstas evolucionan, la fase de filtrado infiere nuevas instancias, propiedades y asociaciones semánticas, y las incluye en la red de validación junto a las entidades inicialmente representadas en la misma. Entre estas entidades, es posible encontrar programas cuyos niveles de activación original no eran suficientemente significativos para ser incluidos en la recomendación; gracias al conocimiento que acaba de aprender el sistema, estos contenidos pueden ver incrementados sus niveles y ser finalmente sugeridos al usuario.

Además de descubrir programas que pasarían inadvertidos en los enfoques existentes, las capacidades de razonamiento semántico incorporadas en nuestra metodología permiten diversificar las recomendaciones basadas en contenido sin recurrir a técnicas colaborativas. La ausencia de estas últimas no sólo elimina un gran número de limitaciones (problemas de latencia, escalabilidad, *sparsity problem*), sino que permite elaborar sugerencias de calidad sin tener que conocer los perfiles (y niveles de interés) de un número elevado de usuarios.

Sin embargo, tal como adelantábamos en la Sección 9.1, fue precisamente la falta de medidas efectivas contra la *sobreespecialización* la que provocó la adopción generalizada del filtrado colaborativo y, sobre todo, de los enfoques híbridos en los sistemas actuales. Este escenario inspiró la segunda de las estrategias de personalización incluidas en nuestra metodología, basada en aplicar procesos de razonamiento semántico para mejorar los (numerosos) sistemas híbridos propuestos en la literatura.

Estrategia híbrida

Si bien los sistemas híbridos actuales permiten atenuar las limitaciones (individuales) de las estrategias que combinan, éstas aún pueden ser combatidas de forma más efectiva; éste es el camino explorado en la segunda de las estrategias descritas en este trabajo.

En primer lugar, esta técnica aplica una *fase de filtrado basado en contenido* y recomienda los programas objetivo a todos aquellos espectadores interesados en contenidos similares. Los usuarios que no cumplan tal requisito se convierten en candidatos para la segunda *etapa* de la estrategia, *basada en filtrado colaborativo*. De esta forma, la primera fase sugiere los contenidos objetivo que son similares a las preferencias del usuario activo, mientras que la segunda etapa recomienda aquellos programas que interesan a sus vecinos.

1. *Fase basada en contenido*: Para poder comparar las preferencias del usuario activo frente a los contenidos objetivo, nuestra metodología define una métrica de similitud semántica. Dicha métrica está basada en inferir relaciones semánticas entre los contenidos comparados, de forma que cuanto más significativas sean éstas, mayor será el valor de similitud cuantificado en el enfoque.

   Han sido varios los enfoques propuestos en la literatura que también recurren a este tipo de métricas para medir la similitud semántica entre conceptos genéricos. En la mayoría de las propuestas, tales conceptos están organizados en taxonomías, de ahí que las únicas relaciones consideradas sean los enlaces jerárquicos establecidos entre las clases a las que éstos pertenecen. Otros enfoques más avanzados extienden este tipo de métricas a formalizaciones más complejas, y cuantifican la similitud entre los conceptos de una ontología considerando tanto

las propiedades explícitas definidas en la misma, como las relaciones jerárquicas mencionadas antes.

Nuestra metodología extiende este tipo de enfoques y explora tanto el conocimiento explícitamente formalizado en la ontología OWL del sistema, como el conocimiento oculto en la misma, resultando una métrica de similitud con dos componentes claramente diferenciadas:

- *Similitud semántica jerárquica*: Al igual que en los enfoques tradicionales, el valor de esta componente depende únicamente de las relaciones jerárquicas establecidas en la ontología OWL entre los géneros a los que pertenecen los contenidos comparados (ver Sección 9.1.1). Gracias a estas relaciones, nuestro enfoque no sólo detecta similitud entre programas clasificados bajo un mismo género, sino también entre contenidos que pertenecen a géneros (distintos) que comparten un antecesor común en la jerarquía.
- *Similitud semántica inferencial*: Esta componente infiere asociaciones semánticas (implícitas) entre los programas comparados[2]. Para detectar tales asociaciones, nuestro enfoque se basa en las relaciones existentes entre sus respectivos atributos semánticos. En concreto, la métrica propuesta infiere que los contenidos comparados son similares cuando comparten los mismos atributos (relación de igualdad), o bien cuando éstos, aun siendo distintos, pertenecen a una misma clase (relación de hermandad). En la inferencia de estas relaciones también intervienen los procesos de razonamiento geográfico-temporales apuntados en la Sección 9.1.1, desde los que es posible inferir similitud entre contenidos relacionados con un mismo emplazamiento geográfico (e.g. regiones y países) y/o temporal (e.g. eventos históricos).

Nuestro enfoque permite combatir de forma efectiva el denominado *new user ramp-up*, causante de la baja calidad de las recomendaciones que reciben los usuarios que acaban de llegar al sistema. Gracias a esta métrica, el recomendador puede completar las preferencias —típicamente inmaduras— de estos usuarios, descubriendo tanto contenidos pertenecientes a los géneros que le interesaron en el pasado (o a otros relacionados jerárquicamente con ellos), como programas que tienen atributos semánticos *similares* (que no necesariamente idénticos) a los registrados en sus perfiles.

Tras sugerir los contenidos objetivo a todos los usuarios que han visto programas parecidos a éstos, nuestro enfoque da una segunda oportunidad a aquellos espectadores que no cumplen tal condición. Estos usuarios se convierten así en candidatos a ser evaluados en la fase colaborativa de la estrategia.

2. *Fase colaborativa*: Esta etapa predice el nivel de interés de los usuarios candidatos en relación a los contenidos objetivo, basándose en las preferencias de sus

[2]Conviene aclarar que las asociaciones consideradas en esta métrica son menos indirectas que las inferidas en la estrategia basada en contenido descrita previamente. El objetivo de esta última era descubrir contenidos asociados semánticamente con los que interesaron al usuario, eliminado con ello el carácter *sobreespecializado* de las recomendaciones tradicionales. Por el contrario, en la estrategia híbrida, es el enfoque colaborativo el que diversifica las sugerencias, de forma que la fase basada en contenido únicamente debe cuantificar el parecido entre los contenidos objetivo (seleccionados de antemano) y las preferencias del espectador. Exigiendo asociaciones directas entre los contenidos comparados, garantizamos la fiabilidad de los valores de similitud cuantificados por la métrica.

respectivos vecinos. La metodología desarrollada incorpora el razonamiento semántico tanto en la fase de creación de los vecindarios, como en el proceso de predicción de los índices de interés.

- *Creación del vecindario*: Nuestro enfoque incorpora un mecanismo para delimitar el vecindario de cada usuario candidato, capaz de mitigar los efectos del *sparsity problem* asociado a los sistemas colaborativos convencionales. Para ello, en lugar de seleccionar los vecinos de un usuario basándose únicamente en los programas incluidos en sus respectivos perfiles, nuestra metodología considera la jerarquía de géneros definida en los mismos (ver Sección 9.1.2). Como consecuencia, nuestro enfoque es capaz de inferir que las preferencias de dos usuarios son similares aun cuando no hayan clasificado los mismos contenidos en sus perfiles; la única exigencia que impone la metodología es que estos contenidos estén clasificados bajo los mismos géneros en la ontología OWL (por ejemplo, en el vecindario de un usuario que ha visto un partido de fútbol, puede estar incluido otro espectador que haya visto un partido de baloncesto, ya que ambos espectadores comparten interés por el género deportivo).

  Por otra parte, este enfoque también beneficia la escalabilidad del sistema recomendador, dado que un incremento en el número de programas registrados en los perfiles de los usuarios, no se traduce necesariamente en nuestro modelo en un aumento de la complejidad del proceso de formación de sus respectivos vecindarios. Esto último es debido a que muchos de los contenidos considerados podrían estar clasificados bajo los mismos géneros en la ontología.
- *Predicción del nivel de interés*: Según lo explicado en la Sección 9.1.3, a la hora de predecir el nivel de interés de los usuarios candidatos en relación a los contenidos objetivo, los enfoques colaborativos actuales únicamente considerarían las contribuciones de aquellos vecinos que ya hubiesen clasificado estos programas en sus perfiles.

  A diferencia de este tipo de aproximaciones, nuestra metodología examina el vecindario completo de cada usuario candidato. Así, si un vecino ya conoce el contenido objetivo considerado, la predicción se basa en el índice de interés que figura en su perfil; en caso contrario, nuestra estrategia predice este nivel de interés a partir de la similitud semántica existente entre ese programa y las preferencias de dicho vecino. Gracias a este enfoque, nuestra estrategia colaborativa elimina los tiempos de latencia presentes en los sistemas actuales, causantes de que los contenidos que acaban de llegar al sistema sean sugeridos a los usuarios con ciertos retardos innecesarios. Así, en lugar de esperar a que los vecinos del usuario activo clasifiquen los nuevos programas, nuestro enfoque infiere su interés —mediante el razonamiento semántico— y lo incluye directamente en el proceso de predicción.

Una vez desarrolladas la ontología OWL, la técnica de modelado de usuarios y las estrategias de personalización revisadas en esta sección, es necesario identificar su localización en la arquitectura del sistema recomendador AVATAR. Dado el dominio de aplicación del sistema, la arquitectura se ha distribuido entre los proveedores que ofrecen los servicios de personalización, y los usuarios finales que acceden a éstos a través de sus receptores (Set-Top Boxes o STBs), permanentemente conectados a través de un canal de retorno.

9.1.4 Arquitectura

Son dos las características más reseñables de la arquitectura propuesta en el sistema AVATAR. La primera de ellas está relacionada con la adopción de diferentes formatos y tecnologías normalizadas (como MHP, TV-Anytime y OWL) que, además de gozar de gran popularidad, promueven un uso extendido de las capacidades automáticas de personalización proporcionadas por el sistema. La segunda de las características comentadas es el carácter abierto y modular conferido a su arquitectura gracias a la incorporación de múltiples agentes que cooperan activamente durante el proceso de elaboración de las recomendaciones. Algunos de estos agentes se ejecutan en el servidor del proveedor de servicios, mientras que otros residen en el STB del usuario.

- Los primeros se encargan de: (i) mantener la ontología de conocimiento actualizada, incluyendo nuevos contenidos de TV (y sus descripciones semánticas) a medida que estén disponibles, (ii) almacenar y actualizar los perfiles de los usuarios registrados en el sistema, y (iii) ejecutar las dos estrategias de personalización para seleccionar los programas finalmente recomendados.
- Entre las tareas de los agentes que implementan las aplicaciones MHP ejecutadas en el STB, destacamos: (i) la presentación de las recomendaciones personalizadas, (ii) la monitorización de las acciones que llevan a cabo los usuarios, y (iii) el envío de la información que proporcionan (programas aceptados y rechazados) para que el agente correspondiente actualice sus perfiles en el servidor del proveedor.

Esta organización distribuida del servicio de personalización entre el proveedor y el STB del usuario responde a dos razones:

- En primer lugar, los costes computacionales asociados, por una parte, al mantenimiento de la ontología OWL y los perfiles de los usuarios y, por otra, a las estrategias de personalización incluidas en nuestra metodología, resultan excesivos para los receptores actuales. Por el contrario, el proveedor de servicios dispondrá típicamente de servidores con capacidades de almacenamiento y procesado capaces de atender sobradamente este coste computacional.
- Por otro lado, la presencia del filtrado colaborativo como estrategia de personalización obliga al sistema a conocer las preferencias de todos los usuarios registrados en el mismo, de ahí que sea el proveedor quien elabora las recomendaciones finalmente presentadas en el STB del usuario.

La arquitectura abierta y modular propuesta en nuestro trabajo permite que el sistema AVATAR pueda ser modificado fácilmente y, con ello, evolucionar en el tiempo. En este proceso no sólo tiene cabida la incorporación de nuevas estrategias de recomendación, sino también la adopción de estándares futuros que, junto a MHP, TV-Anytime y OWL, favorezcan la reutilización de sus capacidades de personalización.

9.1.5 Implementación

Para poder validar la metodología de razonamiento semántico desarrollada en la tesis, hemos implementado una herramienta en la que se han incluido todos los componentes desgranados en las secciones previas.

Esta herramienta permite modelar las preferencias de los usuarios y seleccionar contenidos personalizados mediante las dos estrategias incluidas en el enfoque. Una vez mostradas estas recomendaciones, la herramienta simula la interacción de los usuarios modelados con el recomendador (aceptando algunos contenidos y rechazando otros) y actualiza convenientemente sus perfiles personales. Para ello, ajusta de forma automática tanto sus niveles de interés como los índices de olvido que permiten la adaptación del sistema a medida que evolucionan sus preferencias.

La ontología OWL de TV ha sido implementada mediante el editor Protégé, una herramienta gráfica desarrollada en la Universidad de Stanford, que permite añadir clases, propiedades e instancias, de una forma cómoda e intuitiva, sin necesidad de trabajar sobre los constructores del lenguaje adoptado. Además de esta interfaz gráfica, Protégé proporciona también una API para acceder a las entidades de la base de conocimiento OWL desde una aplicación Java, de ahí que hayamos elegido este lenguaje para la implementación de la herramienta. Esta librería también incluye métodos para poder conectar la herramienta desarrollada con un razonador externo. Por su implementación altamente optimizada, hemos seleccionado el motor lógico RACER, capaz de comprobar la consistencia de nuestra ontología de conocimiento y realizar inferencias lógicas automáticas sobre la misma.

Para propósitos de prueba y depuración, hemos desarrollado, además, dos herramientas auxiliares:

- La primera permite popular la base de conocimiento OWL añadiendo programas específicos y atributos semánticos de los mismos. Esta herramienta extrae la información necesaria desde tres fuentes: las bases de datos IMDB (*Internet Movie Data Base*) y AMG (*All Movie Guide*), donde se describen una gran variedad de contenidos de TV, y el servidor Web de la BBC (*British Broadcasting Corporation*), desde donde es posible acceder a descripciones semánticas conformes a TV-Anytime.
- La segunda herramienta implementada permite trazar en la ontología OWL las asociaciones semánticas que infiere el sistema durante el proceso de recomendación, ayudando a entender y seguir los razonamientos semánticos utilizados en la metodología. Para ello, la herramienta presenta un extracto de la ontología en el que, junto a las asociaciones inferidas entre las preferencias del usuario y los programas disponibles en el sistema, se representan también las relaciones jerárquicas y las propiedades que las originan.

El prototipo implementado ha sido validado en diferentes escenarios, en los que se ha estudiado la influencia tanto del número de usuarios modelados en el sistema (entre 50 y 75 espectadores) como de la cantidad de programas de TV incluidos en la ontología de conocimiento (entre 400 y 520 contenidos). Las pruebas realizadas revelan que los costes computacionales asociados a nuestra metodología son asumibles en un entorno de aplicación como el propuesto. En escenarios más complejos (con un número mayor de usuarios y/o contenidos), son varias las razones que nos llevan a pensar que estos costes siguen sin limitar la validez del enfoque propuesto:

- La primera de ellas está relacionada con la ausencia de requisitos de tiempo real en el sistema recomendador. En un escenario realista, los usuarios no están comprobando las recomendaciones continuamente, sino que registran sus preferencias en el sistema para que éste les envíe una lista de contenidos personalizados cuando estén disponibles. Dado que, en este contexto, el recomendador

conoce las preferencias de los espectadores y la programación con suficiente antelación, podemos concluir que la complejidad computacional de las estrategias desarrolladas no introduce retardos (en la presentación de las sugerencias) inasumibles para los usuarios.

- Por otra parte, tal como adelantábamos en la sección anterior, los servidores utilizados en los proveedores de servicios disponen de capacidades de almacenamiento y procesado capaces de asegurar la viabilidad computacional de nuestra metodología.
- Por último, recordar en este punto alguna de las medidas incorporadas en nuestro trabajo para asegurar la escalabilidad del sistema.
    - La primera de ellas ya ha sido apuntada al describir la técnica de modelado de usuarios propuesta en el enfoque. Esta técnica modela las preferencias de los espectadores a partir de un repositorio de conocimiento común (la ontología OWL de TV). De esta forma, no es necesario replicar en cada perfil la información que identifica tales preferencias (clases, propiedades e instancias), permitiendo rebajar las capacidades de almacenamiento requeridas en el proveedor de servicios.
    - La reutilización de conocimiento asociada a la estrategia basada en contenido también contribuye a aliviar la carga computacional del sistema. Esta reutilización se ve favorecida gracias a la red de propagación de activación utilizada en el enfoque, en la que se combina el conocimiento aprendido en el pasado con las nuevas preferencias del usuario.
    - Por último, la adopción del enfoque jerárquico para seleccionar el vecindario de los usuarios en la estrategia híbrida, también reporta ciertos beneficios computacionales. Así, al aumentar el número de programas disponibles en los perfiles de los usuarios, no tiene por qué aumentar el coste computacional del enfoque, dado que muchos de estos contenidos pueden pertenecer a las mismas categorías jerárquicas (géneros) en la ontología OWL.

9.2 Líneas de trabajo futuro

Son varios los frentes que podemos acometer para seguir mejorando la funcionalidad de los sistemas recomendadores mediante procesos de razonamiento semántico:

- En primer lugar, apuntamos una línea de trabajo que extendería, de manera natural, la metodogía desarrollada en esta tesis. Nos referimos a un mecanismo que combine, de forma eficiente, las dos estrategias de personalización incluidas en nuestro enfoque. Dado un usuario, este mecanismo debería ejecutar ambas técnicas y decidir cuáles de las recomendaciones ofrecidas son más interesantes para el mismo.

  Para ello, es necesario enriquecer los procesos de razonamiento semántico aplicados en nuestro trabajo, incluyendo ciertas capacidades de aprendizaje y procesamiento automático. Dado que las redes neuronales incorporan tales capacidades, creemos que estos sistemas clasificadores podrían ser adoptados como punto de partida en el proceso de combinación.

- Otra de las líneas de investigación que convendría explorar es la capacidad de seleccionar recomendaciones personalizadas cuando la audiencia objetivo del sistema es múltiple. En este escenario, un grupo de usuarios se reúnen para utilizar a la vez las sugerencias del recomendador, de ahí que el modelado de sus preferencias conjuntas sea un elemento clave en el proceso de personalización.

  A la hora de abordar esta tarea, el sistema podría aprovechar las capacidades inferenciales incorporadas en nuestra metodología. Un posible enfoque consistiría en descubrir asociaciones semánticas entre las preferencias (individuales) de todos los usuarios, y utilizarlas para trazar patrones de (des)interés comunes a todos ellos. Gracias al conocimiento inferido en este proceso, el enfoque modelaría un perfil conjunto mejorado que permitiría que las estrategias de personalización seleccionase únicamente recomendaciones válidas para todos los usuarios que integran la audiencia múltiple del sistema.

  En este escenario, habría que modificar diferentes aspectos de la metodología desarrollada en nuestro trabajo. Además de modelar un perfil conjunto, es necesario estudiar también el proceso de adaptación del recomendador a medida que evolucionan las preferencias de los usuarios, así como el impacto que pueda tener el abandono y/o la incorporación de nuevos integrantes al grupo. En este sentido, habría que identificar tanto el conocimiento que se puede seguir reutilizando durante la elaboración de las sugerencias, como aquel que ya no es válido en el nuevo escenario de aplicación.

- La última línea de trabajo que apuntamos consiste en incluir en el proceso de razonamiento semántico, información sobre el contexto actual del usuario (su estado anímico, su localización geográfica y temporal actual, el tipo de audiencia —única o múltiple— a la que pertenece, etc.). En el dominio de la televisión, algunos de estos parámetros son especialmente interantes de cara al proceso de personalización. En concreto, el sistema recomendador puede adaptar sus sugerencias a la franja horaria que prefiere el usuario para ver televisión o a su estado anímico actual. En el primer caso, el sistema únicamente debe vigilar las horas de emisión de los contenidos que interesan al espectador, mientras que en el segundo, debe disponer además de la infraestructura adecuada para intuir algunas de sus características emocionales. Para implementar este tipo de soluciones, algunas líneas de investigación recientes exploran el campo de los sensores táctices, capaces de detectar las sensaciones de intranquilidad, nerviosismo, excitación o cansancio, entre otras, que experimente el usuario.

  Actualmente, nuestra metodología centra sus razonamientos en la semántica de los contenidos y las preferencias del usuario. Combinando dicha semántica con su contexto específico sería posible reforzar el carácter personalizado de las recomendaciones. Para ello, el sistema podría utilizar un conjunto de reglas lógicas que relacionasen la información presente en el contexto del usuario con el conocimiento inferido en nuestra metodología. Mediante dichas reglas, el sistema podría deducir cuáles de las asociaciones semánticas inferidas son más relevante para el usuario, adaptando así las recomendaciones a su situación particular (por ejemplo, debería primar sugerencias relacionadas con la localización geográfica actual del usuario, omitir algunas de las recomendaciones seleccionadas por no ser válidas para todos los usuarios que le acompañan, etc.).

Parte VI

Apéndices

Apéndice A

TV-Anytime

En este apéndice se presenta una clasificación completa de los atributos semánticos definidos en la especificación de metadatos TV-Anytime, junto a los códigos (alfa)numéricos que los identifican de forma unívoca.

Tabla A.1: Clasificación de géneros definidos en TV-Anytime

Código numérico	Género
3.1	*NON-FICTION*
3.1.1	*News*
3.1.1.1	*Daily news*
3.1.1.2	*Special news*
3.1.1.3	*Special reports*
3.1.1.4	*Commentary*
3.1.1.5	*Periodical/General*
3.1.1.6	*National politics*
3.1.1.7	*Economy/Market/Finantial/Business*
3.1.1.8	*Foreigh/International*
3.1.1.9	*Sports*
3.1.1.10	*Cultural*
3.1.1.11	*Local/regional*
3.1.1.12	*Traffic*
3.1.1.13	*Weather forecasts*
3.1.1.14	*Service information*
3.1.1.15	*Public affairs*
3.1.1.16	*Current affairs*
3.1.2	*Philosophies of life*
3.1.2.1	*Religion*
3.1.2.1.1	*Buddhism*
3.1.2.1.2	*Hinduism*
3.1.2.1.3	*Christianity*

Código numérico	Género
3.1.2.1.4	*Islam*
3.1.2.1.5	*Judaism*
3.1.2.1.6	*Atheism*
3.1.2.1.7	*Agnosticism*
3.1.2.1.8	*Shintoism*
3.1.2.2	*Non-religious philosophies*
3.1.2.2.1	*Communism*
3.1.2.2.2	*Humanism*
3.1.2.2.3	*Capitalism*
3.1.2.2.4	*Socialism*
3.1.2.2.5	*Libertarianism*
3.1.2.2.6	*Republicanism*
3.1.3	*General non-fiction*
3.1.3.1	*Political*
3.1.3.2	*Social*
3.1.3.3	*Economic*
3.1.3.4	*Legal*
3.1.3.5	*Finance*
3.1.3.6	*Education*
3.1.3.7	*International affairs*
3.1.3.8	*Military/Defence*
3.1.4	*Arts/Media*
3.1.4.1	*Music*
3.1.4.2	*Dance*
3.1.4.3	*Theatre*
3.1.4.4	*Opera*
3.1.4.5	*Cinema*
3.1.4.6	*Advertising*
3.1.4.7	*Press*
3.1.4.8	*Plastic arts*
3.1.4.9	*Fine arts*
3.1.4.10	*Experimental arts*
3.1.4.11	*Architecture*
3.1.4.12	*Showbiz*
3.1.4.13	*Television*
3.1.4.14	*Radio*
3.1.4.15	*News media*
3.1.5	*Humanities*
3.1.5.1	*Literature*
3.1.5.2	*Languages*
3.1.5.3	*History*
3.1.5.4	*Culture/Tradition/Anthropology/Ethnic studies*
3.1.5.5	*War/Conflicts*

Código numérico	Género
3.1.6	*Sciences*
3.1.6.1	*Applied sciences*
3.1.6.2	*Nature/Natural sciences*
3.1.6.3	*Animals/Wildlife*
3.1.6.4	*Environment/Geography*
3.1.6.5	*Space/Universe*
3.1.6.6	*Physical sciences*
3.1.6.7	*Medicine*
3.1.6.8	*Technology*
3.1.6.9	*Physiology*
3.1.6.10	*Psychology*
3.1.6.11	*Social*
3.1.6.12	*Spiritual*
3.1.6.13	*Mathematics*
3.1.6.14	*Archaelogy*
3.1.7	*Human interest*
3.1.7.1	*Reality*
3.1.7.2	*Society/Show business/Gossip*
3.1.7.3	*Biography/Notable personalities*
3.1.7.4	*Personal problems*
3.1.7.5	*Investigative journalism*
3.1.7.6	*Museums*
3.1.7.7	*Religious buildings*
3.1.8	*Transport and communications*
3.1.8.1	*Air*
3.1.8.2	*Land*
3.1.8.3	*Sea*
3.1.8.4	*Space*
3.1.9	*Events*
3.1.9.1	*Anniversary*
3.1.9.2	*Fair*
3.1.9.3	*Tradeshow*
3.1.9.4	*Musical*
3.1.9.5	*Exhibition*
3.2	*SPORTS*
3.2.1	*Athletics*
3.2.1.1	*Field*
3.2.1.2	*Track*
3.2.1.3	*Combined athletics*
3.2.1.4	*Running*
3.2.1.5	*Cross-county*
3.2.1.6	*Triathlon*
3.2.2	*Cycling/Bicycle*
3.2.2.1	*Mountain bike*
3.2.2.2	*Bicross*

Código numérico	Género
3.2.2.3	*Indoor cycling*
3.2.2.4	*Road cycling*
3.2.3	*Team sports*
3.2.3.1	*Football (american)*
3.2.3.2	*Football (australian)*
3.2.3.3	*Football (gaelic)*
3.2.3.4	*Football (indoor)*
3.2.3.5	*Beach soccer*
3.2.3.6	*Bandy*
3.2.3.7	*Baseball*
3.2.3.8	*Basketball*
3.2.3.9	*Cricket*
3.2.3.10	*Croquet*
3.2.3.11	*Faustball*
3.2.3.12	*Football (soccer)*
3.2.3.13	*Handball*
3.2.3.14	*Hockey*
3.2.3.15	*Korfball*
3.2.3.16	*Lacrosse*
3.2.3.17	*Netball*
3.2.3.18	*Roller skating*
3.2.3.19	*Rugby*
3.2.3.19.1	*Rugby union*
3.2.3.19.2	*Rugby leage*
3.2.3.20	*Softball*
3.2.3.21	*Volleyball*
3.2.3.22	*Beach volley*
3.2.3.23	*Hurling*
3.2.3.24	*Flying disc/Frisbee*
3.2.4	*Racket sports*
3.2.4.1	*Badmington*
3.2.4.2	*Racketball*
3.2.4.3	*Short tennis*
3.2.4.4	*Soft tennis*
3.2.4.5	*Squash*
3.2.4.6	*Table tennis*
3.2.4.7	*Tennis*
3.2.5	*Martial arts*
3.2.5.1	*Aikido*
3.2.5.2	*Jai-alai*
3.2.5.3	*Judo*
3.2.5.4	*Ju-jitsu*
3.2.5.5	*Karate*
3.2.5.6	*Sumo/Fighting games*
3.2.5.7	*Sambo*
3.2.5.8	*Taekwondo*

Código numérico	Género
3.2.6	*Water sports*
3.2.6.1	*Bodyboarding*
3.2.6.2	*Yatching*
3.2.6.3	*Canoeing*
3.2.6.4	*Diving*
3.2.6.5	*Fishing*
3.2.6.6	*Polo*
3.2.6.7	*Rowing*
3.2.6.8	*Sailing*
3.2.6.9	*Sub-aquatics*
3.2.6.10	*Surfing*
3.2.6.11	*Swimming*
3.2.6.12	*Water polo*
3.2.6.13	*Water skiing*
3.2.6.14	*Windsurfing*
3.2.7	*Winter sports*
3.2.7.1	*Bobsleigh/Tobogganing*
3.2.7.2	*Curling*
3.2.7.3	*Ice-hockey*
3.2.7.4	*Ice-skating*
3.2.7.5	*Luge*
3.2.7.6	*Skating*
3.2.7.7	*Skibob*
3.2.7.8	*Skiing*
3.2.7.9	*Sleddog*
3.2.7.10	*Snowboarding*
3.2.7.11	*Alpine skiing*
3.2.7.12	*Freestyle skiing*
3.2.7.13	*Inline skating*
3.2.7.14	*Nordic skiing*
3.2.7.15	*Ski jumping*
3.2.7.16	*Speed skating*
3.2.7.17	*Figure skating*
3.2.7.18	*Ice-Dance*
3.2.7.19	*Marathon*
3.2.7.20	*Short-track*
3.2.7.21	*Biathlon*
3.2.8	*Motor sports*
3.2.8.1	*Motor/Auto racing*
3.2.8.2	*Motor boating/Motor racing*
3.2.8.3	*Motor cycling*
3.2.8.4	*Formula 1*
3.2.8.5	*Indy car*
3.2.8.6	*Karting*
3.2.8.7	*Rally*
3.2.8.8	*Trucking*

Código numérico	Género
3.2.8.9	*Tractor pulling*
3.2.8.10	*Auto racing*
3.2.9	*Social sports*
3.2.9.1	*Billiards*
3.2.9.2	*Boules*
3.2.9.3	*Bowing*
3.2.9.4	*Chess*
3.2.9.5	*Dance sport*
3.2.9.6	*Darts*
3.2.9.7	*Pool*
3.2.9.8	*Snooker*
3.2.9.9	*Tug-of-war*
3.2.9.10	*Balle pelote*
3.2.9.11	*Basque pelote*
3.2.9.12	*Trickshot*
3.2.10	*Gymnastics*
3.2.10.1	*Assymetric bars*
3.2.10.2	*Beam*
3.2.10.3	*Horse*
3.2.10.4	*Mat*
3.2.10.5	*Parallel bars*
3.2.10.6	*Rings*
3.2.10.7	*Trampoline*
3.2.11	*Equestrian*
3.2.11.1	*Cart*
3.2.11.2	*Dressage*
3.2.11.3	*Horse racing*
3.2.11.4	*Polo*
3.2.11.5	*Jumping*
3.2.11.6	*Crossing*
3.2.11.7	*Trotting*
3.2.12	*Adventure sports*
3.2.12.1	*Archery*
3.2.12.2	*Extreme sports*
3.2.12.3	*Mountaineering*
3.2.12.4	*Climbing*
3.2.12.5	*Orienteering*
3.2.12.6	*Shooting*
3.2.12.7	*Sport acrobatics*
3.2.12.8	*Rafting*
3.2.13	*Strength-based sports*
3.2.13.1	*Body-building*
3.2.13.2	*Boxing*
3.2.13.3	*Combative sports*
3.2.13.4	*Power-lifting*
3.2.13.5	*Weight-lifting*

Código numérico	Género
3.2.13.6	*Wrestling*
3.2.14	*Air sports*
3.2.14.1	*Ballooning*
3.2.14.2	*Hang gliding*
3.2.14.3	*Sky diving*
3.2.14.4	*Delta-plane*
3.2.14.5	*Parachuting*
3.2.14.6	*Kiting*
3.2.14.7	*Aeronautics*
3.2.14.8	*Gliding*
3.2.14.9	*Flying*
3.2.14.10	*Aerobatics*
3.2.15	*Golf*
3.2.16	*Fencing*
3.2.17	*Dog racing*
3.2.18	*Casting*
3.2.19	*Maccabi*
3.2.20	*Modern Pentathlon*
3.2.21	*Sombo*
3.3	*LEISURE/HOBBY*
3.3.1	*Do-it-yourself*
3.3.2	*Cookery*
3.3.3	*Gardening*
3.3.4	*Travel/Tourism*
3.3.5	*Adventure/Expeditions*
3.3.6	*Fishing*
3.3.7	*Outdoor*
3.3.8	*Pet*
3.3.9	*Craft/Handicraft*
3.3.10	*Art*
3.3.11	*Music*
3.3.12	*Board games*
3.3.13	*Computer games*
3.3.14	*Card games*
3.3.15	*Fitness/Keep-fit*
3.3.16	*Personal health*
3.3.17	*Car*
3.3.18	*Motorcycle/Motoring*
3.3.19	*Fashion*
3.3.20	*Life/House Keeping/Lifestyle*
3.3.21	*Technology/Computing*
3.3.22	*Gaming*
3.3.23	*Shopping*
3.3.24	*Adult*
3.3.25	*Road safety*
3.3.26	*Consumer advice*

Código numérico	Género
3.3.27	*Employment advice*
3.3.28	*Boating*
3.3.29	*Parenting*
3.3.30	*Self-help*
3.3.31	*Collectibles*
3.3.32	*Jewellery*
3.3.33	*Beauty*
3.3.34	*Aviation*
3.4	*FICTION*
3.4.1	*General light drama*
3.4.2	*Soap*
3.4.2.1	*Soap Opera*
3.4.2.2	*Soap Special*
3.4.2.3	*Soap Talk*
3.4.3	*Romance*
3.4.4	*Legal Melodrama*
3.4.5	*Medical Melodrama*
3.4.6	*Action*
3.4.6.1	*Adventure*
3.4.6.2	*Disaster*
3.4.6.3	*Mystery*
3.4.6.4	*Detective*
3.4.6.5	*Historical*
3.4.6.6	*Horror*
3.4.6.7	*Science fiction*
3.4.6.8	*War*
3.4.6.9	*Western*
3.4.6.10	*Thriller*
3.4.6.11	*Sports*
3.4.6.12	*Martial arts*
3.4.6.13	*Epic*
3.4.7	*Fantasy/Fairy tale*
3.4.8	*Erotica*
3.4.9	*Docudrama*
3.4.10	*Musical*
3.4.11	*Comedy*
3.4.12	*Effect movies*
3.4.13	*Classical drama*
3.4.14	*Period drama*
3.4.15	*Contemporary drama*
3.4.16	*Religious*
3.4.17	*Poems*
3.4.18	*Biography*
3.4.19	*Psychological drama*

Código numérico	Género
3.5	*AMUSEMENT*
3.5.1	*Game show*
3.5.2	*Quiz/Contest*
3.5.3	*Variety Show*
3.5.4	*Surprise Show*
3.5.5	*Reality Show*
3.5.6	*Candid camera*
3.5.7	*Comedy*
3.5.7.1	*Broken comedy*
3.5.7.2	*Romantic comedy*
3.5.7.3	*Sitcom*
3.5.7.4	*Satire*
3.5.9	*Humour*
3.5.10	*Magic/Hypnotism*
3.5.11	*Circus*
3.5.12	*Dating show*
3.5.13	*Bullfighting*
3.5.14	*Rodeo*
3.5.15	*Airshow*
3.6	*MUSIC*
3.6.1	*Classical music*
3.6.1.1	*Early*
3.6.1.2	*Classical*
3.6.1.3	*Romantic*
3.6.1.4	*Contemporary*
3.6.1.5	*Light classical*
3.6.1.6	*Middle Age*
3.6.1.7	*Renaissance*
3.6.1.8	*Baroque*
3.6.1.9	*Opera*
3.6.1.10	*Solo instruments*
3.6.1.11	*Chamber*
3.6.1.12	*Symphonic*
3.6.1.13	*Vocal*
3.6.1.14	*Choral*
3.6.2	*Jazz*
3.6.2.1	*Early jazz*
3.6.2.2	*Big band/Swing*
3.6.2.3	*Blues*
3.6.2.4	*Bop/Hrd bop/Bebop*
3.6.2.5	*Traditional/Smooth*
3.6.2.6	*Cool/Free*
3.6.2.7	*Avant-garde*
3.6.2.8	*World jazz*

Código numérico	Género
3.6.2.9	*Pop jazz*
3.6.2.10	*Acid jazz/Fusion*
3.6.3	*Background music*
3.6.3.1	*Middle-of-the-road*
3.6.3.2	*Easy listening*
3.6.3.3	*Ambient*
3.6.3.4	*Mood music*
3.6.3.5	*Oldies*
3.6.3.6	*Love songs*
3.6.3.7	*Dance hall*
3.6.3.8	*Soundtrack*
3.6.3.9	*Trailer*
3.6.3.10	*Showtunes*
3.6.3.11	*TV*
3.6.3.12	*Cabaret*
3.6.3.13	*Instrumental*
3.6.3.14	*Sound clip*
3.6.3.15	*Retro*
3.6.4	*Pop-rock*
3.6.4.1	*Pop*
3.6.4.2	*Ballad*
3.6.4.3	*Traditional rock and roll*
3.6.4.4	*Slow rock*
3.6.4.5	*Classic/Dance/Pop-rock*
3.6.4.6	*Folk*
3.6.4.7	*Punk*
3.6.4.8	*New Age*
3.6.4.9	*Symphonic rock*
3.6.4.10	*Art-rock*
3.6.4.11	*Seasonal/Holiday*
3.6.4.12	*Japanese pop-rock*
3.6.4.13	*Karaoke/Singing contests*
3.6.5	*Soul/Gospel*
3.6.5.1	*Blues*
3.6.5.2	*R & B*
3.6.5.3	*Soul*
3.6.5.4	*Gospel*
3.6.6	*Country and Western*
3.6.7	*Rap/Hip Hop/Reggae*
3.6.7.1	*Rap/Christian rap*
3.6.7.2	*Hip Hop/Trip-Hop*
3.6.7.3	*Reggae*
3.6.7.4	*Ska/Gangsta*
3.6.8	*Electronic/Club/Urban/Dance*

Código numérico	Género
3.6.8.1	*Acid punk*
3.6.8.2	*Disco*
3.6.8.3	*Techno*
3.6.8.4	*House/Techno House*
3.6.8.5	*Rave*
3.6.8.6	*Jungle/Tribal*
3.6.8.7	*Trance*
3.6.8.8	*Punk*
3.6.8.9	*Garage/Psychadelic*
3.6.8.10	*Metal*
3.6.8.11	*Drum and Bass*
3.6.8.12	*Pranks*
3.6.8.13	*Grunge*
3.6.8.14	*Dance/Dance-pop*
3.6.9	*World/Traditional/Ethnic/Folk Music*
3.6.9.1	*Africa*
3.6.9.2	*Asia*
3.6.9.3	*Australia/Oceania*
3.6.9.4	*Caribbean*
3.6.9.5	*Europe*
3.6.9.6	*Latin America*
3.6.9.7	*Middle East*
3.6.9.8	*North Africa*
3.6.10	*Hit-Chart/Song Requests*
3.6.11	*Children's Songs*
3.6.12	*Event music*
3.6.12.1	*Wedding*
3.6.12.2	*Sports*
3.6.12.3	*Ceremonial/Chants*
3.6.13	*Spoken*
3.6.14	*Dance*
3.6.14.1	*Ballet*
3.6.14.2	*Tap*
3.6.14.3	*Modern*
3.6.14.4	*Classical*
3.7	*INTERACTIVE GAMES*
3.7.1	*Content games categories*
3.7.1.1	*Action*
3.7.1.2	*Adventure*
3.7.1.3	*Fighting*
3.7.1.4	*Online*
3.7.1.5	*Platform*
3.7.1.6	*Puzzle*
3.7.1.7	*RPG/MUDs*

Código numérico	Género
3.7.1.8	*Racing*
3.7.1.9	*Simulation*
3.7.1.10	*Sports*
3.7.1.11	*Strategy*
3.7.1.12	*Wrestling*
3.7.1.13	*Classic/Retro*
3.7.2	*Style*
3.7.2.1	*Logic based*
3.7.2.2	*Word games*
3.7.2.3	*Positional*
3.7.2.4	*Board games*
3.7.2.5	*Text environments*
3.7.2.6	*Dynamic 2D/3D graphics*
3.7.2.7	*Non-linear videos*

Tabla A.2: Clasificación de alertas de contenido

Código numérico	Alertas de contenido
6.1	*SEX*
6.1.1	*No sex descriptor*
6.1.2	*Obscured or implied sexual activity*
6.1.3	*Frank portrayal of sex and sexuality*
6.1.4	*Scenes of explicit sexual behaviour suitable for adults*
6.1.5	*Sexual violence*
6.2	*NUDITY*
6.2.1	*No nudity descriptor*
6.2.2	*Partial nudity*
6.2.3	*Full frontal nudity*
6.3	*VIOLENCE - HUMAN BEINGS*
6.3.1	*No violence descriptor*
6.3.2	*Deliberate infliction of pain to human*
6.3.3	*Infliction of strong psychological or physical pain to human*
6.3.4	*Deliberate killings of human beings*
6.4	*VIOLENCE - ANIMALS*
6.4.1	*No violence descriptor*
6.4.2	*Deliberate infliction of pain to animals*
6.4.3	*Deliberate killings of animals*
6.5	*VIOLENCE - FANTASY CHARACTERS*
6.5.1	*No violence descriptor*
6.5.2	*Deliberate infliction of pain to fantasy characters*

Código numérico	Alertas de contenido
6.6	*LANGUAGE*
6.6.1	*No language descriptor*
6.6.2	*Occasional use of mild swear words*
6.6.3	*Frequent use of mild swear words and profanities*
6.6.4	*Occasional use of very strong language*
6.6.5	*Frequent use of very strong language*
6.7	*DISTURBING SCENES*
6.7.1	*No disturbing scenes descriptor*
6.7.2	*Traumatic events*
6.7.3	*Mild scenes of blood and gore*
6.7.4	*Severe scenes of blood and gore*
6.7.5	*Extreme horror effects*
6.8	*DISCRIMINATION*
6.8.1	*No discrimination descriptor*
6.8.2	*Deliberate discrimination*
6.9	*ILLEGAL DRUGS*
6.9.1	*No illegal drugs descriptor*
6.9.2	*Portrayal of illegal drug use*
6.9.3	*Portrayal of illegal drug use with instructive detail*
6.10	*STROBING*
6.10.1	*No strobing*
6.10.2	*Strobing that could impact on those suffering from photosensitive epilepsy*

Tabla A.3: Clasificación de idiomas en TV-anytime

Código numérico	Idiomas
om	*(Afan) Oromo*
ab	*Abkhazian*
aa	*Afar*
af	*Afrikaans*
sq	*Albanian*
am	*Amharic*
ar	*Arabic*
hy	*Armenian*
as	*Assamese*
ay	*Aymara*
az	*Azerbaijani*
ba	*Bashkir*
eu	*Basque*
bn	*Bengali*

Código numérico	Idiomas
dz	*Bhutani*
bh	*Bihari*
bi	*Bislama*
br	*Breton*
bg	*Bulgarian*
my	*Burmese*
be	*Byelorussian*
km	*Cambodian*
ca	*Catalan*
zh	*Chinese*
co	*Corsican*
hr	*Croatian*
cs	*Czech*
da	*Danish*
nl	*Dutch*
en	*English*
eo	*Esperanto*
et	*Estonian*
fo	*Faeroese*
fj	*Fiji*
fi	*Finnish*
fr	*French*
fy	*Frisian*
gl	*Galician*
ka	*Georgian*
de	*German*
el	*Greek*
kl	*Greenlandic*
gn	*Guarani*
gu	*Gujarati*
ha	*Hausa*
he	*Hebrew*
hi	*Hindi*
hu	*Hungarian*
is	*Icelandic*
id	*Indonesian*
ia	*Interlingua*
ik	*Inupiak*
iu	*Eskimo*
ga	*Irish*
it	*Italian*
ja	*Japanese*
jw	*Javanese*
kn	*Kannada*

Código numérico	Idiomas
ks	*Kashmiri*
kk	*Kazakh*
rw	*Kinyarwanda*
ky	*Kirghiz*
rn	*Kirundi*
ko	*Korean*
ku	*Kurdish*
lo	*Laothian*
la	*Latin*
lv	*Latvian*
ln	*Lingala*
lt	*Lithuanian*
mk	*Macedonian*
mg	*Malagasy*
ms	*Malay*
ml	*Malayalam*
mt	*Maltese*
mi	*Maori*
mr	*Marathi*
mo	*Moldavian*
mn	*Mongolian*
na	*Nauru*
ne	*Nepali*
no	*Norwegian*
oc	*Occitan*
or	*Oriya*
ps	*Pashto, Pushto*
fa	*Persian*
pl	*Polish*
pt	*Portuguese*
pa	*Punjabi*
qu	*Quechua*
rm	*Rhaeto-Romance*
ro	*Romanian*
ru	*Russian*
sm	*Samoan*
sg	*Sangro*
sa	*Sanskrit*
gd	*Scots Gaelic*
sr	*Serbian*
sh	*Serbo-Croatian*
st	*Sesotho*
tn	*Setswana*
sn	*Shona*

Código numérico	Idiomas
sd	*Sindhi*
si	*Singhalese*
ss	*Siswati*
sk	*Slovak*
sl	*Slovenian*
so	*Somali*
es	*Spanish*
su	*Sudanese*
sw	*Swahili*
sv	*Swedish*
tl	*Tagalog*
tg	*Tajik*
ta	*Tamil*
tt	*Tatar*
te	*Tegulu*
th	*Thai*
bo	*Tibetan*
to	*Tigrinya*
to	*Tonga*
ts	*Tsonga*
tr	*Turkish*
tk	*Turkmen*
tw	*Twi*
ug	*Uigur*
uk	*Ukrainian*
ur	*Urdu*
uz	*Uzbek*
vi	*Vietnamese*
vo	*Volapuk*
cy	*Welsh*
wo	*Wolof*
xh	*Xhosa*
yi	*Yiddish*
yo	*Yoruba*
za	*Zhuang*
om	*Zulu*

Tabla A.4: Clasificación de los créditos que participan en los contenidos

Código	Créditos
V708	*Dubber*
V709	*Key character*
V106	*Key talent*
V43	*Participant*
V813	*Puppeteer*
V710	*Stunt*
V80	*Choreographer*
V484	*Costume designer*
V83	*Director of photography*
V714	*Fight Director*
V487	*Floor Manager*
V490	*Post-production editor*
V715	*Script supervisor*
V716	*Second assistant director*
V717	*Second unit director*
V718	*Sound designer*
V76	*Adaptor*
V2	*Scenario*
V94	*Treatment/Programme proposal*
V807	*Choir*
V810	*Librettist*
V811	*Lyricist*
V719	*Music arranger*
V809	*Music group*
V88	*Orchestra*
V103	*Announcer*
V720	*Causeur*
V32	*Commentor*
V483	*Correspondent*
V486	*Editor/Producer*
V30	*Editor-in-chief*
V31	*Editor-of-the-day*
V96	*Expert*
V97	*Interviewed guest*
V721	*News reader*
V117	*Witness*
V19	*Commissioning broadcaster*
V55	*Manufacturer*
V20	*Production company*
V22	*Production department*
V724	*Lighting technician*
V498	*Broadcast assistant*

Código	Créditos
V725	*Carpenter*
V727	*Dialogue coach*
V728	*Draughtsman*
V485	*Dresser*
V489	*Graphic designer*
V729	*Hairdresser*
V44	*Illustrator*
V730	*Leadman*
V496	*Scenic Operative*
V77	*Set dresser*
V82	*Visual editor*
V734	*Assistant visual editor*
V735	*Clapper loader*
V488	*Graphic assistant*
V740	*Grip*
V741	*Key grip*
V742	*Matte artist*
V45	*Photographer*
V743	*Pyrotechnician*
V494	*Rigger*
V744	*Second assistant camera*
V745	*Sound mixer*
V49	*Sound recordist*
V105	*Special effects*
V746	*Vision mixer*
V748	*Animal trainer*
V749	*Armourer*
V812	*Computer programmer*
V79	*Consultant*
V750	*Greensman*
V751	*Location manager*
V493	*Programme production researcher*
V495	*Runner*
V753	*Sign language*
V754	*Subtitles*
V95	*Translation*
V755	*Transportation manager*
V497	*Assistant producer*
V110	*Casting*
V491	*Production manager*
V492	*Production secretary*

Tabla A.5: Clasificación del tipo de audiencia objetivo de los contenidos

Código numérico	Tipo de audiencia
4.1	*GENERAL AUDIENCE*
4.2	*AGE GROUPS*
4.2.1	*Children*
4.2.1.1	*Age 4-7*
4.2.1.2	*Age 8-13*
4.2.1.3	*Age 14-15*
4.2.2	*Young adults*
4.2.2.1	*Age 16-17*
4.2.2.2	*Age 18-20*
4.2.2.3	*Age 21-24*
4.2.3	*Adults*
4.2.3.1	*Age 25-34*
4.2.3.2	*Age 35-44*
4.2.3.3	*Age 45-54*
4.2.3.4	*Age 55-64*
4.2.3.5	*Age 65+*
4.2.4	*All ages*
4.3	*SOCIAL GROUPS*
4.3.1	*Ethnic*
4.3.1.1	*Immigrant groups*
4.3.1.2	*Indiginous*
4.3.1.3	*Religious*
4.4	*OCCUPATIONAL GROUPS*
4.4.1	*AB*
4.4.1.1	*A (opinion former, judge, member of goverment, etc.)*
4.4.1.2	*B (Industry leader, doctors, lawyers, etc.)*
4.4.2	*C1C2*
4.4.2.1	*C1 ("White collar workers")*
4.4.2.2	*C2 (Skilled manual labourer)*
4.4.3	*DE*
4.4.3.1	*D (General manual labourer)*
4.4.3.2	*E (Unemployed + students)*
4.6	*GENDER*
4.6.1	*Primarily for males*
4.6.2	*Primarily for females*
4.6.3	*For males and females*
4.7	*GEOGRAPHICAL*
4.7.1	*Universal*
4.7.2	*Continental*

Código numérico	Tipo de audiencia
4.7.3	*National*
4.7.4	*Regional*
4.7.5	*Local*
4.8	*EDUCATIONAL LEVEL*
4.8.1	*Primary*
4.8.2	*Secondary*
4.8.3	*Tertiary*
4.8.4	*Post graduate*
4.9	*LIFESTYLE STAGES*
4.9.1	*Single*
4.9.2	*Couple*
4.9.3	*Family with children 0-3*
4.9.4	*Family with children 4-7*
4.9.5	*Family with children 8-15*
4.9.6	*Family with children 16+*
4.9.7	*Empty nester*
4.9.8	*Retired*

Tabla A.6: Clasificación relativa al origen técnico de los contenidos

Código numérico	Origen técnico
5.1	*STUDIO*
5.1.1	*Live*
5.1.2	*As live*
5.1.3	*Edited*
5.2	*MADE ON LOCATION*
5.2.1	*Live*
5.2.2	*As live*
5.2.3	*Edited*
5.3	*CINEMA INDUSTRY ORIGINATED*
5.4	*MADE ON FILM*
5.5	*HOME VIDEO*
5.6	*MULTIMEDIA FORMAT*

Tabla A.7: Clasificación relativa al formato estructural de los contenidos

Código numérico	Formato
2.1	*STRUCTURED*
2.1.1	*Bulletin*
2.1.2	*Magazine*
2.1.3	*Commented event*
2.1.4	*Documentary*
2.1.5	*Debate/Interview*
2.1.6	*Lecture/Speech/Presentation*
2.1.7	*Textual*
2.1.8	*Phone-in*
2.1.9	*DJ with discs*
2.2	*REPRESENTATION/PLAY*
2.2.1	*Fictional portrayal of life*
2.2.2	*Readings*
2.2.3	*Representation with puppets*
2.3	*CARTOON/ANIMATION*
2.3.1	*Anime*
2.3.2	*Computer*
2.3.3	*Cartoon*
2.4	*SHOW*
2.4.1	*Hosted show*
2.4.1.1	*Simple game show*
2.4.1.2	*Big game show*
2.4.2	*Panel-show*
2.4.2.1	*Simple game show*
2.4.2.2	*Big game show*
2.4.3	*Non-hosted show*
2.4.4	*Standup comedians*

Tabla A.8: Clasificación de las características emocionales asociadas a los contenidos

Código numérico	Atmósfera
8.1	*Alternative*
8.2	*Analytical*
8.3	*Astonishing*
8.4	*Ambitious*
8.5	*Black*
8.6	*Breathtaking*
8.7	*Chilling*
8.8	*Coarse*
8.9	*Compelling*
8.10	*Confrontational*

Código numérico	Atmósfera
8.11	*Contemporary*
8.12	*Crazy*
8.13	*Cutting edge*
8.14	*Eclectic*
8.15	*Edifying*
8.16	*Exciting*
8.17	*Fast-moving*
8.18	*Frantic*
8.19	*Fun*
8.20	*Gripping*
8.21	*Gritty*
8.22	*Gutsy*
8.23	*Happy*
8.24	*Heart-rending*
8.25	*Heart-warming*
8.26	*Hot*
8.27	*Humorous*
8.28	*Innovative*
8.29	*Insightful*
8.30	*Inspirational*
8.31	*Intriguing*
8.32	*Irreverent*
8.33	*Laid back*
8.34	*Outrageous*
8.35	*Peaceful*
8.36	*Powerful*
8.37	*Practical*
8.38	*Rollercoaster*
8.39	*Romantic*
8.40	*Rousing*
8.41	*Sad*
8.42	*Satirical*
8.43	*Serious*
8.44	*Sexy*
8.45	*Shocking*
8.46	*Silly*
8.47	*Spooky*
8.48	*Stunning*
8.49	*Stylish*
8.50	*Terrifying*

Tabla A.9: Clasificación de la temática de los contenidos comerciales

Código numérico	Temática
3.50	*COMMERCIAL/PRODUCTS*
3.50.1	*Agriculture, forestry and fishery products*
3.50.1.1	*Products of agriculture, horticulture and market gardening*
3.50.1.2	*Live animals and animal productos*
3.50.1.3	*Forestry and logging products*
3.50.1.4	*Fish and other fishing products*
3.50.2	*Ores and minerals, electricity, gas and water*
3.50.2.1	*Coal and lignite*
3.50.2.2	*Crude petroleum and natural gas*
3.50.2.3	*Uranium and thorium ores*
3.50.2.4	*Metal ores*
3.50.2.5	*Stone, sand and clay*
3.50.2.6	*Other minerals*
3.50.2.7	*Electricity, steam and hot water*
3.50.2.8	*Water*
3.50.3	*Food products, tobacco and textiles*
3.50.3.1	*Meat, fish, fruit, vegetables and oils*
3.50.3.2	*Dairy products*
3.50.3.3	*Grain mill products and starches*
3.50.3.4	*Beverages*
3.50.3.5	*Tobacco products*
3.50.3.6	*Yarn and thread*
3.50.3.7	*Textile articles other than apparel*
3.50.3.8	*Knitted or crocheted fabrics*
3.50.3.9	*Leather products*
3.50.4	*Other transportable goods, except metal products and machinery*
3.50.4.1	*Wood, cork, straw and plaiting materials*
3.50.4.2	*Pulp and paper products*
3.50.4.3	*Coke oven products*
3.50.4.4	*Basic chemicals*
3.50.4.5	*Man-made fibres*
3.50.4.6	*Rubber and plastic products*
3.50.4.7	*Glass products*
3.50.4.8	*Furniture*
3.50.5	*Metal products, machinery and equipment*
3.50.5.1	*Basic metals*
3.50.5.2	*Fabricated metal products*
3.50.5.3	*General purpose machinery*
3.50.5.4	*Special purpose machinery*

Código numérico	Temática
3.50.5.5	*Office and computing machinery*
3.50.5.6	*Electrical machinery*
3.50.5.7	*Radio, television and communication equipment*
3.50.5.8	*Medical appliances and optical instruments*
3.50.5.9	*Transport equipment*
3.50.6	*Intangible assets, land and construction services*
3.50.6.1	*Intangible assets*
3.50.6.2	*Land*
3.50.6.3	*Constructions*
3.50.6.4	*Construction services*
3.50.7	*Distributive trade services, lodgin and transport services*
3.50.7.1	*Wholesade trade services*
3.50.7.2	*Retail trade services*
3.50.7.3	*Lodging*
3.50.7.4	*Land transport services*
3.50.7.5	*Water transport services*
3.50.7.6	*Air transport services*
3.50.7.7	*Supporting and auxiliary transport services*
3.50.7.8	*Postal and courier services*
3.50.7.9	*Electricity, gas and water distribution services*
3.50.8	*Financial, rental and leasing services*
3.50.8.1	*Insurance services*
3.50.8.2	*Real estate services*
3.50.8.3	*Leasing or rental services*
3.50.9	*Business and production services*
3.50.9.1	*Research and development services*
3.50.9.2	*Proffesional, scientific and technical services*
3.50.9.3	*Other professional services*
3.50.9.4	*Telecommunications services*
3.50.9.5	*Support services*
3.50.9.6	*Production services*
3.50.9.7	*Repair services*
3.50.10	*Community, social and personal services*
3.50.10.1	*Public administration*
3.50.10.2	*Educational services*
3.50.10.3	*Health and social services*
3.50.10.4	*Protection services*
3.50.10.5	*Services of membership organizations*
3.50.10.6	*Recreational, cultural and sporting services*
3.50.10.7	*Domestic services*

Tabla A.10: Clasificación relativa a la intención de los contenidos

Código numérico	Intención
1.1	*ENTERTAINMENT*
1.1.1	*Pure entertainment*
1.1.2	*Informative entertainment*
1.2	*INFORMATION*
1.2.1	*Goverment*
1.2.2	*Pure information*
1.2.3	*Infotainment*
1.2.4	*Advice*
1.3	*EDUCATION*
1.3.1	*School programmes*
1.3.1.1	*Primary*
1.3.1.2	*Secondary*
1.3.2.3	*Tertiary*
1.3.2.4	*Further education*
1.4	*PROMOTIONAL*
1.5	*ADVERTISING*
1.6	*RETAIL*
1.7	*FUND-RAISING*
1.8	*ENRICHMENT*
1.8.1	*General enrichment*
1.8.2	*Inspirational enrichment*

Apéndice B

Implementación

Este apéndice describe los detalles de implementación de la metodología de razonamiento semántico desarrollada en la tesis, y presenta las herramientas y tecnologías de libre disposición utilizadas en la evaluación experimental.

B.1 Entorno de ejecución

Para poder validar el trabajo presentado en la tesis, hemos implementado una herramienta que incluye los diferentes componentes de la metodología de razonamiento semántico desarrollada. En el entorno de ejecución del sistema AVATAR hemos omitido el receptor del usuario, debido a que las funciones que desempeña (presentación de recomendaciones y monitorización del espectador) no requieren validación experimental.

Una vez desarrollada la ontología OWL de TV, la herramienta implementada permite:

- modelar las preferencias de los usuarios a partir del conocimiento formalizado en la misma,
- calcular los niveles de interés y olvido asociados a dichas preferencias,
- ejecutar las dos estrategias de personalización propuestas,
- proporcionar realimentación de relevancia sobre las recomendaciones ofrecidas, y
- actualizar convenientemente los perfiles modelados en la herramienta.

En este apéndice, describimos tanto las herramientas utilizadas en la implementación como la evaluación experimental de la metodología. Si bien los escenarios propuestos en estos experimentos son suficientemente realistas, nuestro trabajo también incorpora algunas medidas para favorecer la escalabilidad (y viabilidad computacional) de la metodología, a medida que aumenta el número de usuarios y contenidos de TV en el sistema.

B.2 Herramientas de soporte utilizadas

Para desarrollar la ontología requerida en la metodología hemos utilizado el editor Protégé-OWL [197], una herramienta gratuita, de código abierto, que facilita la creación, visualización y consulta de bases de conocimiento en formato OWL. Su arquitectura abierta y modular —basada en la incorporación de numerosos *plugins* que permiten extender sus funcionalidades básicas— junto a la API Java que proporciona, convierten a esta herramienta en un candidato excelente para la implementación de la metodología.

La API Protégé-OWL[1] es una librería Java de código abierto que, además de proporcionar clases y métodos para crear, modificar y consultar modelos OWL, permite realizar inferencias mediante razonadores lógicos (razonadores DL). Nuestra herramienta, implementada como un *plugin* de Protégé, utiliza este tipo de facilidades para: (i) añadir clases y propiedades a la ontología OWL, (ii) instanciar la base de conocimiento incorporando programas y descripciones semánticas específicas, y (iii) comprobar la consistencia de la ontología y la correcta clasificación de sus instancias mediante los razonadores DL.

Asimismo, la API está optimizada para la implementación de interfaces gráficas. Gracias a estas facilidades, nuestra herramienta implementa diferentes interfaces para: (i) inicializar y actualizar los perfiles de los usuarios, (ii) presentar las recomendaciones, y (iii) desarrollar herramientas auxiliares que permiten explorar los razonamientos semánticos aplicados sobre la base de conocimiento OWL.

B.2.1 Implementación de la ontología de TV

De acuerdo a lo explicado en el capítulo 5, las clases y propiedades incluidas en nuestra ontología OWL son extraídas a partir de los metadatos TV-Anytime, definidos en formato XML. Para automatizar este proceso hemos recurrido a JAXP[2] (*Java API for XML Processing*), una librería que, gracias a un conjunto de analizadores sintácticos (*parsers*) DOM[3], permite construir un árbol a partir de cada fichero XML y acceder a su estructura jerárquica, cómoda y rápidamente, mediante un conjunto de métodos predefinidos.

Junto a este conocimiento, nuestra formalización debe importar las ontologías GEO y OWL-Time para poder asociar a los contenidos información geográfica y temporal. Para ello, incluimos desde el editor Protégé-OWL los espacios de nombres (*namespaces*) asociados a cada una de ellas, junto a los prefijos utilizados para referenciar sus clases y propiedades desde la ontología de TV.

B.2.2 Instanciación de la base de conocimiento

Una vez definidas las clases y propiedades del modelo OWL, es necesario incluir en el mismo las instancias específicas que identifican tanto contenidos de TV como atributos semánticos. Algunos de estos atributos se extraen de la propia especificación TV-Anytime, de ahí que el procedimiento descrito en la sección previa siga siendo válido para el proceso de instanciación.

[1] Disponible en *http://protege.stanford.edu/plugins/owl/api/*
[2] Disponible en *http://java.sun.com/webservices/jaxp/*
[3] Disponible en *http://www.w3.org/DOM/*

Por el contrario, para incorporar programas de TV y descripciones semánticas de los mismos, hemos implementado un segundo *plugin* Protégé, que extrae esta información desde tres bases de datos:

- IMDB (*Internet Movie DataBase*[4]) y AMG (*All Movie Guide*[5]): A pesar de sus nombres, IMDB y AMG son dos bases de datos que no sólo contienen descripciones de películas sino también de una gran variedad de contenidos de TV. Tales descripciones se estructuran en forma de registros y aportan información de diversa naturaleza: título de los programas, sinopsis, créditos, alertas de contenido, clasificación por edades, palabras claves (geográficas, cronológicas y temáticas), géneros, país y fecha de estreno, premios, contenidos relacionados (por participar en ellos los mismos créditos o tratar una misma temática), etc.

  Dado que estas descripciones también están contempladas en la especificación TV-Anytime, los contenidos de ambas bases de datos se traducen fácilmente a la ontología OWL. Para ello, hemos desarrollado un procedimiento automático que analiza (sintácticamente) la información HTML disponible en IMDB y AMG, y la introduce en el modelo OWL mediante los métodos de la API Protégé-OWL.

  Tal como apuntábamos en el capítulo 5, las palabras clave y sinopsis de los contenidos requieren un procesamiento adicional, orientado a identificar instancias específicas tanto de las clases definidas en las ontologías GEO y OWL-Time utilizadas en la metodología, como de la clase genérica *Topics* incluida en la misma.

  - Para analizar las sinopsis de los programas, hemos recurrido a mecanismos semiautomáticos en los que:
    - en primer lugar, se elimina la información irrelevante para el proceso de instanciación (artículos, verbos, terminaciones comunes, etc.);
    - a continuación, se clasifican (manualmente) los términos resultantes, definiéndolos como instancias concretas en la ontología que corresponda;
    - por último, se completa el conocimiento asociado a dichas instancias (por ejemplo, si una instancia identifica un intervalo temporal, es necesario definir su duración en la ontología OWL-Time).
  - El proceso de instanciación a partir de las palabras clave es inmediato, gracias a su organización en tres categorías genéricas que identifican, precisamente, información temática, geográfica y temporal.

- BBC (*British Broadcasting Corporation*): Además de los contenidos descritos en IMDB y AMG, también hemos incluido en la ontología OWL las descripciones TV-Anytime accesibles desde el servidor Web de la BBC. Para traducir los programas disponibles al modelo OWL, recurrimos de nuevo a los *parsers* implementados en JAXP y a la API Protégé, tal como se describió en la Sección B.2.1.

Una vez implementada la ontología OWL, la herramienta desarrollada aprovecha las capacidades de los razonadores DL para asegurar la consistencia de la base de conocimiento utilizada en la metodología.

[4]Disponible en *http://www.imdb.com/*
[5]Disponible en *http://allmovie.com/*

B.2.3 Inferencia mediante razonadores lógicos

Si bien los razonadores DL disponibles en la Web Semántica no son capaces de descubrir las asociaciones semánticas consideradas en nuestro trabajo, sus capacidades lógicas son útiles para realizar inferencias sencillas sobre ontologías OWL DL, aprovechables en procesos de razonamiento más complejos.

Tal como adelantábamos en la Sección B.2, la API Protégé-OWL permite realizar este tipo de razonamientos lógicos. Para ello, la librería soporta el lenguaje DIG (*DL Implementation Group*), una especificación que describe cómo debe establecerse la comunicación entre una aplicación y un razonador DL externo.

Los razonadores DL conformes a DIG son accesibles a través del protocolo HTTP y proporcionan una interfaz basada en este lenguaje. DIG incluye representaciones (en formato XML) de las clases, propiedades e instancias definidas en la ontología, así como axiomas para identificar relaciones jerárquicas, de equivalencia o disyunción entre clases. En definitiva, este lenguaje incorpora constructores para describir la ontología y también para preguntar al razonador la información descubierta durante sus inferencias (e.g. nuevas subclases en la ontología, relaciones de pertenencia entre clases e instancias, nuevas propiedades asociadas a instancias, etc.).

Una de las ventajas de DIG es que las aplicaciones pueden comunicarse con el razonador sin necesidad de conocer sus detalles específicos ni los protocolos de interacción utilizados. Esto permite que Protégé-OWL pueda utilizar cualquier razonador DIG para realizar las inferencias lógicas mencionadas previamente. Por sus prestaciones y su implementación altamente optimizada, nuestra herramienta experimental incorpora el razonador (DIG) externo RACER. Para utilizar esta herramienta mediante los métodos disponibles en la API de razonamiento de Protégé-OWL, se han seguido los siguientes pasos:

- En primer lugar, definimos la URL del razonador para poder establecer la comunicación mediante HTTP.
- A continuación, creamos una instancia del razonador y la inicializamos con el modelo OWL creado en las secciones previas.
- Finalmente, utilizamos los métodos de la API para acceder a la información inferida por RACER. Esta información no sólo permite comprobar la consistencia de la ontología, sino que también se utiliza para completar el modelo OWL inicialmente especificado. En este último proceso, se incluyen los razonamientos geográficos y temporales apuntados en el capítulo 5.

B.2.4 Inspección de los razonamientos semánticos aplicados

Para facilitar la evaluación experimental, hemos implementado una herramienta auxiliar (incluida también como un *plugin* de Protégé) que permite trazar en la ontología OWL los razonamientos aplicados en el sistema, explicando las asociaciones semánticas en las que se basan sus recomendaciones personalizadas.

Para implementar esta herramienta, aprovechamos las facilidades de visualización de ontologías ofrecidas en Protégé. En concreto, este editor proporciona hasta tres *plugins* de visualización diferentes:

- Onto Viz[6]: Está basado en el software Graphviz (*Graph Visualization Software*) y permite visualizar a la vez diferentes grafos no conectados entre sí. Estos últimos son estáticos y no interactivos, lo que convierte a Onto Viz en una herramienta poco apropiada para navegar y visualizar redes de gran tamaño como las manejadas en nuestra metodología.

- Jambalaya[7]: Esta herramienta utiliza la tecnología SHriMP (*Simple, Hierarchical, Multi-Perspective Views*) para mostrar diferentes vistas de grafos anidados obtenidos desde la estructura de la ontología. Entre sus capacidades se incluyen tanto la visualización como la adquisición de conocimiento, de ahí que su funcionalidad resulte excesiva para nuestros propósitos de implementación.

- TGVizTab[8]: Este *plugin* está basado en la tecnología TouchGraph, un entorno Java de código abierto que permite crear y navegar a través de grafos genéricos de forma dinámica e interactiva. En concreto, los nodos del grafo son las clases e instancias de la ontología, mientras que los enlaces del mismo identifican las propiedades y relaciones establecidas entre ellos. Para facilitar una exploración clara e intuitiva para el usuario, el *plugin* incorpora opciones de configuración para controlar el conocimiento representado en el grafo (e.g. mostrar sólo algunas de las propiedades de una instancia o clase determinada, representar subgrafos aislados dando la opción de presentar u ocultar los enlaces que permiten relacionarlos, etc.).

  La flexibilidad que introducen estas opciones (accesibles a través de la API Protégé-OWL), junto a la capacidad de visualizar la ontología de forma dinámica e interactiva, convierten a TGVizTab en el *plugin* más adecuado para nuestra implementación. Las primeras permiten presentar un grafo en el que únicamente se representan las preferencias del usuario y el conjunto de contenidos con los que están asociadas semánticamente. Al interactuar con este grafo y seleccionar sus nodos, la herramienta muestra aquellas propiedades y relaciones jerárquicas de la ontología que originan las asociaciones implícitas visualizadas en el *plugin*.

  Esta información es útil en nuestra evaluación experimental porque permite entender los razonamientos semánticos utilizados en la metodología y, además, valorar la naturaleza de las asociaciones inferidas, advirtiendo si son o no suficientemente significativas para el proceso de personalización.

B.3 Evaluación experimental

En los experimentos realizados se modelaron las preferencias de varios usuarios ficticios y se simuló su interacción con la herramienta implementada (aceptando algunos de los programas sugeridos y rechazando otros), para así poder estudiar el efecto de su realimentación de relevancia en los razonamientos semánticos aplicados en la metodología. En concreto, hemos modelado hasta 75 perfiles en la herramienta y hemos incluido hasta 520 contenidos en la ontología OWL de TV.

Mediante la herramienta implementada no sólo se han elegido los valores iniciales de todos los parámetros utilizados en la metodología, sino que también se ha validado

[6]Disponible en *http://protege.cim3.net/cgi-bin/wiki.pl?OntoViz*
[7]Disponible en *http://www.thechiselgroup.org/∼chisel/projects/jambalaya/jambalaya.html*
[8]Disponible en *http://www.ecs.soton.ac.uk/∼ha/TGVizTab/*

el proceso de ajuste dinámico (propuesto en la mayoría de ellos) para establecer su valor a lo largo del tiempo.

Nuestra evaluación se ha centrado en tres aspectos fundamentales:

- En primer lugar, analizamos la técnica de modelado de usuarios definida en el trabajo, validando en los experimentos tanto el proceso de adaptación de las recomendaciones a las preferencias de los usuarios, como el mecanismo propuesto para olvidar intereses obsoletos e identificar los actuales.
- Tras las pruebas realizadas, hemos comprobado que las asociaciones utilizadas en el proceso de recomendación conducen a contenidos que, presumiblemente, interesarán al usuario, ya que permiten establecer relaciones semánticas útiles entre estos programas y la mayoría de sus preferencias. La naturaleza de estas asociaciones ha sido evaluada con ayuda de la herramienta que traza los razonamientos semánticos utilizados en las estrategias de personalización desarrolladas.
- Gracias a esta herramienta, también hemos podido validar la hipótesis de que el razonamiento semántico descubre contenidos que los enfoques actuales (métodos basados en contenido y filtrado colaborativo) omitirían.
    - En concreto, algunos de estos contenidos nunca serían seleccionados mediante el filtrado tradicional basado en contenido, dado que, en lugar de compartir los atributos semánticos definidos en el perfil del usuario, estos programas estaban (semánticamente) asociados con sus preferencias.
    - Por otra parte, los contenidos descubiertos en nuestra metodología también pasarían desapercibidos en el filtrado colaborativo, dado que, en algunos casos, no figuraban en ninguno de los perfiles de los usuarios modelados en la herramienta recomendadora implementada.

Si bien las pruebas realizadas revelan que los costes computaciones asociados a la metodología de razonamiento son asumibles en un entorno de aplicación como el propuesto, nuestro trabajo también incorpora diferentes mecanismos (resumidos en la Sección 9.1.5) para garantizar la viabilidad computacional del enfoque en escenarios más complejos (con un número mayor de usuarios y/o contenidos).

Bibliografía

[1] Adomavicius G. y Tuzhilin A. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):739–749, 2005.

[2] Agrawal R. y Srikant R. Fast Algorithms for Mining Association Rules. En *20th International Conference on Very Large Data Bases (VLDB-94)*, págs. 487–499, 1994.

[3] Aleman-Meza B., Halaschek-Wiener C., Sanket Sahoo S., Sheth A. y Budak Arpinar I. Template based Semantic Similarity for Security Application. En *Technical Report LSDIS Lab, Computer Science Department (University of Georgia)*, 2005.

[4] Amazon. *Disponible en http://www.amazon.com*. 2001.

[5] Anyanwu K. y Sheth A. ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. En *12th International World Wide Web Conference (WWW-03)*, págs. 115–125, 2003.

[6] Anyanwu K., Maduko A. y Sheth A. SemRank: Ranking Complex Relationship Search Results on the Semantic Web. En *14th International World Wide Web Conference (WWW-05)*, págs. 117–127, 2005.

[7] Arbib M. A. *The handbook of brain theory and neural networks*. Cambridge, 1995.

[8] Ardissono L., Gena C., Torasso P., Bellifemine F., Difino A. y Negro B. User modeling and recommendation techniques for personalized electronic program guides. En *Personalized Digital Television: Targeting Programs to Individual Viewers*, págs. 3 – 26. Kluwer Academics Publishers, 2004.

[9] Armstrong R., Freitag D., Joachims T. y Mitchell T. WebWatcher: A Learning Apprentice for the World Wide Web. En *AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments*, págs. 6–12, 1995.

[10] Asnicar F. y Tasso C. IfWeb: A Prototype of User-Models-Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web. En *6th International Conference on User Modeling (UM-97)*, págs. 14–24, 1997.

[11] Baader F. y Hollunder B. Kris: Knowledge representation and inference system. En *Special Issue on Implemented Knowledge Representation and Reasoning Systems, SIGART Bulletin 2*, págs. 8–14, 1991.

[12] Baader F., Calvanese D., McGuinness D., Nardi D. y Schneider P. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[13] Bailey J., Bry F., Furche T. y Schaffert S. Web and Semantic Web Query Languages: A Survey. En *1st International Summer School*, págs. 35–133. Springer, 2005.

[14] Balabanovic M. y Shoham Y. Combining Content-Based and Collaborative Recommendation. *Communications of the ACM*, 40(3):1–9, 1997.

[15] Bartfai G. Hierarchical Clustering with ART Neural Networks. En *IEEE World Conference on Computational Intelligence (WCCI-94)*, págs. 940–944, 1994.

[16] Basu C., Hirsh H. y Cohen W. Recommendation as Classification: Using Social and Content-based Information in Recommendation. En *15th National Conference on Artificial Intelligence (AAAI-98)*, págs. 714–720, 1998.

[17] Baudisch P. y Bruekner L. TV Scout: guiding users from printed TV program guides to personalized TV recommendation. En *1st Workshop on Personalization in Future TV*, págs. 151 – 160, 2001.

[18] Beckett D. RDF Syntax Specification. Disponible en *http://www.w3.org/TR/rdf-syntax-grammar*. 2004.

[19] Berners-Lee T., Hendler J. y Lassila O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *The Scientific American*, 279(5):34–43, 2001.

[20] Berry M. Computational Information Retrieval. En *Proceedings in Applied Mathematics*, págs. 185–205, 2000.

[21] Berry M., Dumai S. y Shippy A. A case study of latent semantic indexing. En *Technical Report CS-95-271, U.T. Knoxville*, 1995.

[22] Billsus D. y Pazzani M. Learning Collaborative Information Filters. En *15th International Conference on Machine Learning (ICML-98)*, págs. 46–54, 1998.

[23] Billsus D. y Pazzani M. A personal news agent that talks, learns and explains. En *3rd International Conference on Autonomous Agents (Agents-99)*, págs. 268–275, 1999.

[24] Björkman M., Aroyo L, Bellekens M., Dekker T., Loef E. y Pulles R. Personalised Home Media Centre Using Semantically Enriched TV-Anytime Content. En *4th European Conference on Interactive Television (EuroITV-06)*, págs. 195–173, 2006.

[25] Blanco Fernández Y., Pazos Arias J. J., Gil Solla A., Ramos Cabrer M., López Nores M. y Barragáns Martínez B. AVATAR: Modeling Users by Dynamic Ontologies in a TV Recommender System based on Semantic Reasoning. En *3rd European Conference on Interactive Television (EuroITV-05)*, págs. 173–183, 2005.

[26] Blanco Fernández Y., Pazos Arias J. J., Gil Solla A., Ramos Cabrer M., López Nores M. y Bermejo Múñoz J. AVATAR: Enhancing the Personalized Television by Semantic Inference. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) - Special issue on Personalization Techniques for Recommender Systems and Intelligent User Interfaces*. En prensa.

[27] Blanco Fernández Y., Pazos Arias J. J., Gil Solla A. y Ramos Cabrer M. AVATAR: A Flexible Approach to Improve the Personalized TV by Semantic Inference. En *1st Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSIUI-05)*, 2005.

[28] Blanco Fernández Y., Pazos Arias J. J., López Nores M., Gil Solla, A. and Ramos Cabrer M. AVATAR: An improved solution for Personalized TV based on Semantic Inference. En *IEEE International Conference on Consumer Electronics (ICCE-06)*, págs. 145–146, 2006.

[29] Blanco Fernández Y., Pazos Arias J. J., López Nores M., Gil Solla A. y Ramos Cabrer M. AVATAR: An improved solution for personalized TV based on semantic inference. *IEEE Transactions on Consumer Electronics*, 52(1):223–232, 2006.

[30] Boone G. Concept Features in RE:Agent, an Intelligent Email Agent. En *2nd International Conference on Autonomous Agents (Agents-98))*, págs. 141–148, 1998.

[31] Bousquet C., Jualent M. C., Chatelier G. y Degoulet P. Using semantic distance for the efficient coding of medical concepts. En *The American Medical Informatics Association 2000 Symposium (AMIA-2000)*, págs. 96–100, 2000.

[32] Brachman R.J. y Schmolze G. An overview of the KL-ONE Knowledge Representation system. *Cognitive Science*, 9(2):171–216, 1985.

[33] Bradley K. y Smyth B. Improving recommendation diversity. En *12th National Conference in Articial Intelligence and Cognitive Science (AICS-01)*, págs. 75–84, 2001.

[34] Bray T., Paoli J. y Sperberg-McQueen C. *Extensible Markup Language (XML). World Wide Web Consortium (W3C). Disponible en http://www.w3.org/TR/1998/REC-xml-19980210.html*.

[35] Breese J., Heckerman D. y Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. En *14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, págs. 43–52, 1998.

[36] Brickley D. y Guha R. RDF vocabulary description language 1.0: RDF Schema. Disponible en *http://www.w3.org/TR/rdf-schema*. 2004.

[37] Brin S. y Page L. The anatomy of a large-scale hypertextual Web search engine. En *7th International World Wide Web Conference (WWW-98)*, págs. 107–117, 1998.

[38] Broadcast and On-line Services: Search, select, and rightful use of content on personal storage systems. *ETSI TS 102 822 series*. 2003.

[39] Broekstra J., Kampman A. y van Harmelen F. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. En *1st International Semantic Web Conference (ISWC-02)*, págs. 54–69, 2002.

[40] Buczak A., Kurapati K. y Zimmerman J. Personalization: Improving Ease-of-Use, Trust and Accuracy of a TV Show Recommender. En *2nd Workshop on Personalization in Future TV (TV-02)*, págs. 9–17, 2002.

[41] Budanitsky A. y Hirst G. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. En *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, págs. 18–25, 2001.

[42] Burke R. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[43] Cassel L., Banerjee S., Narayana A. y Wolz U. The Web Host Access Tool (WHAT): Exploring Context Sensitive Information Access. En *1st Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSIUI-05)*, págs. 34–43, 2005.

[44] CDNow. *Disponible en http://www.cdnow.com.* 2001.

[45] Chen H. Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. *Journal of the american society for information science*, 46(3):196–216, 1995.

[46] Chen H. y Lynch K. J. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(1):885–902, 1992.

[47] Chen H., Lynch K. J., BAsu K. y Ng T. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Experts (special series on Artificial Intelligence in Text-based Information Systems)*, 8(1):25–34, 1993.

[48] Chen L. y Sycara K. WebMate: A Personal Agent for Browsing and Searching. En *2nd International Conference on Autonomous Agents (AGENTS-98)*, págs. 132–139. ACM Press, 1998.

[49] Chorianopoulos K. y Spinellis D. User Interface Development for Interactive Television: Extending a Commercial DTV Platform to the Virtual Channel API. *Computers and Graphics*, 28(2):157–166, 2004.

[50] Christakou C. y Stafylopatis A. A Hybrid Movie Recommender System Based on Neural Networks. En *5th International Conference on Intelligent Systems Design and Applications (ISDA-05)*, págs. 500–505, 2005.

[51] Cohen S., Mamou J., Kanza Y. y Sagiv Y. XSEarch: A Semantic Search Engine for XML. En *29th International Conference on Very Large Data Bases (VLDB-03)*, págs. 45–56, 2003.

[52] Colucci S., Di Noia T., Di Sciascio E., Donini F., Mongiello M. y Mottola M. A Formal Approach to Ontology-based Semantic Match of Skills Descriptions. *Journal of Universal Computer Science*, 9(12):1437–1454, 2003.

[53] Cooley R., Tan P. N. y Srivastava J. WebSift: The Web Site Information Filter System. En *KDD Workshop on Web Mining*, págs. 24–32, 1999.

[54] CORESE. Disponible en *http://www-sop.inria.fr/acacia/soft/corese/*.

[55] Cotter P. y Smyth B. A personalized TV listing service. *Communications of the ACM*, 43(8):107–111, 2000.

[56] Crestani F. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[57] Cunningham P., Bergmann R., Schmitt S., Traphöner R., Breen S. y Smyth B. WebSell: Intelligent Sales Assistants for the World Wide Web. En *eBusinees and eWork Conference and Exhibition (E-01)*, págs. 31–41, 2001.

[58] DAML: The DARPA Agent Markup Language. *Disponible en http://www.daml.org*. 2000.

[59] DAML+OIL Revised Language Specification. *Disponible en http://www.daml.org/2001/03/daml+oil-index*. 2001.

[60] Davies J., Duke A. y Stonkus A. OntoShare: Using Ontologies for Knowledge Sharing. En *11th International WWW Conference (WWW-02)*, págs. 42–53, 2002.

[61] Davies J., Fensel D. y van Harmelen F. *Towards the Semantic Web: Ontology-driven Knowledge Management*. John Wiley & Sons, 2003.

[62] Davies J., Weeks R. y Krohn U. QuizRDF: Search Technology for the Semantic Web. En *11th International World Wide Web Conference (WWW-02)*, págs. 24–32, 2002.

[63] Demiriz A. An association mining-based product recommender. En *INFORMS Miami 2001 Annual Meeting Cluster: Data Mining*, págs. 34 – 44, 2001.

[64] Diestel R. *Graph Theory*. Springer-Verlag, 2000.

[65] Digital Video Broadcasting Consortium. The Multimedia Home Platform specification. ETSI standard TS 102 812 v1.2.1, 2003.

[66] Ding L., Finin T., Joshi A., Peng Y., Scott R., Sachs J., Pan R., Reddivari P. y Doshi V. Swoogle: A Semantic Web Search and Metadata Engine. En *ACM Conference on Information and Knowledge Management (CIKM-04)*, págs. 272–281, 2004.

[67] Dolog P., Henze N., Nejdl W. y Sintek M. Towards the Adaptive Semantic Web. En *International Workshop on Principles and Practice of Semantic Web Reasoning*, págs. 1–12, 2003.

[68] dos Santos J., Goularte R., Faria G. y Moreira E. Modeling of User Interaction in Context-aware Interactive Television Applications on Distributed Environments. En *1st Workshop on Personalization in Future TV (TV-01)*, págs. 42–51, 2001.

[69] Efron M. y Geisler G. Is it all About Connections? Factors Affecting the Performance of a Link-Based Recommender System. En *SIGIR Workshop on Recommender Systems*, págs. 34–43, 2001.

[70] Excite: Information Retrieval Technology and Intelligent Concept Extraction Searching. *Disponible en http://www.excite.com*.

[71] Fensel D., van Harmelen F., Horrocks I., McGuinness D. y Patel-Schneider P. SOIL: An ontology infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2):38–45, 2001.

[72] Fierro R. D. y Berry M. W. *Efficient Computation of the Riemannian SVD in TLS Problems in Information Retrieval*, págs. 349–360. Kluwer Academic Publishers, 2002.

[73] Foltz P. Using latent semantic indexing for information filtering. En *ACM SIGOIS and IEEE CS TC-OA Conference on Office information systems*, págs. 40–47, 1990.

[74] Froogle. *Disponible en http://froogle.google.com*.

[75] Fukushima K. Cognitron: A Self-Organizing Multilayered Neural Network. *Biological Cybernetics*, 20(1):121–136, 1975.

[76] Funakoshi K. y Ohguro T. Evaluation of integrated content-based collaborative filtering. En *ACM SIGIR Workshop on Recommender Systems*, págs. 42–51, 2001.

[77] Ganesan P., Garcia-Molina H. y Widom J. Exploiting Hierarchical Domain Structure to Compute Similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.

[78] Giunchiglia F., Shvaiko P. y Yatskevich M. S-Match: an Algorithm and an Implementation of Semantic Matching. En *1st European Semantic Web Conference (ESWC-04)*, págs. 61–75, 2004.

[79] Goldberg D., Nichols D., Oki B. y Terry D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[80] Good N., Schafer J., Konstan J., Borchers A., Sarwar B., Herlocker J. and Riedl J. Combining Collaborative Filtering with Personal Agents for Better Recommendations. En *17th National Conference on Artificial Intelligence (AAAI-99)*, págs. 439–446, 1999.

[81] Google. *Disponible en http://www.google.es*.

[82] Gruber T. R. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[83] Guha R., McCool R. y Miller E. Semantic Search. En *12th International World Wide Web Conference (WWW-03)*, págs. 700–709, 2003.

[84] Gujjar D., Thawani A., Gopalan S. y Sridhar V. Path Analysis Based Adaptive Navigation for iTV Applications. En *3rd European Conference on Interactive Television (EuroITV-05)*, págs. 183–191, 2005.

[85] Haarslev V. y Möller R. Racer system description. En *1st International Joint Conference on Automated Reasoning (IJCAR-01)*, págs. 701–706, 2001.

[86] Haase P., Ehrig M., Hotho A. y Schnizler B. Personalized Information Access in a Bibliographic Peer-to-Peer System. En *AAAI Workshop on Semantic Web Personalization*, págs. 1–12, 2004.

[87] Halascheck C. *A Flexible Approach for Ranking Complex Relationships on the Semantic Web*. PhD thesis, Department of Computer Science. University of Georgia, 2004.

[88] Hampshire J. B. y Perlmutter B. A. Equivalence proofs for multilayer perceptron classifiers and the Bayesian discriminant function. En *Connectionist Models Summer School*, págs. 159–173, 1990.

[89] Han S. y Karypis G. Feature-based recommendation system. En *14th International Conference on Information and Knowledge Management (CIKM-05)*, págs. 446 – 452, 2005.

[90] Haveliwala T. Efficient computation of PageRank. En *Technical Report 1999-31*, 1999.

[91] Hayes C. y Cunningham P. Smart Radio - a Proposal. En *Technical Report, Trinity College Dublin, Computer Science, TCD-CS-1999-24*, 1999.

[92] Hayes C. y Cunningham P. Smart radio - community based music radio. *Knowledge-Based Systems*, 14(3):197–201, 2001.

[93] Herlocker J., Konstant J., Borchers A. y Riedl J. An Algorithmic Framework for Performing Collaborative Filtering. En *International Conference on Research and Development in Information Retrieval (SIGIR-99)*, págs. 230–237, 1999.

[94] Hill W., Stead L., Rosenstein M. y Furnas G. Recommending and Evaluating Choices in a Virtual Community of Use. En *International Conference on Human Factors in Computing Systems (CHI-95)*, págs. 194–201, 1995.

[95] Hinton G. E. y Sejnowski T. J. Learning and relearning in Boltzmann Machines. En *Parallel Distributed Processing*, págs. 282–317. MIT Press, 1986.

[96] Hobbs J. R. y Pan F. *Time Ontology in OWL. Disponible en http://www.w3.org/2001/sw/BestPractices/OEP/Time-Ontology*. 2006.

[97] Holger H. *The Topic Maps Handbook. White paper*. 2003.

[98] Hopfield J. J. Neural network and physical systems with emergent collective computational abilities. *National Academy of Science*, 79(4):2254–2258, 1982.

[99] Hornik K., Stinchcombe M. y White H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[100] Horrocks I. Using an expressive description logic: FaCT or fiction? En *6th International Conference on Principles of Knowledge Representation and Reasoning (KR-98)*, págs. 636–645, 1998.

[101] Hristidis V., Papakonstantinou Y. y Balmin A. Keyword proximity search on XML graphs. En *19th International Conference on Data Engineering (ICDE-03)*, págs. 367–378, 2003.

[102] Huang Z., Chen H. y Zeng D. Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering. *ACM Transactions on Information Systems*, 22(1):116–142, 2004.

[103] Huberman B. y Kaminsky M. Beehive: A System for Cooperative Filtering and Sharing of Information. En *Technical Report, Dynamics of Computation Group*, 1996.

[104] InferEd. Disponible en *http://www.intellidimension.com/*.

[105] Jena. Disponible en *http://jena.sourceforge.net/*.

[106] Jennings A. y Higuchi H. A User Model Neural Network for a Personal News Service. *User Modeling and User-Adapted Interaction*, 3(1):1–25, 1993.

[107] Jensen F. V. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[108] Jiang J. J. y Conranth D. W. Semantic Similarity based on corpus statistics and lexical taxonomy. En *International Conference on research in Computational Linguistics (ROCLING-97)*, págs. 19–33, 1997.

[109] Joachims T., Freitag D. y Mitchell T. WebWatcher: A Tour Guide for the World Wide Web. En *16th International Joint Conference on Artificial Intelligence (IJCAI-97)*, págs. 770–777, 1997.

[110] Kacholia V., Pandit S., Chakrabarti S., Sudarshan S., Desai R. an Karambelkar H. Bidirectional Expansion for Keyword Search on Graph Databases. En *31st International Conference on Very Large Data Bases (VLDB-05)*, págs. 487–499, 2005.

[111] Kamba T., Bharat K. y Albers M.C. The Krakatoa Chronicle - an Interactive, Personalized Newspaper on the Web. En *4th International World Wide Web Conference (WWW-95)*, págs. 159–170, 1995.

[112] Kamba T., Sakagami H. y Koseki Y. ANATAGONOMY: A Personalized Newspaper on the World Wide Web. *International Journal of Human-Computer Studies*, 46(6):789–803, 1997.

[113] KAON. Disponible en *http://kaon.semanticweb.org/*.

[114] Karvounarakis G., Alexaki S., Christophides V., Plexousakis D. y Scholl M. RQL: A Declarative Query Language for RDF. En *11th International World Wide Web Conference (WWW-02)*, págs. 195–205, 2002.

[115] KIM. Disponible en *http://www.ontotext.com/kim/index.html*.

[116] Kleinberg J. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):48–60, 1999.

[117] Knight K. Connectionist ideas and algorithms. *Communications on the ACM*, 33(11):59–74, 1990.

[118] Kobsa A., Koenemann J. y Pohl W. Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review*, 16(2):111–155, 2001.

[119] Kohonen T. *Self-Organization and Associative Memory*. Springer Verlag, 1989.

[120] Konstan J., Miller B., Maltz D., Herlocker J., Gordon L. y Riedl J. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[121] Kornfeld M. y Reimers U. *DVB-H, the emerging standard for mobile data communication*. EBU Technical Review, 2005.

[122] Kovács A. y Ueno H. Recommending in Context: A Spreading Activation Model that is Independent of the Type of Recommender System and Its Contents. En *2nd Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSIUI-06)*, págs. 49–59, 2006.

[123] Kowari. Disponible en *http://kowari.org/*.

[124] KPOntology. Disponible en *http://kpontology.isoco.net/*.

[125] Krulwich B. LifeStyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. *AI Magazine*, 18(2):37–45, 1997.

[126] Krulwich B. y Burkey C. Contact Finder: Extracting indications of expertise and answering questions with referrals. En *Technical Report, FS-95-03. The AAAI Press*, 1995.

[127] Krulwich B. y Burkey C. Learning User Information Interests through Extraction of Semantically Significant Phrases. En *AAAI Spring Symposium on Machine Learning Learning in Information Access*, págs. 22–27, 1996.

[128] Landauer T. y Littman M. Fully automatic cross-language document retrieval using Latent Semantic Indexing. En *6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, págs. 31–38, 1990.

[129] Landauer, T., Foltz, P. W. y Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2):259–284, 1998.

[130] Lang K. NewsWeeder: Learning to Filter News. En *12th International Conference on Machine Learning (ICML-95)*, págs. 331–339, 1995.

[131] Le Grand B., Soto M. y Dodds M. XML Topic Maps and Semantic Web Mining. En *XML Conference and Exhibition*, págs. 5–40, 2001.

[132] Lee J., Kim M. y Lee Y. Information Retrieval based on Conceptual Distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.

[133] Lekakos G., Chorianopoulus K. y Spinellis D. Information Systems in the Living Room: A Case Study of Personalized Interactive TV Design. En *9th European Conference on Information Systems (ECIS-01)*, págs. 687–695, 2001.

[134] Lemire D., Boley H., McGrath S. y Ball M. Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation. *International Journal of Interactive Technology and Smart Education*, 2(3):179–188, 2005.

[135] Librelotto G. R., Ramalho J. an Henriques P. TM-builder: An ontology builder based on XML Topic Maps. *CLEI Electronic Journal*, 7(2):24–38, 2004.

[136] Lieberman H. Letizia: An Agent that Assists Web Browsing. En *14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, págs. 924–929, 1995.

[137] Lieberman H., Van Dyke N. y Vivacqua A. S. Let's Browse: A Collaborative Web Browsing Agent. En *International Conference on Intelligent User Interfaces (IUI-99)*, págs. 924–929, 1999.

[138] Lin D. An Information-Theoretic Definition of Similarity. En *15th International Conference on Machine Learning (ICML-98)*, págs. 296–304, 1998.

[139] Lin W., Álvarez S. y Ruiz C. Collaborative recommendation via adaptive association rule mining. En *ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD-00)*, págs. 40 – 47, 2000.

[140] MacGregor R. Inside the LOOM Description Classifier. En *Special Issue on Implemented Knowledge Representation and Reasoning Systems, SIGART Bulletin 2*, págs. 88–92, 1991.

[141] Maguitman A., Menczer F., Roinestad H. y Vespignani A. Algorithmic Detection of Semantic Similarity. En *14th International World Wide Web Conference (WWW-05)*, págs. 107–116, 2005.

[142] Mak H., Koprinska I. y Poon J. INTIMATE: A Web-Based Movie Recommender Using Text Categorization. En *IEEE/WIC International Conference on Web Intelligence (WI-03)*, págs. 602 – 612, 2003.

[143] Maloof M. y Michalski R. Selecting Examples for Partial Memory Learning. *Machine Learning*, 41(1):27–52, 2000.

[144] Marc de Graauw. *Using Topic Maps To Extend Relational Databases*. XML.com, 2003.

[145] Martin-Smith P., Pelayo F. J., Díaz A. F., Ortega J., Prieto A. A Learning Algorithm to Obtain Self-Organizing Maps Using Fixed Neighbourhood Kohonen Networks. En *International Workshop on Artificial Neural Networks (IWANN-93)*, págs. 297–304, 1993.

[146] MBMS white paper. Telia Sonera. *Disponible para consulta en el enlace http://www.medialab.sonera.fi/workspace/MBMSWhitePaper.pdf*. 2004.

[147] McGuinness D. y van Harmelen F. OWL Web Ontology Language Overview. W3C Recommendation. 2004.

[148] Melville P., Mooney R. y Nagarajan R. Content-Boosted Collaborative Filtering for Improved Recommendations. En *18th National Conference on Artificial Intelligence (AAAI-2002)*, págs. 187–192, 2002.

[149] Middleton S. *Capturing Knowledge of User Preferences with Recommender Systems*. PhD thesis, Department of Electronic and Computer Science. University of Southampton, 2003.

[150] Miller B., Albert I., Lam S., Konstan J. y Riedl J. Movielens unplugged: experiences with an occasionally connected recommender system. En *ACM Conference on Intelligent User Interfaces*, págs. 263–266, 2003.

[151] Miller L., Seaborne A. y Reggiori A. Three Implementations of SquishQL, a Simple RDF Query Language. En *1st International Semantic Web Conference (ISWC-02)*, págs. 423–435, 2002.

[152] MindSwap. Disponible en *http://www.mindswap.org/2004/SWOOP/*.

[153] Minio M. y Tasso C. User Modeling for Information Filtering on Internet Services: Exploiting an Extended Version for the UMT Shell. En *Workshop on User Modeling for Information Filtering on the World Wide Web*, págs. 45–55, 1996.

[154] Minsky M. A framework for representing knowledge. *The Psychology of Computer Vision*, págs. 211–277, 1975.

[155] Mitchell T., Caruana R., Freitag D., McDermott J. y Zabowski D. Experience with a Learning Personal Assistant. *Communications of the ACM*, 37(7):81–91, 1994.

[156] Mitchell T. M., Mahadevan S. y Steinberg L. Leap: A Learning Apprentice for VLSI Design. En *4th International Joint Conference on Artificial Intelligence (IJCAI-85)*, págs. 573–580, 1985.

[157] Mladenic D. Personal WebWatcher: Implementation and Design. En *Technical Report IJS-DP-7472, Department of Intelligent Systems*, 1996.

[158] Mobasher B., Cooley R. y Srivastava J. Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8):142–151, 2000.

[159] Mobasher B., Dai H., Luo T., Nakagawa M., Sun Y., Wiltshire J. Discovery of Aggregate Usage Profiles for Web Personalization. En *ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD-00)*, págs. 61 – 68, 2000.

[160] Mobasher B., Jin X. y Zhou Y. Semantically Enhanced Collaborative Filtering on the Web. En *Web Mining: applications and techniques*, págs. 57–76. IDEA Group Publishing, 2004.

[161] Moldovan D. y Mihalcea R. A WordNet-based interface to Internet search engines. En *11th International Florida Artificial Intelligence Research Symposium Conferenceof (FLAIRS-98)*, págs. 275–279, 1998.

[162] Moldovan D. y Mihalcea R. Using WordNet and Lexical Operators to improve Internet Searches. *IEEE Internet Computing*, 4(1):34–43, 2000.

[163] Montaner M., López B. y De la Rosa J. A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review*, 19(4):285–330, 2003.

[164] Moody J. y Darken C. J. Fast Learning in Networks of Locally Tuned Processing Units. *Neural Computation*, 1(2):281–294, 1989.

[165] Mooney R. y Roy L. Content-based book recommending using learning for text categorization. En *Workshop on Recommender Systems: Algorithms and Evaluation*, págs. 195–204, 1999.

[166] Morato J., Marzal M., Lloréns J. y Moreiro J. WordNet applications. En *2nd International WordNet Conference (GWC-04)*, págs. 270–278, 2004.

[167] Morita M. y Shinoda Y. Information filtering based on user behavior analysis and best match text retrieval. En *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, págs. 272–281, 1994.

[168] Moukas A. Amalthea: Information Filtering and Discovery Using in a Multiagent Evolving System. *Journal of Applied AI*, 11(5):437–457, 1997.

[169] Moviefinder Recommender System. Disponible en *http://www.moviefinder.com*.

[170] Mukherjea S. y Bamba B. BioPatentMiner: An Information Retrieval System for BioMedical Patents. En *30th International Conference on Very Large Data Bases (VLDB-04)*, págs. 1066–1077, 2004.

[171] Murthy S. K. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

[172] Nemrava J. Using WordNet glosses to refine Google queries. En *Conference on Databases, Texts, Specifications, and Objects (DATESO-06)*, págs. 14–18, 2006.

[173] O'Hara K., Alani H. y Shadbolt N. Identifying Communities of Practices: Analyzing Ontologies as Networks to Support Community Recognition. En *Information Systems: The e-Business Challenge (IFIP-WCC 2002)*, págs. 89–102, 2002.

[174] OilEd. Disponible en *http://oiled.man.ac.uk/*.

[175] OntoAgent. *http://www.aifb.uni-karlsruhe.de/WBS/aeb/ontoagent/*.

[176] OntoEdit. Disponible en *http://www.ontoknowledge.org/tools/ontoedit.shtml*.

[177] OntoJava. Disponible en *http://www.aifb.uni-karlsruhe.de/WBS/aeb/ontojava/*.

[178] OntoLingua. Disponible en *http://www.ksl.stanford.edu/software/ontolingua/*.

[179] Ontología Teknowledge Ontology of Geography. Información adicional disponible en *http://reliant.teknowledge.com/DAML/Geography.owl*. 2001.

[180] Ontomat. Disponible en *http://annotation.semanticweb.org/ontomat/index.html*.

[181] OntoWeaver. Disponible en *http://kmi.open.ac.uk/projects/akt/ontoweaver/*.

[182] OntoWebber. *http://www-db.stanford.edu/OntoAgents/OntoWebber/*.

[183] Orwant L. J. Heterogeneous Learning in the Doppelganger User Modelling System. *User Modeling and User-Adapted Interaction*, 42(2):107–130, 1995.

[184] O'Sullivan D., Smyth B., Wilson D. y McDonald K. Improving the Quality of the Personalized Electronic Program Guide. *User Modeling and User-Adapted Interaction*, 14(1):5–36, 2004.

[185] Page L., Brin S., Motwani R. y Winograd T. The PageRank citation ranking: Bringing order to the Web. En *Technical Report, Stanford Digital Library Technologies Project*, 1998.

[186] Paolucci M., Kawamura T., Payne T. y Sycara K. Semantic Matching of Web Services Capabilities. En *1st International Semantic Web Conference (ISWC-02)*, págs. 333–347, 2002.

[187] Papadimitriou C., Raghavan P., Tamaki H. y Vempala S. Latent Semantic Indexing: A Probabilistic Analysis. En *17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, págs. 159–168, 1998.

[188] Patel-Scheneider P.F., McGuiness D., Brachman R., Resnick L.A. y Borgida A. The classic knowledge representation system: Guiding principles and implementation rationale. En *Special Issue on Implemented Knowledge Representation and Reasoning Systems, SIGART Bulletin 2*, págs. 108–113, 1991.

[189] Pazzani M. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5):393–408, 1999.

[190] Pazzani M. y Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, 27(3):313–331, 1997.

[191] Pazzani M., Muramatsu J. y Billsus D. Syskill & Webert: Identifying Interesting Web Sites. En *13th National Conference on Artificial Intelligence (AAAI-96)*, págs. 54–61, 1996.

[192] Pellet. Disponible en *http://www.mindswap.org/2003/pellet/*.

[193] Pepper S. y Moore G. *XML Topic Maps (XTM) 1.0. TopicMaps.Org. Disponible en http://www.topicmaps.org/xtm/index.html*. 2001.

[194] Perry M., Janik M., Ramakrishnan C., Ibañez C., Arpinar B. y Sheth A. Peer-to-Peer Discovery of Semantic Associations. En *2nd International Workshop on Peer-to-Peer Knowledge Management (P2PKM-05)*, págs. 1–12, 2005.

[195] Plugin de Protégé con soporte para SWRL (SWRL-Protégé Plugin). Disponible en *http://protege.stanford.edu/plugins/owl/swrl/*.

[196] pOWL. Disponible en *http://powl.sourceforge.net/*.

[197] Protégé. Disponible en *http://protege.stanford.edu/*.

[198] PROTON (PROTo ONtology). Disponible en *http://proton.semanticweb.org/*.

[199] Quillian R. Semantic memory. En *Semantic Information Processing*. The MIT Press, 1968.

[200] Quinlan J. R. Learning Efficient Classification Procedures and their Application to Chess End Games. En *Machine Learning: An Artificial Intelligence Approach*, págs. 463–482, 1983.

[201] Rada R., Mili H., Bicknell E. y Blettner M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[202] Rafter R., Bradley K y Smyth B. Personalized retrieval for online recruitment services. En *22nd Annual Colloquium on Information Retrieval*, págs. 58–69, 1999.

[203] Ramakrishnan C., Milnor W., Perry M. y Sheth A. Discovering informative connection subgraphs in multi-relational graphs. *SIGKDD Explorations Newsletter*, 7(2):56–63, 2005.

[204] RDFCrawler. Disponible en *http://ontobroker.semanticweb.org/rdfcrawl/*.

[205] Replay TV. Disponible en *http://www.replaytv.com*.

[206] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11(4):95–130, 1999.

[207] Rich E. User Modeling via Stereotypes. *International Journal of Cognitive Science*, 3(1):329–354, 1979.

[208] Riordan A. y Sorensen H. An intelligent agent for high-precision text filtering. En *4th International Conference on Information and Knowledge Management (CIKM-95)*, págs. 205–211. ACM Press, 1995.

[209] Rocha C., Schawabe D. y Poggi M. A Hybrid Approach for Searching in the Semantic Web. En *13th International World Wide Web Conference (WWW-04)*, págs. 374–383, 2004.

[210] Rodriguez M. A. *Assessing semantic similarity among spatial entity classes*. PhD thesis, University of Maine, 2000.

[211] Roget's Thesaurus. *Disponible en http://www.gutenberg.org*.

[212] Rosenblatt F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, 1962.

[213] RuleML. Disponible en *http://www.ruleml.org/*.

[214] Rumelhart D. E. y MacClelland J. L. *Parallel Distributed Processing. Vol 1*. MIT Press, 1986.

[215] Sahami M. Learning Non-Linearly Separable Boolean Functions With Linear Threshold Unit Trees and Madaline-Style Networks. En *11th National Conference on Artificial Intelligence (AAAI-93)*, págs. 335–341, 1993.

[216] Salton G. y McGill M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[217] Sarwar B., Karypis G., Konstan J. y Riedl J. Application of Dimensionality Reduction in Recommender System: A Case Study. En *Web Mining for E-Commerce: Challenges and Opportunities (WebKDD-00)*, págs. 45–58, 2000.

[218] Sarwar B., Karypis G., Konstan J. y Riedl J. Item-based collaborative filtering recommendation algorithms. En *10th International WWW Conference (WWW-01)*, págs. 285–295, 2001.

[219] Schmitz C., Hotho A., Jäschke R. y Stumme G. Content Aggregation on Knowledge Bases using Graph Clustering. En *3rd European Semantic Web Conference (ESWC-06)*, págs. 61–75, 2006.

[220] Schwab I., Kobsa A. y Koychev I. *Learning user interests through positive examples using content analysis and collaborative filtering*. Draft from Fraunhofer Institute for Applied Information Technology, Germany, 2001.

[221] Seaborne A. RDQL: A Data Oriented Query Language for RDF Models. En *Disponible en http://www.hpl.hp.com/semweb/rdql-grammar.html*, 2001.

[222] Sesame. Disponible en *http://www.openrdf.org/*.

[223] Shannon C. y Weaver W. The mathematical theory of communication. En *University of Illinois Press*, 1948.

[224] Shardanand U. *Social Information Filtering for Music Recommendation*. PhD thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, 1994.

[225] Shardanand U. y Maes P. Social Information Filtering: Algorithms for Automating "Word of Mouth". En *International Conference on Human Factors in Computing Systems (CHI-95)*, págs. 210–217, 1995.

[226] Sheth A., Bertram C., Avant D., Hammond B., Kochut K. y Warne Y. Managing Semantic Content for the Web. *IEEE Internet Computing*, 6(4):80–87, 2002.

[227] Sheth B. y Maes P. Evolving Agents for Personalized Information Filtering. En *9th International Conference on Artificial Intelligence for Applications (AIA-93)*, págs. 331–339, 1993.

[228] Sintek M. y Decker S. TRIPLE -A Query, Inference and Transformation Language for the Semantic Web. En *1st International Semantic Web Conference (ISWC-02)*, págs. 364–378, 2002.

[229] Smyth B. y Cotter P. Surfing the digital wave: generating personalized tv listings using collaborative, case-based recommendation. En *3rd International Conference on Case-Based Reasoning (CBR-99)*, págs. 561 – 571, 1999.

[230] Smyth B. y Cotter P. A personalized TV listing service for the Digital TV age. *Journal of Knowledge-Based Systems*, 13(2):53–59, 2000.

[231] Smyth B. y McClave P. Similarity vs. diversity. En *4th International Conference on Case-Based Reasoning (ICCBR-01)*, págs. 347–361, 2001.

[232] Sorensen H. y McElligott M. PSUN: A Profiling System for Usenet News. En *Workshop on Intelligent Information Agents*, págs. 121–127, 1995.

[233] Staab S. y Studer R. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

[234] Stefani A. y Strapparava C. Personalizing Access to Web Sites: The SiteIF Project. En *2nd Workshop on Adaptive Hypertext and Hypermedia (HYPERTEXT-98)*, págs. 69–75, 1998.

[235] Stojanovic N., Struder R. y Stojanovi L. An approach for the ranking of query results in the Semantic Web. En *2nd International Semantic Web Conference (ISWC-03)*, págs. 500–516, 2003.

[236] Sure Y., Erdman M., Angele J., Staab S., Studer R., Wenke D. OntoEdit: Collaborative Ontology Development for the Semantic Web . En *1st International Semantic Web Conference (ISWC-02)*, págs. 221–236, 2002.

[237] Sure Y., Maedche A. y Staab S. Leveraging corporate skill knowledge: from proper to ontoproper. En *4th International Conference on Practical Aspects of Knowledge Management (PAKM-00)*, págs. 34–44, 2000.

[238] Semantic Web Rule Language (SWRL). Disponible en *http://www.w3.org/Submission/SWRL/*.

[239] Takefuji Y. y Szu H. Design of Parallel Distributed Cauchy Machines. En *International Joint Conference on Neural Networks (IJCNN-89)*, págs. 529–532, 1989.

[240] Tatemura J. Virtual Reviewers for Collaborative Exploration of Movie Reviews. En *ACM Intelligent User Interfaces Conference (IUI-2000)*, págs. 272–275, 2000.

[241] TiVo, Inc. TiVo: TV your way. Disponible en *http://www.tivo.com*, year = 2002.

[242] Trastour D., Bartolini C. y Priest C. Semantic Web Support for the Business-to-Busines E-Commerce Lyfecycle. En *11th International World Wide Web Conference (WWW-02)*, págs. 89–98, 2002.

[243] TV-Anytime on DVB. Carriage and signalling of TV-Anytime information in DVB transport streams. European Standard ETSI TS 102 323 V1.2.1, 2006.

[244] TV-Anytime Specification Series S-3 on Metadata. *Disponible en http://www.tv-anytime.org*. 2001.

[245] TV-Anytime. Broadcast and on-line services: Search, select, and rightful use of content on personal storage systems. European Standard ETSI TS 102 822 series, 2006.

[246] Web Service Modeling Ontology Group. Disponible en *http://www.wsmo.org/*.

[247] Webb G. y Kuzmycz M. Feature based Modeling: A Methodology for Producing Coherent, Consistent, Dynamically Changing Models of Agents'Competencies. *User Modeling and User-Adapted Interaction*, 4(1):117–150, 1996.

[248] WebODE. Disponible en *http://webode.dia.fi.upm.es/WebODEWeb/index.html*.

[249] WebOnto. Disponible en *http://kmi.open.ac.uk./projects/webonto*.

[250] Weigand L. A Multilingual Ontology-based Lexicon for News Filtering - The TREVI Project. En *15th International Joint Conferences on Artificial Intelligence (IJCAI-97)*, págs. 160–165, 1997.

[251] White B. A. y Elmasry M. I. The digi-neocognitron: a digital neocognitron neurall network model for VLSI. *IEEE Transactions on Neural Networks*, 3(1):167–172, 1992.

[252] Widmer G. y Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23(1):69–101, 1996.

[253] Widrow B. y Winter R. Adaptive switching circuits. *IRE WESCON Convention Record*, 4(1):96–104, 1960.

[254] WordNet. *Disponible en http://wordnet.princeton.edu*.

[255] WSMO Studio. Disponible en *http://www.wsmostudio.org/*.

[256] Yan T. y García-Molina H. Sift - a Tool for Wide-Area Information Dissemination. En *USENIX Technical Conference*, págs. 177–186, 1995.

[257] Yoshida S., Yukawa T. y Kuwabara K. Constructing and Examining Personalized Cooccurrence-based Thesauri on Web Pages. En *12th International World Wide Web Conference (WWW-03)*, págs. 48–52, 2003.

[258] Zhuge H. y Zheng L. Ranking Semantic-linked Network. En *12th International World Wide Web Conference (WWW-03)*, págs. 72–76, 2003.

[259] Ziegler C. N., Schmidt-Thieme L. y Lausen G. Exploiting Semantic Product Descriptions for Recommender Systems. En *2nd ACM SIGIR Semantic Web and Information Retrieval Workshop (SWIR-04)*, págs. 25–29, 2004.

[260] Zimmerman J., Kurapati K., Buczak A., Schafer D., Gutta S. y Martino J. TV Personalization System: Design of a TV Show Recommender Engine and Interface. En *Personalized Digital Television: Targeting Programs to Individual Viewers*, págs. 27 – 51. Kluwer Academics Publishers, 2004.