Universidade Vigo

DEPARTAMENTO DE ENXEÑERÍA TELEMÁTICA
E.E. TELECOMUNICACIÓN

TESIS DOCTORAL

ESTRATEGIAS PARA EL INCREMENTO DE LA FIABILIDAD Y CONFIANZA EN SISTEMAS DE RECOMENDACIÓN COLABORATIVA. UN ENFOQUE SEMÁNTICO PARA LA PERSONALIZACIÓN DE LA PUBLICIDAD EN COMERCIO ELECTRÓNICO.

OPTA AL GRADO DE DOCTOR POR LA UNIVERSIDAD DE VIGO

Autora: Manuela I. Martín Vicente

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

2012

Vigo, a 23 de marzo de 2012

Los abajo firmantes, D. Manuel Ramos Cabrer (DNI 33305656-T) y D. Alberto Gil Solla (DNI 35310169-V), ambos profesores Titulares de Universidad y adscritos al Departamento de Ingeniería Telemática de la Universidad de Vigo, declaramos que hemos codirigido el trabajo de tesis doctoral de Dña. Manuela Isabel Martín Vicente, y que damos nuestro consentimiento para proceder al depósito de la tesis e iniciar los trámites conducentes a su defensa.

Manuel Ramos Cabrer

Alberto Gil Solla

TESIS DOCTORAL

ESTRATEGIAS PARA EL INCREMENTO DE LA FIABILIDAD Y CONFIANZA EN SISTEMAS DE RECOMENDACIÓN COLABORATIVA. UN ENFOQUE SEMÁNTICO PARA LA PERSONALIZACIÓN DE LA PUBLICIDAD EN COMERCIO ELECTRÓNICO.

Autora: Manuela I. Martin Vicente

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

TRIBUNAL CALIFICADOR

Presidente:
Vocales:

Secretario:

CALIFICACIÓN:

Vigo, a

Resumen

Desde su nacimiento en la década de los años 60, la capacidad de transmisión y almacenamiento de datos en Internet ha crecido hasta límites insospechados, dando lugar a una *sobrecarga de información* que se ha extendido a dominios de muy diversa naturaleza —como 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 es relevante para ellos (ciertos productos comerciales, contenidos de televisión, cursos educativos, etc.) en una tarea tediosa, difícil de asumir sin la utilización de herramientas de búsqueda intuitivas y eficientes.

Con el afán de realizar un filtrado de dicha información de forma personalizada y automática, surgen en los años 90 los sistemas recomendadores. Precisamente el objetivo último de estos mecanismos es ofrecer a cada usuario, de entre la miríada de productos disponibles, aquellos que mejor se adapten a sus preferencias o necesidades.

Desde sus orígenes, los sistemas recomendadores han despertado gran interés en la comunidad investigadora. En particular, se han convertido en herramientas muy apreciadas en el dominio del comercio electrónico y el marketing debido a su rendimiento económico, ya que la selección de forma individualizada de productos comerciales, identificando artículos que encajen con los gustos de cada usuario o le puedan resultar de utilidad (generando incluso nuevas necesidades a los consumidores), puede aumentar considerablemente los ingresos por ventas.

La evolución de estos sistemas ha sido notable, principalmente en lo referente a las técnicas de filtrado empleadas y a la incorporación de información adicional acerca de los usuarios en el proceso de recomendación, mejorando la calidad de las sugerencias ofrecidas. Sin embargo, a pesar de los avances logrados hasta la fecha en el campo de la personalización, se identifican carencias importantes en las estrategias adoptadas en los sistemas recomendadores actuales. Por una parte, su utilización en dominios de aplicación caracterizados por una gran heterogeneidad de productos disponibles, como es el caso del comercio electrónico, menoscaba la calidad de las sugerencias ofrecidas a los usuarios. Por otro lado, la incorporación de información en el proceso de recomendación se ve limitada por la necesidad de obtener dicha información de forma explícita por parte de los propios usuarios.

El trabajo presentado en esta tesis se centra en el desarrollo de mecanismos efectivos que combatan tales limitaciones, mejorando la calidad de las recomendaciones de forma transparente a los usuarios —sin requerir que éstos proporcionen ningún tipo de información adicional de forma expresa—.

Para ello, la metodología desarrollada se enmarca en una prometedora línea de investigación surgida en la última década en el campo de los sistemas recomendadores:

los enfoques de personalización semánticos, que combinan técnicas de inferencia de conocimiento, tradicionalmente vinculadas a la Web Semántica, con las estrategias de modelado de usuarios y filtrado de la información utilizadas en dichos sistemas. A su vez, nuestro enfoque aprovecha la masiva adopción de las tecnologías Web 2.0 en la actualidad e incorpora información adicional, extraída de redes sociales, en el proceso de recomendación.

Si bien las estrategias propuestas en esta tesis no están vinculadas a un dominio determinado, su aplicación es especialmente beneficiosa en dominios donde existe una gran diversidad en la naturaleza de los productos disponibles. Por este motivo, el comercio electrónico es el contexto elegido para ilustrar el enfoque propuesto y describirlo con suficiente claridad.

Palabras clave: Sistemas recomendadores, comercio electrónico, razonamiento semántico, Web 2.0

Abstract

The data storage and transmission capabilities of Internet have tremendously grown since its inception in the 60s, giving rise to the so-called *information overload* problem in several application domains —such as e-commerce, Digital Television or distance learning via electronic means (*e-learning*, *t-learning*, *m-learning*). In this scenario, users have to deal with overwhelming amounts of data, turning the selection of relevant information (e.g. certain commercial products, TV contents or educational courses) into a cumbersome task, extremely hard to tackle without the assistance of intuitive and efficient searching tools.

Aiming to overcome this problem, recommender systems arose in the 90s with the purpose of automatically selecting such information in a personalized way. Indeed, the final goal of these tools is offering to each user, from among the myriad of products available, those that they may like or find useful.

Recommender systems have attracted great interest of the research community. Particularly, marketing and e-commerce application domains have been increasingly explored in the literature, due to their business interest, since identifying those products that match each user's preferences (even creating new needs for consumers) can considerably increase sales revenue.

The developments in this research field have been remarkable, especially regarding information filtering techniques and the incorporation of additional data about the users in the personalization process, improving recommendation results. However, significant limitations are identified in the strategies adopted by recommender systems nowadays. On the one hand, the great diversity of products available in heterogeneous application domains, such as e-commerce, makes the current recommendation techniques spoil the quality of the selected recommendations. On the other hand, the incorporation of extra data in order to enhance the personalization process, usually requires the users to explicitly provide such information.

The work presented in this PhD thesis is focused on the development of effective mechanisms that overcome those limitations, improving recommendation results transparently to the users —without requiring them to provide any kind of additional data explicitly.

With this goal in mind, the developed methodology is framed in a promising research direction that have arisen in the last decade within the recommender systems field: the semantic personalization approaches, which combine reasoning techniques from the Semantic Web with the recommendation strategies —in both the user modeling and the information filtering techniques. Our approach also makes use of the present widespread adoption of Web 2.0 technologies to obtain additional data from social networks, in order to include them in the recommendation process.

Although the strategies proposed in this thesis are not linked to any specific domain at all, they overcome limitations related to the use of recommender systems in heterogeneous application domains. That is why we have chosen e-commerce as the context to illustrate our approach and describe its capabilities clearly enough.

Key words: Recommender systems, e-commerce, semantic reasoning, Web 2.0

Índice general

Índice de figuras IX

I Introducción 1

1. Ámbito y objetivos de la tesis 3
1.1. Antecedentes y motivación 3
1.2. Ámbito general del trabajo 6
1.3. Objetivos de la tesis 7
1.4. Organización de la memoria 9

II Estado del arte 11

2. Estado del arte 13
2.1. Introducción 13
2.2. Modelado de usuarios 14
2.2.1. Representación del perfil del usuario 14
2.2.2. Generación del perfil inicial 20
2.2.3. Realimentación de relevancia 21
2.2.4. Técnicas para la adaptación del perfil de usuario 23
2.3. Métodos de filtrado de información 24
2.3.1. Filtrado demográfico 24
2.3.2. Métodos basados en contenido 25
2.3.3. Filtrado colaborativo 27
2.3.4. Enfoques híbridos 32
2.4. Problemas abiertos en los sistemas recomendadores colaborativos . . 33
2.4.1. Problema de falsos vecindarios 33
2.4.2. Verificación de la fiabilidad de los usuarios 35
2.4.3. Cuantificación de la experiencia de los usuarios 37
2.4.4. Problema de *cold start users* 38
2.4.5. *Sparsity problem* 38

2.5. Nuevos enfoques de personalización 39
2.5.1. Enfoques semánticos 39
2.5.2. Enfoques vinculados a la Web 2.0 44
2.6. Conclusiones 45

III Estrategias 47

3. Marco de razonamiento semántico 49
3.1. Un sistema recomendador colaborativo semántico 49
3.2. Ontología en el dominio del comercio electrónico 50
3.2.1. Singularidades de las ontologías de productos heterogéneos . 51
3.3. Modelado de usuarios basado en perfiles-ontología 53
3.3.1. Perfiles-ontología 53
3.3.2. Generación del perfil inicial 55
3.3.3. Realimentación de relevancia 56
3.3.4. Cálculo de los niveles de interés 56
3.3.5. Identificadores de usuario 58
3.3.6. Adaptación ante cambios en las preferencias de los usuarios . 59
3.4. Conclusiones 61

4. Estrategia para combatir el problema de falsos vecindarios 63
4.1. Introducción 63
4.2. Similitud semántica entre productos 65
4.2.1. Similitud semántica jerárquica 65
4.2.2. Similitud semántica inferencial 68
4.3. Similitud entre usuarios 72
4.3.1. Medidas convencionales 72
4.3.2. Un enfoque semántico 73
4.3.3. Mitigando el problema de falsos vecindarios 76
4.4. Ejemplo de aplicación 81
4.4.1. Problema de falsos vecindarios 83
4.4.2. Estrategia semántica propuesta 83
4.5. Beneficios de la estrategia 86

5. Estrategia para la inclusión de un modelo de confianza implícita 89
5.1. Introducción 89
5.2. Cuantificación de relaciones de confianza entre pares de usuarios . . . 91
5.2.1. Relaciones de confianza 91
5.2.2. Definición de éxito 91
5.2.3. Enfoque semántico vs. enfoque sintáctico 92
5.2.4. Inferencia de relaciones de confianza condicionadas a clase . . 93

5.3. Construcción de redes de confianza implícita 95
5.4. Cuantificación de medidas de reputación 96
5.5. Selección de vecindario *trust-aware* 98
5.6. Ejemplo de aplicación 99
5.6.1. Relaciones de confianza condicionadas a la clase *Rock* 100
5.6.2. Reputación de CN_4 102
5.6.3. Selección de vecindario *trust-aware* 102
5.7. Beneficios de la estrategia 103

6. Estrategia para la cuantificación de la experiencia práctica de los usuarios 105
6.1. Introducción 105
6.2. Experiencia práctica 107
6.2.1. Variabilidad cuantitativa 108
6.2.2. Variabilidad cualitativa 110
6.3. Incorporación de medidas de experiencia en el proceso de recomendación 114
6.3.1. Métrica de predicción tradicional 114
6.3.2. Métrica de predicción propuesta 115
6.3.3. Criterio de recomendación: rango del umbral $\beta^{C_p}_{Pred_U}$ 115
6.4. Ejemplo de aplicación 116
6.4.1. Variabilidad cuantitativa de N_1 y N_2 117
6.4.2. Variabilidad cualitativa de N_1 y N_2 117
6.4.3. Experiencia práctica 120
6.5. Beneficios de la estrategia 122

7. Hibridación del proceso de selección de vecindarios con una componente basada en redes sociales 123
7.1. Introducción 123
7.2. Cuantificación de relaciones de confianza personal entre usuarios . . . 125
7.2.1. Factor de interacción 126
7.2.2. Factor de privacidad 127
7.3. Selección de vecindarios híbrida 128
7.3.1. Completitud del perfil 128
7.3.2. Madurez 132
7.4. Reuniendo las piezas del puzle 133
7.5. Beneficios de la estrategia 137

IV Conclusiones 139

8. Conclusiones y líneas futuras 141
8.1. Conclusiones 141
8.1.1. Estrategia para combatir el problema de falsos vecindarios . . 142

8.1.2. Estrategia para la inferencia de redes de confianza implícitas . 142
8.1.3. Estrategia para la cuantificación de la experiencia práctica de los usuarios 143
8.1.4. Estrategia para la cuantificación de relaciones de confianza personal entre usuarios 144
8.1.5. Selección de vecindarios híbrida 144
8.2. Líneas de trabajo futuro 145

Índice de figuras

3.1. Visión general del proceso de recomendación colaborativo 50
3.2. Micro extracto de nuestra ontología de productos comerciales 52
3.3. item-IDs en los perfiles de dos usuarios 54
3.4. Ventana temporal utilizada en la función de olvido gradual 60

4.1. Similitud semántica jerárquica entre productos 66
4.2. Asociación ρ-*path* entre los productos i_i e i_j 68
4.3. Algunas asociaciones semánticas utilizadas en el cálculo de la similitud semántica inferencial 69
4.4. Asociación ρ-*join* entre los productos i_i e i_j 70
4.5. Similitud entre usuarios a fin de mitigar el problema de falsos vecindarios 77
4.6. Subárboles en función del nivel de abstracción 79
4.7. Extracto de la ontología considerado en el ejemplo de aplicación . . . 84

5.1. Representación gráfica de una relación de confianza 92
5.2. Clasificación jerárquica de la clase *Bebop* 94
5.3. Construcción de redes de confianza implícita 97
5.4. Extracto de la red de confianza implícita condicionada a *Rock* 100
5.5. Contribuciones de CN_1 y CN_2 en recomendaciones a U —clase *Rock*— 101

6.1. Disimilitud semántica entre dos instancias 111
6.2. Disimilitud semántica entre $Gruyere_A$ y $Parmesano_A$ 121

7.1. Relaciones de confianza personal 125
7.2. Heterogeneidad 129
7.3. Completitud del perfil 132
7.4. Proceso de selección de vecindarios híbrida 136

Parte I

Introducción

Capítulo 1

Ámbito y objetivos de la tesis

"Todos conocemos el extraordinario poder de las tecnologías de la información y la comunicación. Del comercio a la telemedicina, de la enseñanza a la protección del medio ambiente, tenemos en nuestras manos, en nuestro escritorio y en el cielo por encima de nosotros la capacidad de mejorar la calidad de vida de miles de millones de personas."

***Kofi Annan**, séptimo Secretario General de las Naciones Unidas.*
Cumbre Mundial sobre la Sociedad de la Información, 2003.

1.1 Antecedentes y motivación

Vivimos en la denominada *Era de la Información*, en la que la revolución tecnológica digital —encabezada por Internet— ha posibilitado el intercambio libre de información a nivel global y el acceso inmediato a contenidos y servicios de todo tipo. Cada vez más, las TIC (*Tecnologías de la Información y la Comunicación*) forman parte importante de la vida diaria de un gran sector de la población mundial, teniendo un fuerte impacto en el modo de trabajo y aprendizaje, las relaciones interpersonales y grupales, el ocio y el comercio, entre otros.

Tanto es así que, en un corto espacio de tiempo, numerosas actividades cotidianas han experimentado significativas transformaciones. Por ejemplo, el uso del correo electrónico y las aplicaciones de mensajería instantánea han dejado obsoleto al correo tradicional en lo que a comunicación se refiere. Igualmente, Internet se ha erigido como la principal fuente de consulta de conocimiento de estudiantes y profesionales, relegando a los centros de documentación físicos a un segundo plano. Por su parte, la Televisión Digital ha abierto un amplio abanico de posibilidades para los espectadores, tanto en oferta de contenidos (cada vez es mayor el número de canales accesibles a través de las diversas plataformas de Televisión Digital), como respecto a las nuevas capacidades interactivas del medio.

Mención aparte merecen las tecnologías Web 2.0 (tales como blogs, *wikis*, foros de debate y redes sociales), por el gran impacto que han generado en la sociedad, permitiendo a los usuarios Web ser *productores* de información, interactuando y colaborando entre sí, y no sólo *consumidores* pasivos de contenidos.

Por otro lado, cabe destacar la creciente implantación y expansión del comercio electrónico que, en los últimos años, se ha convertido en una interesante alternativa

o complemento a los hábitos de consumo tradicionales, llevando incluso a grandes empresas —con una amplia red de tiendas físicas— a abrir sus respectivas tiendas en línea. Esta proliferación se debe en gran medida a la ubicuidad y carácter global que lo caracterizan. En efecto, mediante sitios web de comercio electrónico el consumidor puede acceder a la compra de productos y servicios desde cualquier lugar, en cualquier momento del día, procedentes de cualquier rincón del mundo.

Desde su nacimiento en la década de los años 60, la capacidad de transmisión y almacenamiento de datos en Internet ha crecido hasta límites insospechados[1], dando lugar a una *sobrecarga de información* en la Red. Ciertamente, se llega a un punto en el que el usuario dispone de demasiadas fuentes para tomar una decisión o estar informado sobre un determinado tema.

A fin de mitigar este problema, surgen a principios de los años 90 los motores de búsqueda, herramientas de recuperación de información por palabras clave. Son sistemas que realizan una búsqueda y ordenación de las páginas web, imágenes u otros archivos que contienen, en alguna parte de su texto, las palabras clave especificadas por el usuario. Es el buscador Google [50] el que logra la hegemonía en el ámbito y la mantiene actualmente. Sin embargo, a pesar de su evolución, los usuarios aún sufren a día de hoy la frustración de iniciar una búsqueda y encontrar un sinfín de resultados poco o nada relacionados con su objetivo, entre los que se ocultan los enlaces que realmente les interesan.

Por otro lado, el problema de sobrecarga de información, que se gestó inicialmente en la Red, se ha extendido a dominios de muy diversa naturaleza —desde el comercio electrónico, a la Televisión Digital o a 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 productos que son relevantes para ellos (ciertos artículos comerciales, contenidos de televisión, cursos educativos, etc.) en una tarea tediosa y frecuentemente no exitosa, incluso con la utilización de motores de búsqueda —cuyos resultados, en ocasiones excesivos, dependen además de las palabras clave introducidas por el usuario—. Baste con pensar, por ejemplo, en la dificultad de exploración de sitios web de comercio electrónico en los que la oferta de productos de todo tipo es exorbitante.

Con el afán de realizar un filtrado de dicha información de forma *personalizada*, surgen a mediados de los años 90 los sistemas recomendadores [100]. Precisamente el objetivo último de estos mecanismos es ofrecer a cada usuario, de entre la miríada de productos disponibles, aquellos que mejor se adapten a sus preferencias o necesidades.

Los sistemas recomendadores

Desde sus orígenes, los sistemas recomendadores han despertado gran interés en la comunidad investigadora. En particular, se han convertido en herramientas muy apreciadas en el dominio del comercio electrónico y el marketing debido a su rendimiento económico, ya que la selección de forma individualizada de productos comerciales, identificando artículos que encajen con los gustos de cada usuario o le puedan resultar de utilidad —generando incluso nuevas necesidades a los consumidores—, puede aumentar considerablemente los ingresos por ventas [25].

En efecto, es posible encontrar actualmente en la Red sistemas recomendadores

[1] En el año 2000 se alcanzó la cifra de mil millones de páginas web indexadas y actualmente se ha superado el hito de un billón de URLs únicas [3].

de música (como Last.fm [72]), de cine (como FilmAffinity [41]) o de productos comerciales en general (como Amazon [4], cuyo sitio web de comercio electrónico es uno de los más populares a nivel mundial), utilizados por un gran número de usuarios diariamente.

La evolución de los sistemas recomendadores en la literatura ha sido notable, principalmente en lo referente a las técnicas de filtrado empleadas [1] y a la incorporación de información adicional acerca de los usuarios en el proceso de recomendación [88, 22]:

- Las técnicas basadas en *contenido*, consistentes en ofrecer a un cierto usuario productos similares a los que le han gustado en el pasado, se impusieron a las que únicamente consideraban características *demográficas* (como edad, sexo o estado civil del usuario), debido principalmente a la falta de individualización de estas últimas. Asimismo, a causa de su tendencia a seleccionar productos excesivamente parecidos a los que el usuario en cuestión ya conoce, los métodos basados en contenido perdieron popularidad frente a las estrategias de filtrado *colaborativo* —consistentes en recomendar a cada usuario productos que les han gustado a otros individuos con preferencias similares a las suyas—, siendo relegados a formar parte únicamente de aquellos enfoques que combinan ambas técnicas con el propósito de aunar sus ventajas y mitigar sus deficiencias.

- La calidad de las sugerencias ofrecidas por un sistema recomendador colaborativo dependen, por tanto, de la selección del grupo de usuarios afines a aquél que recibe las recomendaciones del sistema en un momento dado —ya que las preferencias de dicho grupo de usuarios determinarán que un cierto producto sea o no finalmente sugerido—. Es así que, con el objetivo de mejorar la fiabilidad de los sistemas y la precisión de las recomendaciones ofrecidas, numerosos trabajos se han centrado recientemente en la incorporación de información adicional acerca de los usuarios que condicionan cada recomendación.

  En concreto, cabe destacar la utilización de modelos de confianza y métricas de experiencia[2] en el proceso. El primer tipo de enfoques tratan de evitar la selección de usuarios poco fiables a la hora de determinar el grupo de usuarios afines uno dado, considerando no sólo la similitud de intereses, también las relaciones de confianza (o desconfianza) entre ellos. Íntimamente relacionados con estos enfoques, las propuestas que incorporan métricas de experiencia consideran los conocimientos que tienen en el dominio los usuarios que influyen en el proceso de recomendación.

Sin embargo, a pesar de los avances logrados hasta la fecha en el campo de la personalización, se identifican carencias importantes en las estrategias adoptadas en los sistemas recomendadores actuales.

Por una parte, su utilización en dominios de aplicación caracterizados por una gran heterogeneidad de productos disponibles —como es el caso del comercio electrónico— menoscaba la calidad de las sugerencias ofrecidas a los usuarios. En efecto, en tales dominios, las técnicas colaborativas convencionales son propensas a una selección contraproducente de usuarios afines al no tener en cuenta que, dada la diversidad existente en la naturaleza de los productos, los gustos de dos individuos pueden ser similares en productos de una índole pero totalmente contrarios en otros [75].

[2] Entendida la experiencia como “la aptitud o conocimiento que una persona tiene en un cierto campo”.

Por otro lado, la incorporación de métricas de experiencia y modelos de confianza significativos en el proceso de recomendación colaborativo se ve limitada por la necesidad de obtener dicha información de forma explícita por parte de los propios usuarios. Precisamente, el requerir que los usuarios especifiquen lo fiables que consideran a otros usuarios [88], o datos que permitan estimar su nivel de experiencia [22], es un gran inconveniente de las propuestas existentes, ya que se trata de información que los individuos pueden no querer proporcionar o, incluso, es desconocida por ellos.

1.2 Ámbito general del trabajo

El trabajo presentado en esta tesis se centra en el desarrollo de mecanismos efectivos que combatan las limitaciones que acabamos de mencionar, de forma *transparente* a los usuarios. Precisamente, el leitmotiv común a todas las contribuciones incluidas en la tesis es mejorar la calidad de las recomendaciones sin requerir que los usuarios proporcionen ningún tipo de información adicional de forma expresa. Cabe destacar el interés de este objetivo general, ya que no siempre es posible que los usuarios proporcionen el conocimiento que se requiere y, por otra parte, es habitual que éstos no estén dispuestos a proporcionar según qué información —bien por pereza, bien por cuestiones de privacidad—.

Con ese propósito en mente, la metodología desarrollada se enmarca en una prometedora línea de investigación surgida en la última década en el campo de los sistemas recomendadores: la integración de técnicas de inferencia de conocimiento, tradicionalmente vinculadas a la Web Semántica, con las estrategias de personalización utilizadas en dichos sistemas. A su vez, nuestro enfoque aprovecha la masiva adopción de las tecnologías Web 2.0 en la actualidad e incorpora información adicional, extraída de redes sociales, en el proceso de recomendación colaborativo.

Tal como acabamos de adelantar, los enfoques de personalización semánticos aprovechan la experiencia ganada en el campo de la Web Semántica para enriquecer el proceso de recomendación. En esencia, 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. Con este objetivo, se rescata la noción de ontología del campo de la Inteligencia Artificial [29]: una formalización consensuada y reutilizable en la que se identifican los conceptos y relaciones típicas en un dominio de aplicación. De ahí que la base sobre la que se apoya un sistema recomendador semántico, como el planteado en nuestro trabajo, sea una ontología que formaliza las descripciones semánticas de los productos propios del dominio de aplicación considerado.

La explotación de procesos de razonamiento semántico sobre una ontología de dominio posibilita la incorporación de gran cantidad de conocimiento útil en el sistema recomendador. Es así que encontramos en la literatura interesantes enfoques que emplean técnicas de inferencia complejas a fin de superar problemas derivados de la naturaleza sintáctica de las herramientas de personalización tradicionales. Aunque dichos enfoques no combaten las limitaciones identificadas en los sistemas recomendadores actuales referidas en esta tesis —razón por la cual hemos de desarrollar nuevos mecanismos y métricas apropiados a la tarea—, algunas de esas técnicas pueden ser adaptadas a las necesidades de nuestro contexto para aprovechar sus beneficios.

Si bien la metodología desarrollada no está vinculada a un dominio determinado,

y puede ser explotada en aplicaciones de personalización de muy diversa índole, su aplicación es especialmente beneficiosa en dominios donde existe una gran diversidad en la naturaleza de los productos disponibles. Para facilitar la comprensión del lector e ilustrar el enfoque propuesto con suficiente claridad, se ha optado por el comercio electrónico como contexto específico de aplicación. Por un lado, se trata de un dominio caracterizado por una gran heterogeneidad de productos. Por otra parte, tal como comentamos anteriormente, los sistemas recomendadores de productos comerciales (ya sea para su utilización en sitios de venta en línea o para la personalización de campañas de marketing) tienen un interés añadido debido a los beneficios económicos que pueden reportar.

1.3 Objetivos de la tesis

Considerando lo expuesto hasta ahora, detallamos a continuación los objetivos concretos perseguidos en esta tesis:

- Tal como adelantamos en la sección previa, la utilización de estrategias de filtrado colaborativo tradicionales en dominios de aplicación caracterizados por la heterogeneidad de sus productos repercute negativamente en la calidad de las recomendaciones. Recordemos que el éxito de una recomendación colaborativa depende principalmente de la selección de usuarios afines a aquel que recibe las sugerencias por parte del sistema, dado que las preferencias de dichos usuarios determinan el que un cierto producto sea recomendado o no. La carencia identificada en las técnicas tradicionales radica en que se limitan a seleccionar un único grupo de usuarios con intereses similares, independientemente del producto considerado en cada caso. En efecto, este modo de proceder es desfavorable en dominios cuyos productos son de muy distinta naturaleza, como es el caso del comercio electrónico —donde el sistema recomendará desde libros a música o productos alimenticios—, ya que los usuarios pueden tener intereses semejantes en unos tipos de productos y totalmente opuestos en otros.

  Con el propósito de combatir este problema, nuestro enfoque realiza dicha selección de forma *dinámica*, identificando para cada usuario un grupo u otro de individuos afines en función del producto considerado en la recomendación. Para ello, la estrategia desarrollada explota técnicas de razonamiento semántico sobre la ontología de dominio con un doble objetivo: por un lado, delimitar las preferencias de los usuarios a tener en cuenta en cada caso de forma flexible y significativa, según los diferentes tipos de productos existentes; y, por otra parte, descubrir relaciones ocultas en la base de conocimiento entre los intereses de los usuarios comparados, identificando similitudes que pasarían desapercibidas si únicamente se atiende a los términos que los designan.

- En cuanto a la incorporación de modelos de confianza en las estrategias colaborativas a fin de evitar la selección de usuarios afines maliciosos o poco fiables —y, por tanto, incrementar la fiabilidad de los sistemas y la calidad de las recomendaciones—, la mayoría de los enfoques propuestos en la literatura se ven limitados por la necesidad de obtener la información acerca de las relaciones de confianza entre los usuarios de forma explícita, requiriendo a los mismos que especifiquen cuánto confían en los otros usuarios. En consecuencia, al margen de precisar información que los usuarios pueden no estar dispuestos a proporcionar,

no son aptos para ser utilizados en sistemas abiertos donde los individuos, por norma general, no se conocen los unos a los otros.

Por este motivo, se propone en nuestro trabajo una estrategia para la inferencia de relaciones de confianza entre los usuarios de forma *implícita*, a partir de los resultados obtenidos en sugerencias previas. Dado que un proceso de recomendación colaborativo siempre involucra, por un lado, al usuario que recibe la recomendación y, por otro lado, al grupo de usuarios afines que determinan que un cierto producto sea o no sugerido al primero, es posible extraer ciertos vínculos entre ellos. Con el objetivo de inferir relaciones de confianza significativas entre los usuarios del sistema, nuestra estrategia explota la semántica formalizada en la ontología de dominio.

- Tal como mencionamos anteriormente, la necesidad de información explícita limita también la definición de métricas sobre la experiencia de los usuarios. Nuestra metodología incluye una estrategia para la cuantificación de la experiencia *práctica* de los usuarios a partir de la información contenida en cualquier sistema recomendador, empleando técnicas de razonamiento sobre la base de conocimiento de nuestro sistema. Más concretamente, dicha estrategia obtiene medidas de experiencia en base a la variedad de productos de un cierto tipo que los usuarios han probado en el pasado, atendiendo a las descripciones semánticas de los mismos definidas en la ontología de dominio considerada.

- Otro de los objetivos perseguidos en nuestro trabajo es combatir uno de los mayores problemas que han afectado a los sistemas recomendadores desde sus orígenes, aún abierto en el estado del arte: la deficiente calidad de las recomendaciones a usuarios nuevos en el sistema, debido a la falta de información acerca de sus preferencias. Aunque en menor medida que a los sistemas recomendadores tradicionales, éste es un problema que también afecta a los enfoques semánticos —como el adoptado en esta tesis—. En efecto, la capacidad de nuestro sistema para descubrir nuevo conocimiento acerca de los usuarios se ve reducida cuanto menor sea la información que tengamos sobre ellos. Por tanto, para mitigar este problema, es necesario incluir en el proceso de recomendación información externa al sistema.

  La utilización de estrategias de recomendación colaborativas basadas en confianza, en las que las sugerencias a un usuario dependen únicamente de las preferencias de los individuos en los que confía (al margen de la similitud entre ellos), se ha perfilado como una posible solución en la literatura. Sin embargo, tal como ya hemos comentado, las propuestas actuales se ven limitadas por la necesidad de obtener dicha información de forma explícita.

  Dada la masiva adopción de las tecnologías Web 2.0 en la actualidad, es razonable suponer que gran parte de los usuarios de nuestro sistema recomendador estarán dados de alta en una o más redes sociales. Teniendo en cuenta esta premisa, la metodología presentada en la tesis incluye una estrategia que obtiene relaciones de confianza *personal* entre usuarios a partir de la información disponible en los perfiles sociales a los que tiene acceso. En particular, la estrategia desarrollada toma como base los enlaces existentes en las redes sociales —que relacionan a cada usuario con sus *contactos*— y cuantifica los correspondientes valores de confianza mediante el análisis del historial de interacciones entre los individuos y las configuraciones de privacidad que han especificado en sus respectivos perfiles.

- El objetivo final de nuestro trabajo es la combinación de los distintos tipos de información inferida acerca de los usuarios de la forma más apropiada en cada caso, en función de la evolución de los usuarios en el sistema. De este modo, cuando un usuario acaba de darse de alta en el sistema y, en consecuencia, la información disponible sobre sus intereses es insuficiente, las sugerencias ofrecidas dependerán de las preferencias de los individuos del círculo social de dicho usuario que obtienen los mayores valores de confianza personal. A medida que el sistema adquiere más información acerca de los nuevos usuarios, se vale de las estrategias semánticas de nuestra metodología para seleccionar el grupo de usuarios afines más conveniente —respecto al producto considerado en cada recomendación—, y así obtener sugerencias de mayor calidad.

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. En particular, se establece una clasificación de los recomendadores propuestos en la literatura, revisando las técnicas más extendidas, sus ventajas y limitaciones, así como los nuevos enfoques de personalización —haciendo especial hincapié en los planteamientos semánticos—. Además, se analizan los métodos propuestos en el área del modelado de usuarios.

- Parte III: Antes de abundar en los detalles de la metodología desarrollada, se describen (en el Capítulo 3) los dos elementos clave del marco de razonamiento semántico en nuestro enfoque: la ontología de dominio considerada y los perfiles de usuario, modelados en base al conocimiento formalizado en la ontología.

  Las bases teóricas y los detalles algorítmicos de las estrategias incluidas en la metodología se presentan en los siguientes capítulos. En concreto, en el Capítulo 4 se describe la estrategia propuesta para combatir el problema asociado a la utilización de técnicas de filtrado colaborativo tradicionales en dominios de aplicación heterogéneos. Después, el Capítulo 5 expone nuestra estrategia para la inferencia de relaciones de confianza entre los usuarios de forma implícita, y su estructuración en una red de confianza a fin de obtener medidas de fiabilidad de los usuarios en términos generales. Posteriormente, en el Capítulo 6 se describe la estrategia para la cuantificación de la experiencia práctica de los usuarios y su incorporación en el proceso de recomendación colaborativo. Por último, el Capítulo 7 presenta la estrategia propuesta para la obtención de relaciones de confianza personal a partir de la información disponible en redes sociales, y además, se detalla en él la técnica utilizada para combinar las distintas contribuciones de la forma más apropiada en cada caso, en función de la evolución de los usuarios en el sistema recomendador.

- Parte IV: Finalmente, en el capítulo 8 se describen las principales conclusiones extraídas del trabajo de esta tesis y se apuntan algunas líneas de investigación futura.

Parte II

Estado del arte

Capítulo 2

Estado del arte

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, actualizarlos a partir de la realimentación proporcionada ante una sugerencia del sistema y asegurar su adaptación a posibles cambios en las preferencias de éstos, (ii) los métodos de filtrado de información adoptados en este tipo de herramientas, (iii) problemas abiertos en los sistemas recomendadores colaborativos y (iv) nuevos enfoques de personalización.

2.1 Introducción

Los sistemas recomendadores surgen como respuesta al problema de *sobrecarga de información* presente en numerosos dominios de aplicación (e.g., el comercio electrónico, la Televisión Digital, la educación a distancia mediante medios telemáticos, etc.), con el fin de ayudar a los usuarios a distinguir la información relevante para ellos de la que no lo es, ofreciéndoles sugerencias de forma individualizada. En particular, el objetivo perseguido por estas herramientas de personalización es seleccionar, de entre la miríada de productos disponibles (e.g., productos comerciales en el dominio del comercio electrónico, programas de televisión en el dominio de la Televisión Digital, cursos educativos en plataformas de educación a distancia, etc.), aquellos que mejor encajen con los intereses de cada usuario.

La necesidad de herramientas de personalización se ha manifestado claramente desde hace ya mucho tiempo, de ahí la gran diversidad de enfoques que han ido surgiendo en el campo de los sistemas recomendadores desde su aparición [1, 100]. En este capítulo establecemos una clasificación completa de los mismos, en base a dos criterios:

- Comenzaremos nuestra revisión del estado del arte centrándonos en las técnicas de *modelado de usuarios* propuestas en la literatura.
- A continuación, atenderemos a las técnicas de *filtrado de información* empleadas por los sistemas, esto es, las diferentes estrategias utilizadas para identificar qué productos pueden interesar a cada usuario. Destacaremos las técnicas más relevantes, así como sus ventajas y limitaciones.

Una vez finalizada dicha clasificación, identificamos los problemas que actualmen-

te despiertan el interés de la comunidad investigadora en este campo y revisamos nuevos enfoques de personalización con un gran potencial: la incorporación de técnicas tradicionalmente vinculadas a la Web Semántica en el proceso de recomendación y la explotación de tecnologías Web 2.0.

2.2 Modelado de usuarios

Los perfiles de los usuarios son elementos cruciales en un sistema recomendador, dado que permiten conocer las preferencias o necesidades de los usuarios en que se basan las sugerencias elaboradas y, por tanto, de la información registrada en ellos depende en gran medida el resultado de las recomendaciones.

En esta sección analizamos diferentes métodos tanto para generar los perfiles de los usuarios como para mantenerlos a lo largo del ciclo de vida del recomendador, entre los que destacamos:

a) mecanismos de representación de perfiles de usuarios,

b) métodos para la creación de perfiles iniciales,

c) técnicas para que los usuarios puedan proporcionar al sistema realimentación de relevancia sobre la precisión de las sugerencias ofrecidas, y

d) los mecanismos necesarios para asegurar que los perfiles se adaptan a los posibles cambios que se produzcan en las preferencias de los usuarios, garantizando con ello la calidad de las recomendaciones.

2.2.1 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,
- modelos basados en clasificadores automáticos, y
- modelos basados en una ontología de dominio.

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], uno de los recomendadores más exitosos y populares en el campo del comercio electrónico. Otro enfoque similar ha sido adoptado en el sistema WebSell [28], 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 [98]. Por último, destacar el sistema Tapestry [47], encargado de filtrar de forma automática los correos electrónicos recibidos por el usuario, basándose en los mensajes que éste ha aceptado antes y en sus anotaciones personales.

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 los usuarios 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 [24], una herramienta que recupera aquellos documentos que son relevantes para las consultas formuladas por cada usuario. 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 [104], Let's Browse [77], WebWatcher [8, 60], Letizia [76] y Labour [125], o los sistemas de entrega de noticias personalizadas NewsWeeder [71], NewT [128], Beehive [57] o Anatagonomy [63]. 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 [91], Nakif [42], Movielens [93], Recommendation Explorer [37] y Virtualrev's [135].

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 [56] y los sistemas Ringo [126, 127] y Smart Radio [52, 53], que ofrecen listas de música personalizadas de acuerdo a los gustos musicales de los usuarios.

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 [66].

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 [68], 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, el sistema Blu-IS descrito en [15] modela las preferencias de los usuarios complementando las características demográficas con dos elementos adicionales: información contextual, que registra datos como la localización del usuario, la franja horaria a la que prefiere ver televisión, su estado anímico o el número de personas que le acompañan; y un historial de visionado, que contiene una relación de los programas que han interesado al usuario en el pasado.

Redes semánticas ponderadas

Las redes semánticas fueron propuestas por Quillian en el año 1968 [113]. Estas estructuras se asocian con los orígenes de la *representación de conocimiento*, una disciplina vinculada al campo de la Inteligencia Artificial y, por tanto, estrechamente relacionada con el dominio de la Web Semántica. Según la propia definición de Quillian, las redes semánticas son sistemas que permiten expresar conocimiento sirviéndose, principalmente, de conceptos, propiedades y relaciones jerárquicas.

Aunque originariamente las redes semánticas se organizan y estructuran de forma heterogénea (careciendo de un lenguaje formal adoptado de forma generalizada entre la comunidad investigadora), 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 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 [134] 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. Un enfoque similar ha sido utilizado en el recomendador ifWeb [94, 9].

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 [120] representa los perfiles de los usuarios 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.

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, 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**. Una red Bayesiana es un grafo acíclico dirigido [31] en el

que los nodos representan variables proposicionales y los arcos identifican dependencias causales entre ellos, de forma que un arco entre un nodo origen (llamado nodo padre) y un nodo destino (nodo hijo) en la red significa que el último depende del primero. Así, el valor de cada nodo es función de los valores de sus nodos padres según una tabla de probabilidades condicionales, calculadas mediante el teorema de Bayes [59].

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 es interesante para un usuario, basándose en las relaciones probabilísticas existentes entre los atributos de dicho producto y las preferencias representadas en su red Bayesiana.

- **Árboles de decisión**. Como su propio nombre indica, un árbol de decisión [114] es una estructura en forma de árbol formada por un conjunto de nodos y arcos dirigidos que los conectan entre sí. Los nodos identifican preguntas sobre las variables consideradas en el problema de clasificación, mientras que los arcos representan las respuestas a esas preguntas, esto es, los posibles valores que pueda tomar cada variable. También son utilizados como técnicas de modelado de usuarios, en cuyo caso 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.

  Entre otros, este enfoque ha sido adoptado en los sistemas InfoFinder [69, 70], que recupera documentos para los usuarios basándose en las consultas previas, y en el recomendador Web Syskill & Webert [111, 110], que combina los árboles de decisión con el método basado en vectores de características, descrito en la Sección 2.2.1 de esta misma sección.

- **Reglas de aprendizaje inductivo**. El descubrimiento de reglas de asociación mediante aprendizaje inductivo [98] se ha relacionado tradicionalmente con la realización de estudios de mercado, ya que dichas reglas identifican patrones comunes entre los datos contenidos en diferentes transacciones —del tipo "el 80 % de los usuarios que compran un producto A, también compran B y C"—.

  En la literatura se han propuesto algoritmos (como Apriori [2]) que extraen automáticamente las reglas de asociación (de la forma $X \Rightarrow Y$) a partir de un conjunto de entrenamiento, formado por varias transacciones que contienen atributos concretos. 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, como el recomendador WebSift [26], en el dominio de la Web, y PTV-Plus [129, 108], Movielens [93] y Recommender [11], en el dominio de la TV personalizada.

- **Redes neuronales**. Las redes neuronales artificiales [6] son sistemas de aprendizaje y procesamiento automático de la información, inspirados en el funcionamiento del sistema nervioso de los animales.

Gracias a sus capacidades de aprendizaje y clasificación automática, se han aplicado también en el campo de los sistemas recomendadores, como en el sistema Re:Agent [18], 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 [58], donde sus autores entrenan una red neuronal para cada usuario, con el fin de aprender y representar sus preferencias.

Modelos basados en una ontología de dominio

Los nuevos enfoques de personalización vinculados con la Web Semántica, en los que nos centraremos en la Sección 2.5.1, requieren una representación de las preferencias de los usuarios que posibilite el aprovechamiento del conocimiento de dominio en el proceso de recomendación. Como ya señalamos anteriormente, la forma más generalizada y exhaustiva de formalización de dicho conocimiento es la ontología, piedra angular de la Web Semántica, en la que se identifican los conceptos y relaciones típicas de un dominio de aplicación de forma consensuada y reutilizable. No obstante, revisamos también propuestas de modelado de usuarios que emplean una simplificación de la misma: la taxonomía —se trata de la estructura jerárquica en la que se clasifican los conceptos del dominio—.

En la primera herramienta de personalización propuesta para la Web Semántica, denominada Quickstep y presentada por Middleton en su tesis doctoral [92], la ontología utilizada modela el dominio académico, representando información sobre investigadores, publicaciones, proyectos, conferencias, intereses de investigación, etc. Esta ontología es utilizada en Quickstep para extraer conocimiento sobre las preferencias de los usuarios y construir sus perfiles personales, basándose principalmente en la jerarquía en la que se organizan las categorías temáticas sobre las que pueden versar las publicaciones de los investigadores. En particular, los perfiles son modelados como vectores que almacenan las categorías temáticas de la jerarquía que son relevantes para el usuario, inferidas a partir de los artículos que dicho usuario haya consultado recientemente en la Web, junto con su nivel de interés concreto.

Por su parte, el sistema recomendador de contenidos de televisión AVATAR, presentado por Blanco en su tesis doctoral [16], va un paso más allá en este nuevo paradigma de modelado de usuarios, definiendo el concepto de *perfiles-ontología*. El enfoque propuesto no sólo emplea 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, incluyendo en los perfiles las características semánticas asociadas a los contenidos de televisión, así como sus respectivos niveles de interés. Para ello, dicha metodología aprovecha la capacidad de reutilización de conocimiento asociada a este tipo de conceptualizaciones, almacenando en cada perfil referencias a las entidades formalizadas en la ontología en lugar de réplicas de la misma. De este modo, es posible mantener de forma eficiente un modelo semántico completo sobre los intereses de los usuarios, ya que dichas referencias permiten acceder directamente a todas las descripciones semánticas de los contenidos de televisión registrados en cada perfil.

En el dominio del comercio electrónico, cabe destacar el enfoque descrito en [149], que será retomado en la Sección 2.5.1 por su interés a la hora de mitigar una de las limitaciones asociadas a los sistemas recomendadores colaborativos desde su aparición. Esta aproximación, utilizada para sugerir nuevos libros a los usuarios, emplea como base de conocimiento una taxonomía en la que se identifican las diferentes categorías (llamadas descriptores) bajo las que se pueden clasificar los libros. En ella, 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. El enfoque calcula los niveles correspondientes a los descriptores más específicos, y después los propaga por la jerarquía hasta alcanzar el nodo raíz.

2.2.2 Generación del perfil inicial

Una vez acordado el modelo de representación de las preferencias de los usuarios, es necesario recurrir a mecanismos que permitan inicializar sus perfiles personales antes de 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.

Reconocimiento automático de preferencias

Algunos recomendadores no necesitan crear un perfil inicial para cada usuario, 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 [28], Fab [10] y Webmate [24].

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.

Estereotipos

Esta técnica, propuesta por Elaine Rich en el año 1989 [119], 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 2.2.1.

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 en el sistema LifeStyle Finder [68], 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 [7], 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.).

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 2.2.1.

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 [71], ACR News [98], PSUN [132] y NewT [128], que selecciona noticias personalizadas para sus usuarios, la herramienta de filtrado de correos electrónicos Re:Agent [18], y los recomendadores Ringo [126] y Movielens [93], vinculados al dominio de la música y la TV personalizada, respectivamente.

2.2.3 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 con el sistema, monitorizando las acciones del usuario cuando recibe recomendaciones. Por último, algunos sistemas optan por un enfoque híbrido en el que se combinan los dos modelos anteriores.

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 [144, 18]. Es obvio que este tipo de enfoques no necesitan la realimentación de relevancia; sin embargo, obligan al usuario a adoptar un rol activo que no siempre está dispuesto a asumir.

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 [24], y las herramientas de filtrado de correo electrónico SIFT Netnews [144] y NewT [128].

- 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 [104], ifWeb [94] y Syskill & Webert [111], los sistemas de filtrado de noticias personalizadas INFOrmer [120] y PSUN [132], el recomendador de música Ringo [126] y los sistemas de TV personalizada Movielens [93] y Recommender [11].

- Por último, existen sistemas en los que los usuarios proporcionan realimentación mediante comentarios textuales. Este es el caso del recomendador de noticias personalizadas Grouplens [67] y el sistema de filtrado de correo electrónico Tapestry [47].

Realimentación implícita

La realimentación de relevancia puede ser inferida por el sistema recomendador a partir de las acciones que lleva a cabo el usuario [96].

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: los enlaces por los que éstos han navegado [76, 97, 26, 98], su historial de consumo (por ejemplo, productos comprados mediante comercio electrónico) [4, 23, 68], el tiempo empleado en ver un programa de TV o en consultar una determinada página Web [103, 67, 66, 63], 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 [62], marcar una página Web, minimizar la ventana en la que ésta se visualiza [63], ordenar la grabación de un programa de TV, avanzar, retroceder o detener su reproducción [151, 7], etc.

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. 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 [68], Anatagonomy [63], Tapestry [47], Grouplens [67], y los recomendadores de comercio electrónico CDNow [23] y Amazon [4].

2.2.4 Técnicas para la adaptación del perfil de usuario

Las preferencias de los usuarios 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.

Actualización manual por parte de los usuarios

Tal como adelantamos en las secciones previas, algunos sistemas exigen que los usuarios modifiquen sus propios perfiles, incorporando así los posibles cambios que hayan sufrido sus preferencias desde la última interacción con el recomendador [144, 18]. 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 la cantidad de preferencias registradas en el perfil es excesiva.

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, 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 [57], Anatagonomy [63], CDNow [23], infoFinder [69], Let's Browse [77], Syskill & Weber [111], INFOrmer [120], LifeStylefinder [68], WebWatcher [60], Letizia [76], Bellcore Video Recommender [56], Recommender [11], PTV [27], PTVPlus [129], TV Recommender Show [151], PPG [7], Movielens [93] y Moviefinder [105], entre otros.

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* [139].

Dicha función asigna un peso a cada producto definido en los perfiles de los usuarios, 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 [134] e ifWeb [94], y en el sistema de filtrado de documentos LaboUr [125].

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 [95], de forma que el sistema olvida todas aquellas preferencias que quedan fuera de la misma [82].

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) [141]. Así, cuando se reduce la precisión, el sistema estrecha la ventana temporal, eliminando con ello el efecto negativo de aquellas preferencias de los usuarios que hayan quedado obsoletas.

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, 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 [132] y NewT [128], y en los recomendadores Web Fab [10] y Amalthea [104].

2.3 Métodos de filtrado de información

Los métodos de filtrado de información son los mecanismos empleados por los sistemas recomendadores para seleccionar los productos que mejor se ajustan a las preferencias o necesidades de cada usuario. Tal como veremos a lo largo de la sección, en la literatura se puede apreciar una evolución en la utilización los diferentes métodos de filtrado: del filtrado demográfico a los métodos basados en contenido, y de estos últimos al filtrado colaborativo, siendo los métodos basados en contenido mayormente relegados a formar parte de los enfoques híbridos que combinan ambas estrategias.

En adelante, para referirnos a aquellos productos que el sistema debe decidir si son o no adecuados para cada usuario, emplearemos el término de *productos objetivo*.

2.3.1 Filtrado demográfico

El filtrado demográfico emplea las características personales de los usuarios (edad, sexo, estado civil, ocupación profesional, 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 [68]. 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 2.2.1 y 2.2.2). 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.

2.3.2 Métodos basados en contenido

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

Son varias las técnicas utilizadas para realizar esta comparación:

- **Comparación sintáctica basada en palabras claves**. Este método consiste en comparar sintácticamente las características definidas en el perfil del usuario, 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 en diferentes herramientas de personalización Web, como los buscadores de Internet convencionales, y los sistemas ifWeb [9] y SiteIf [134].

- **Categorización de texto**. 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. En su aplicación a sistemas recomendadores basados en contenido, los enfoques propuestos en la literatura procesan las descripciones textuales de las preferencias de los usuarios, identificando las categorías genéricas a las que pertenecen. En consecuencia, un producto objetivo es recomendado al usuario únicamente 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 [81] y las herramientas presentadas en [102] y [14], 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 2.2.1. 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 (vectores opuestos) la similitud entre ambos será -1 [122].

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

La literatura revela la adopción de este enfoque en diferentes dominios de aplicación: filtrado de noticias (Anatagonomy [63] y NewT [128]), TV personalizada (Movielens [93] y Moviefinder [105]), y recomendadores Web (Amalthea [104], Let's Browse [77], Letizia [76], Webmate [24], Fab [10] y WebWatcher [8]), entre otros.

- **Vecinos más cercanos**. Los algoritmos que utilizan este enfoque calculan la distancia entre los productos objetivo y las preferencias del usuario, 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. Dependiendo del mecanismo elegido para representar los perfiles de los usuarios (ver Sección 2.2.1), 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 [125], WebSell [28], Re:Agent [18], Syskill & Webert [110] y Entree [21].

- **Clasificación automática**. En estos 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 2.2.1, 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. Como resultado, los productos recomendados son aquellos que comparten *exactamente* las características que interesan a dicho usuario.

  Encontramos en la literatura diversas propuestas que siguen este enfoque, como el sistema de filtrado de correo electrónico Re:Agent [18], que emplea una red neuronal para clasificar los mensajes recibidos en dos carpetas etiquetadas como *trabajo* y *otras*; el recomendador Syskill & Webert [110], 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 [11] y Movielens [93], que clasifican películas mediante reglas de aprendizaje inductivo.

Si bien, por su propia naturaleza, las estrategias basadas en contenido son muy precisas, adolecen de varias limitaciones:

- El inconveniente más significativo de estas estrategias es que suelen recomendar productos excesivamente parecidos a los que el usuario ya conoce, e incluso, demasiado similares entre sí [131]. 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 [130, 116, 19]. Estas propuestas tratan de diversificar las recomendaciones ofrecidas seleccionando un conjunto de productos que no sólo son parecidos a los que interesan al usuario considerado, sino que además son diferentes entre sí. Sin embargo, 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 de dicho usuario.

- Otra limitación 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 *cold start users*, un problema relacionado con la llegada de nuevos usuarios al sistema. En este escenario, el recomendador dispone, generalmente, de muy poco conocimiento sobre sus preferencias personales, lo que se traduce en una baja precisión en las recomendaciones ofrecidas.

Tal como veremos a continuación, los enfoques colaborativos, por su propia naturaleza, permiten superar la falta de diversidad asociada a los métodos basados en contenido —de ahí su proliferación en la literatura—.

2.3.3 Filtrado colaborativo

A diferencia de los métodos basados en contenido, a la hora de ofrecer una recomendación a un usuario, el filtrado colaborativo no considera únicamente sus preferencias personales, sino también las de un grupo de usuarios con intereses similares a los suyos, llamado en adelante *vecindario*. Por su parte, emplearemos el término *usuario activo* para referirnos al individuo que recibe las sugerencias personalizadas del recomendador en un momento dado.

Se han definido dos técnicas diferentes dentro de las estrategias colaborativas:

- *Filtrado colaborativo basado en usuario*. Este modelo sigue una intuición evidente: los productos que gustan a individuos con preferencias similares a uno

dado serán también apreciadas por éste. Esto es, sugiere a cada usuario activo aquellos productos que han interesado a sus vecinos. Para formar el vecindario, la estrategia considera que dos usuarios tienen preferencias similares si han clasificado los mismos productos en sus perfiles y les han asignado niveles de interés parecidos.

- *Filtrado colaborativo basado en ítem*. Mediante este método, un producto es recomendado a un usuario activo si es considerado vecino de los definidos en su perfil personal. En este escenario, dos productos son vecinos si los usuarios que han clasificado uno de ellos tienden a clasificar el otro, asignándole niveles de interés parecidos. Dicho de otro modo, este modelo extrae, de entre todos los productos registrados en los perfiles de los usuarios del sistema, aquellos que aparecen, junto a otros que han interesado al usuario activo en el pasado (y con valoraciones similares), en un número significativo de perfiles.

Si bien esta última 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 [123], es de difícil aplicación en dominios con una gran cantidad de productos susceptibles de ser recomendados.

En cualquier caso, el objetivo es predecir el nivel de interés del usuario activo en relación a un producto objetivo dado en base a su vecindario y, en consecuencia, sugerirlo (o no) a dicho usuario. Por tanto, se identifican dos fases en un proceso de recomendación colaborativo:

1. *Creación del vecindario*: En los enfoques colaborativos basados en usuario, esta primera fase selecciona a usuarios cuyas preferencias son similares a las del usuario activo. Por el contrario, los enfoques basados en ítem identifican productos registrados en el perfil del usuario *vecinos* al producto objetivo, según el criterio explicado anteriormente.

2. *Predicción*: 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 ítem considera el nivel de interés del usuario activo en relación a aquellos productos de su perfil vecinos al objetivo.

Fase 1: Creación del vecindario

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 ítem.

- **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 ítem).

  En los enfoques de filtrado colaborativo basado en usuario, cada perfil se representa mediante un vector, cuyas componentes son las valoraciones dadas por el

usuario 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.

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

En esta expresión, $\bar{A}$ 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)[1].

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, los enfoques colaborativos basados en ítem, en lugar de computar la similitud entre dos usuarios concretos, la calculan entre dos productos (ítems). 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 [123]. 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 [56] y Movielens [93], los sistemas de filtrado de noticias personalizadas GroupLens [67] y NewsWeeder [71], los recomendadores de listas musicales Ringo [126] y SmartRadio [52] y los sistemas Web Fab [10], WebWatcher [60] y WebSell [28].

- **Clustering**: Mediante el método de estereotipos introducido por Rich, descrito en la Sección 2.2.2, 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, dado 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 [107]. 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

[1] Esta modificación tiene el propósito de evitar que la medida de similitud resultante se vea distorsionada cuando los usuarios no usan el mismo rango de valores en sus clasificaciones.

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 [20], 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 2.2.1). Típicamente, este tipo de enfoques representan los datos de entrada sobre los que trabaja el clasificador mediante las matrices descritas en la Sección 2.2.1, 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 [20, 13, 11, 48, 99, 79, 30]. El trabajo presentado en [20] adopta un modelo basado en redes Bayesianas, el enfoque descrito en [13] 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.

Una vez medida la similitud entre las preferencias del usuario activo y los usuarios candidatos a vecinos, sólo falta decidir qué candidatos forman parte del vecindario y cuáles no. 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 candidatos cuya correlación con el usuario activo supere un cierto umbral, podrán formar parte de su vecindario. De forma análoga, los enfoques basados en ítem 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 ítem), este método selecciona los n candidatos más parecidos al usuario 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 [48, 150].

- **Centroide**: En primer lugar, este enfoque selecciona el candidato más cercano al usuario activo y calcula su centroide [124]. A continuación, incluye otros candidatos en el vecindario del usuario 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 ítem para comparar los productos en lugar de los usuarios que los consumen.

Fase 2: 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 ítem 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 ítem 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 [27, 108] 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 2.2.1). 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.

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

donde:

- M es el tamaño del vecindario de U,
- N_k es el k-ésimo vecino de U,
- $sim(U, N_k)$ es la similitud (correlación) entre las preferencias de los usuarios U y N_k, y
- $R_{N_k}(i)$ es el nivel de interés del vecino N_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 [67] es uno de los sistemas que adopta este enfoque. En el dominio de la TV personalizada, algunos ejemplos representativos son MovieLens [93], Moviefinder™ [105], TiVo™ [136], ReplayTV™ [117] y TV Scout [12].

La media ponderada de clasificaciones también puede ser utilizada en los sistemas colaborativos basados en ítem. 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.

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

donde:

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

2.3.4 Enfoques híbridos

Antes de incidir en los problemas que actualmente despiertan el interés de la comunidad investigadora, en el campo de los sistemas recomendadores colaborativos, cabe destacar los enfoques que combinan varias de las estrategias de filtrado presentadas a

lo largo de esta sección, a fin de aunar sus ventajas y mitigar sus deficiencias. Son 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 [109]. Este autor propone calcular la similitud entre los usuarios del sistema utilizando tanto las descripciones 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 [23] y WebSell [28] (comercio electrónico); Fab [10], WebWatcher [8] y LaboUr [125] (Web personalizada); Anatagonomy [63] y NewsWeeder [71] (filtrado de noticias); y Recommender [11], PTV [27], PTVPlus [129] y Nakif [42] (TV personalizada).

Para una revisión exhaustiva de los enfoques híbridos se recomienda la consulta de [21], donde Burke define hasta siete métodos diferentes para combinar estrategias de personalización de diferente naturaleza.

2.4 Problemas abiertos en los sistemas recomendadores colaborativos

Tal como adelantábamos en la sección previa, revisamos a continuación las principales limitaciones identificadas en los sistemas recomendadores colaborativos actuales, así como las propuestas existentes en la literatura para combatirlos.

2.4.1 Problema de falsos vecindarios

La utilización de sistemas colaborativos en dominios de aplicación caracterizados por una gran heterogeneidad en la naturaleza de sus productos —como puede ser el comercio electrónico, donde el producto objetivo de la recomendación puede variar desde un libro a un coche o un producto alimenticio— revela una importante carencia, que menoscaba la calidad de las sugerencias ofrecidas a los usuarios. Se trata del problema de *falsos vecindarios*.

Tal problema es debido al carácter invariable de los procesos de selección de vecindarios descritos en la Sección 2.3.3. En efecto, los enfoques tradicionales forman un único vecindario para cada usuario activo, considerando en el proceso de selección los niveles de interés asignados a *todos* los productos incluidos en los perfiles de los usuarios. En consecuencia, la decisión de recomendar (o no) al usuario activo un cierto producto objetivo está basada en las preferencias de un mismo vecindario, independientemente del tipo de producto considerado.

En dominios de aplicación donde existe una gran diversidad de productos disponibles, dichas técnicas son propensas a una selección contraproducente de vecinos, dado que dos usuarios pueden ser afines en un tipo de productos pero no en otros. El problema surge cuando los candidatos a vecinos tienen preferencias similares a las del usuario activo en gran cantidad de productos (por lo cual han sido seleccionados como vecinos) pero difieren en los del tipo del producto objetivo. Por ejemplo, dos usuarios

que tienen los mismos gustos en libros, coches, equipamiento deportivo y música, pero no comparten gustos gastronómicos, serán falsos vecinos siempre que el producto objetivo sea un producto alimenticio.

En particular, el considerar todos los productos incluidos en los perfiles de los usuarios, al cuantificar la similitud existente entre el usuario activo y los candidatos a vecinos, tiene dos consecuencias negativas en la creación del vecindario:

- Por un lado, son seleccionados como parte del vecindario usuarios cuyas preferencias difieren de las del usuario activo respecto al producto objetivo de ciertos procesos de recomendación —los denominados falsos vecinos—. Conforme a lo explicado anteriormente, se trata de aquellos usuarios que tienen gustos similares a los del usuario activo en gran cantidad de productos, pero no en los del tipo del producto considerado en la recomendación actual. Dado que es poco probable que dos usuarios tengan preferencias similares en productos de toda índole, es pertinente hablar de falsos vecindarios: cada uno de los usuarios que forman el vecindario único de un usuario activo será en algún momento falso vecino.

- Por otra parte, se desechan candidatos que, pese a no coincidir en la mayoría de productos, tienen intereses comunes en relación al tipo de productos considerado en según qué procesos de recomendación. En otras palabras, no se incluirán en el vecindario usuarios con un gran potencial para obtener buenos resultados en el proceso de recomendación en curso —y en todas aquellas recomendaciones de productos de esa índole—.

Este problema es identificado por Li *et al.* en [75]. En dicho trabajo los autores concluyen que, para solventar el problema, la similitud entre las preferencias del usuario activo y los candidatos a vecinos debe calcularse en base a aquellos productos relacionados con el producto considerado en cada recomendación. Esto es, el vecindario de un cierto usuario activo ha de ser diferente para productos objetivo distintos. La solución propuesta por Li *et al.* consiste en emplear en la selección de vecindario los enfoques basado en ítem y en usuario simultáneamente. El primero, para identificar qué productos son similares; y el segundo, para calcular medidas de similitud entre usuarios respecto a los grupos de productos identificados previamente.

Encontramos en la literatura propuestas alternativas, como la presentada en [112], que trata de mejorar la calidad de las recomendaciones separando los productos en grupos más pequeños, con el objetivo de minimizar la varianza de las similitudes entre usuarios dentro de cada grupo. Sin embargo, este enfoque adolece de las limitaciones inherentes a los métodos de *clustering*, tales como la dependencia de los grupos de productos resultantes a la agrupación inicial y al orden de productos considerado.

En cualquier caso, la mayor limitación de estos enfoques radica en su naturaleza sintáctica. En efecto, las técnicas empleadas, más o menos sofisticadas, están basadas en la comparación de cadenas de texto y, en consecuencia, únicamente son capaces de detectar similitudes entre productos identificados exactamente con los mismos términos. La incorporación de la semántica de los productos, formalizada en una ontología de dominio, posibilitaría la comparación de las preferencias de los usuarios y productos disponibles de manera flexible y significativa, gracias a la clasificación jerárquica de los mismos y sus características asociadas. Además, sería posible aplicar técnicas de razonamiento semántico sobre dicha ontología, capaces de descubrir nuevo conocimiento acerca de los intereses de los usuarios y las similitudes existentes entre ellos.

2.4.2 Verificación de la fiabilidad de los usuarios

Al fundamentar el proceso de formación de vecindarios únicamente en las preferencias incluidas en los perfiles de los usuarios, las estrategias tradicionales son incapaces de detectar qué información proviene de usuarios confiables y cuál no, por lo que los resultados de la recomendación (que dependen directamente de los candidatos a vecinos seleccionados) pueden verse distorsionados debido a la influencia de usuarios maliciosos o no fiables.

Por este motivo, los denominados enfoques *trust-aware* han crecido en popularidad en la comunidad investigadora en los últimos años. Dichas propuestas complementan (o incluso sustituyen) las estrategias de creación de vecindarios tradicionales, basadas en medidas de similitud entre los intereses de los usuarios, con mecanismos de selección que atienden a las relaciones de confianza existentes entre el usuario activo y los candidatos a vecinos. La idea es, por tanto, llevar a cabo el proceso de formación del vecindario para un cierto usuario activo considerando no sólo si los candidatos a vecinos tienen preferencias similares a las de dicho usuario, también teniendo en cuenta si son o no fiables.

Si bien ha habido discrepancias entre los investigadores en su definición básica [46], se asumen de forma generalizada en la literatura tres propiedades principales del concepto de confianza:

- **Transitividad**. Aunque la confianza no es considerada transitiva en el sentido matemático del término, se considera que hay una cierta transmisión entre personas. Esto es, si A confía en B y B confía en C, es probable que A tenga un cierto nivel de confianza en C.
- **Asimetría**. Una relación de confianza es naturalmente asimétrica: la confianza que tiene A en B no tiene por qué ser la misma que B tiene en A. Por ejemplo, A puede confiar plenamente en B, mientras que éste desconfía de A.
- **Personalización**. La confianza conlleva una opinión personal. En efecto, dos personas pueden tener opiniones muy diferentes acerca de la fiabilidad de una tercera.

En cuanto a la incorporación de modelos de confianza en el proceso de personalización, ésta suele implicar la utilización de una *red de confianza* entre los usuarios del sistema recomendador, en la que los nodos representan a los usuarios y los arcos simbolizan las relaciones de confianza (o desconfianza) existentes entre ellos. Esta estructuración en forma de red posibilita la propagación de los valores de confianza entre pares de usuarios a través de la misma [148] y, por tanto, facilita la extracción de métricas de confianza global o reputación —i.e., medidas que indican lo fiable que es considerado un usuario dado por el conjunto de la comunidad [61]—.

Enfoques *trust-aware* explícitos

La mayoría de los enfoques *trust-aware* propuestos en la literatura hasta la fecha dependen de la existencia de redes de confianza *explícita* ente usuarios, esto es, redes formadas a partir de relaciones de confianza especificadas expresamente por los usuarios. Dicho de otro modo, son estrategias que requieren que los usuarios indiquen en qué otros usuarios confían, en cuáles no, y precisen cuánto confían en ellos (o desconfían de ellos).

A efectos de evaluación, es habitual en estos enfoques la utilización de los datos de confianza registrados en la web Epinions.com [39], recientemente adquirida por el conocido sitio de subastas en línea Ebay [35] —i.e., se basan en el supuesto de que los usuarios del sistema recomendador estén dados de alta en dicha web—. Epinions es un sitio de comercio electrónico en el que los usuarios pueden evaluar y/o escribir reseñas sobre productos. Además, los usuarios pueden asignar valores de confianza a otros usuarios, en base, por ejemplo, a lo útiles y fiables que han considerado las reseñas escritas por ellos en el pasado.

Massa y Avesani [88] fueron de los pioneros en proponer el uso de redes de confianza explícita en el proceso de selección de vecindarios. Su trabajo, que emplea un modelado de usuarios basado en matrices de clasificaciones (ver Sección 2.2.1), estudia el efecto de incorporar una matriz de relaciones de confianza (que representa los valores de confianza existentes entre pares de usuarios) en la calidad de las recomendaciones. En particular, los autores contemplan la posibilidad de, bien combinar dicha matriz de confianza con la matriz de similitudes entre usuarios, o bien condicionar la selección de vecindarios únicamente a la información contenida en matriz de confianza.

Por otro lado, en [89] Massa y Avesani presentan Moletrust, un algoritmo de propagación de confianza a través de la red. Dicho algoritmo, apoyándose en la propiedad de transitividad antes mencionada, posibilita la incorporación de medidas de confianza indirecta entre usuarios en el proceso. Este enfoque es tomado como punto de partida en [145], en el que los autores buscan mejorar el enfoque basándose en la hipótesis de que las redes de confianza tienen topología *small-world*[2]

Por su parte, Golbeck [46, 45] propone un algoritmo alternativo de propagación de valores de confianza, explorando la red desde el punto de vista del usuario para el que se infieren relaciones de confianza indirecta. Este algoritmo es aplicado en FilmTrust [44], un sistema recomendador de cine asentado sobre un sitio web que combina una red social y un sistema de valoración y reseñas de películas. La estrategia de personalización colaborativa propuesta considera únicamente la red de confianza a la hora de predecir el nivel de interés de un cierto usuario activo en una película, atendiendo a las opiniones de los usuarios vinculados a éste en la red. En caso de que ninguno de estos usuarios haya valorado la película, se consideran las opiniones de los usuarios conectados a éstos (en base a las relaciones de confianza indirecta inferidas para el usuario activo), y así sucesivamente.

En general, todos los enfoques *trust-aware* explícitos propuestos en la literatura comparten una misma limitación: precisamente la necesidad de obtener dicha información de confianza de forma explícita por parte de los usuarios. Por un lado, los usuarios pueden no estar dispuestos a proporcionar tal información. Por otro lado, aun bajo el supuesto de que éstos se presten a suministrar los valores de confianza requeridos, no son aptos para ser utilizados en sistemas abiertos donde los individuos, por norma general, no se conocen los unos a los otros.

Enfoques *trust-aware* implícitos

Conscientes de la limitación de acabamos de mencionar, algunos autores optan por la incorporación de modelos de confianza en el proceso de selección de vecindarios de forma *implícita*.

[2]Dicha hipótesis asume que la mayoría de los nodos no están conectados entre sí pero hay un número reducido de saltos entre cualquier par de nodos de la red.

- Por un lado, encontramos en la literatura varios trabajos que obtienen relaciones de confianza implícita entre usuarios a partir de la similitud de sus preferencias, como [145], [101] o [80], entre otros. Estos métodos están fundamentados en la conclusión alcanzada por Ziegler y Golbeck en [147], la cual sostiene que existe una correlación entre la confianza que se establece entre dos personas y la similitud de sus intereses, cuando la red de confianza considerada está limitada a un dominio específico.

  Sin embargo, los enfoques mencionados pasan por alto una de las propiedades inherentes de la confianza: su asimetría. En efecto, al inferir relaciones de confianza a partir de medidas de similitud, dichas relaciones son inevitablemente no-direccionales, puesto que la similitud es simétrica.

- Un enfoque diferente es el presentado por O'Donovan y Smyth en [106]. La estrategia propuesta obtiene relaciones de confianza implícita a partir de los resultados obtenidos en recomendaciones previas. En particular, una relación de confianza inferida de un cierto usuario a otro dependerá de las contribuciones del segundo en recomendaciones al primero. Es decir, dicha estrategia asume que en los procesos de recomendación se establecen automáticamente relaciones implícitas entre el usuario activo y cada uno de sus vecinos, que serán positivas o negativas en función de si las participaciones del vecino considerado han sido o no exitosas. Así, si un usuario ha contribuido (como vecino) a un gran número de recomendaciones que han agradado a otro, será considerado fiable por éste.

  Sin embargo, la naturaleza sintáctica de este enfoque posibilita únicamente la definición de dos niveles de confianza: el denominado en [106] *nivel de ítem*, que permite inferir relaciones de confianza respecto a la recomendación de productos específicos (e.g., la fiabilidad de John para recomendar un Toyota Corolla); y el denominado *nivel de perfil*, que representa la confianza como un todo, sin estar condicionada a productos específicos (e.g., la fiabilidad de John).

2.4.3 Cuantificación de la experiencia de los usuarios

Igualmente relacionada con el análisis de la aptitud de los usuarios que son seleccionados para formar parte del vecindario, surge otra línea de investigación con el propósito de mejorar la calidad de las estrategias de personalización colaborativas: la incorporación de medidas de *experiencia* en el proceso de recomendación.

En efecto, valorar la experiencia[3] de los vecinos respecto al producto objetivo de una recomendación, teniendo en mayor consideración las opiniones de aquellos vecinos más expertos en el cálculo de predicción del interés del usuario activo (ver Sección 2.3.3), puede mejorar la calidad de las recomendaciones.

Sin embargo, el mayor problema de este tipo de enfoques es la falta de conocimiento acerca de las aptitudes de los usuarios en un sistema recomendador, puesto que la información disponible suele estar limitada al historial de consumo registrado en sus perfiles. Es por este motivo que la mayoría de dichos trabajos dependen de información explícita adicional para estimar la experiencia de los usuarios.

- Así, por ejemplo, el trabajo presentado en [22] depende de las calificaciones explícitas y subjetivas que dan los usuarios del sitio web Epinions (ver Sección 2.4.2)

[3] Entendida como "la aptitud o conocimiento que una persona tiene en un cierto campo".

a las reseñas sobre productos escritas por otros usuarios, para inferir la experiencia de estos últimos.

- Por su parte, los autores de [143] basan su métrica de experiencia en la calidad de los ítems valorados por los usuarios en sus perfiles, considerando a un usuario más experto cuantos más ítems de calidad conoce. La mayor limitación de este enfoque radica en cómo determinar la mencionada calidad.
- En cuanto a la incorporación de conocimiento del dominio de aplicación en la cuantificación de la experiencia de los usuarios, destacamos el trabajo de Dima *et al.* [32]. Si bien está limitado a ser utilizado en un recomendador de películas, su enfoque define la experiencia de un usuario en base al conocimiento que éste tiene acerca de ciertos géneros, actores o directores.

2.4.4 Problema de *cold start users*

El problema de *cold start users* es una limitación que ha afectado a los sistemas recomendadores desde su aparición, y que sigue abierto en el estado del arte.

Como su propio nombre indica, está relacionado con la llegada de nuevos usuarios al sistema. En este escenario, el recomendador dispone, típicamente, de muy poco conocimiento sobre sus preferencias personales. En los sistemas recomendadores colaborativos, esta situación dificulta enormemente la formación del vecindario asociado a un usuario activo que acaba de darse de alta en el sistema. Como consecuencia, se ve comprometida la calidad de las sugerencias ofrecidas a dicho usuario.

Para combatir este problema, es necesario incorporar en el proceso de personalización información acerca de los usuarios externa al sistema recomendador. A tal fin, los enfoques *trust-aware* explícitos que basan la creación del vecindario únicamente en la red de confianza entre los usuarios se han propuesto como posible solución. Se trata de estrategias que seleccionan como vecinos a individuos en los que confía el usuario, de forma directa o indirecta, al margen de la similitud entre ellos.

Sin embargo, tal como comentamos en la Sección 2.4.2, los enfoques existentes están limitados a aquellos casos en que los usuarios estén dispuestos a proporcionar la información de confianza requerida, especificando explícitamente cuánto confían en otros usuarios.

2.4.5 *Sparsity problem*

Esta limitación está relacionada con la noción de similitud adoptada, en los enfoques colaborativos tradicionales, en el proceso de selección de vecindarios —recordemos que dos usuarios son considerados similares si tienen registrados en sus perfiles los *mismos* productos, valorados con niveles de interés similares—.

Los efectos del *sparsity problem* se manifiestan cuando los recomendadores colaborativos son aplicados en dominios en los que existe una gran cantidad de productos, y son más perjudiciales cuanto mayor es el número de productos disponibles en el sistema. Ante tal diversidad, es menos probable que dos perfiles contengan exactamente los mismos productos y, por consiguiente, es más difícil encontrar vecinos para un usuario activo dado.

Si bien éste es un problema que sigue llamando la atención de la comunidad investigadora en la actualidad, la incorporación de la semántica en los enfoques de perso-

nalización colaborativos permite aliviar sus efectos negativos notablemente, tal como veremos en la próxima sección. En efecto, la explotación de las descripciones semánticas asociadas a los productos (formalizadas en una ontología de dominio) posibilita la detección de similitudes en las preferencias de dos usuarios aun cuando sus perfiles no contienen productos identificados *exactamente* con los mismos términos.

2.5 Nuevos enfoques de personalización

En esta sección, destacamos dos líneas de investigación con un gran potencial en el campo de la personalización. En primer lugar, nos centramos en los enfoques que incorporan técnicas de razonamiento, tradicionalmente vinculadas a la Web Semántica, en sistemas recomendadores. Haremos especial hincapié en este tipo de propuestas puesto que gran parte de la metodología presentada en esta tesis sigue un enfoque semántico, dado su enorme potencial para combatir muchas de las limitaciones identificadas en los sistemas recomendadores actuales. Por otro lado, revisamos enfoques que aprovechan tecnologías de la Web 2.0, las cuales se han instaurado en los últimos años como valiosas fuentes de información acerca de los usuarios.

2.5.1 Enfoques semánticos

Antes de concentrarnos en determinadas propuestas de personalización semánticas, existentes en la literatura, abundamos en los detalles de las que son el pilar fundamental de la Web Semántica y, por tanto, de dichas estrategias de recomendación semánticas: las ontologías.

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" [51]. En esta definición, convertida 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" [140]. Ésta es precisamente la noción de ontología utilizada en el dominio de la Web Semántica [133].

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.

Por su parte, la madurez alcanzada en la elaboración de ontologías se debe en gran medida a la capacidad expresiva de OWL (*Web Ontology Language*), el lenguaje definido y estandarizado por el consorcio W3C, actualmente en su versión OWL 2 [142].

Los elementos clave de una ontología 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[4].

En una ontología que formaliza el dominio de aplicación de un sistema recomendador, se distinguen dos tipos de clases:

- *Clases de productos*, utilizadas para organizar jerárquicamente las categorías y subcategorías bajo las que se pueden clasificar los productos objetivo de las recomendaciones. Dicha estructura jerárquica recibe el nombre de taxonomía.
- *Clases de atributos*, que identifican a las características semánticas de los productos, también organizadas jerárquicamente.

Asimismo, de acuerdo con esta distinción, se diferencian en ellas dos tipos de instancias: las *instancias de productos* y las *instancias de atributos* (en adelante, *atributos*), que representan en la ontología a productos específicos y sus correspondientes características semánticas concretas, respectivamente.

Tal como acabamos de adelantar, la taxonomía es la estructura jerárquica en la que se organizan las clases de productos. Se trata de una estructura en forma de árbol en la que las clases se ordenan a partir de una única clase raíz (*owl:Thing*) de forma que, a medida que se desciende por la jerarquía, las entidades son más específicas; mientras que, al ascender hacia la raíz, se encuentran clases más generales (abstractas).

Habida cuenta de su organización jerárquica, las taxonomías modelan relaciones de herencia o inclusión (*subclassOf*), según el *principio de subsunción*:

Siendo $\mathcal{T}$ una taxonomía, y C_a y C_b dos clases incluidas en ella, se dice que C_a *subsume* a C_b si C_b es subclase de C_a (es decir, si el concepto C_b *está incluido en* el concepto 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 [31]. Siendo a y b dos instancias pertenecientes a las clases C_a y C_b, respectivamente:

- **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 $\mathcal{T}$.
- **Clases hojas**: Son clases de las que no dependen otras más específicas, es decir, aquellas que carecen de subclases en la estructura jerárquica —i.e., las que ocupan el último nivel de la misma, las de mayor profundidad—.

[4] 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.

- **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 $\mathcal{T}$.

Enfoque de Middleton

El primer enfoque de personalización basado en ontologías fue presentado por Middleton en su tesis doctoral [92]. 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 una notable mejora en la calidad de las sugerencias ofrecidas, con respecto a los enfoques tradicionales.

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. Tal como mencionamos en la Sección 2.2.1 dicha ontología es utilizada para extraer conocimiento sobre las preferencias de los usuarios y construir sus perfiles personales.

Cabe destacar que las categorías incluidas en los perfiles son 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. 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.

Además del modelado de usuarios basado en ontologías, con el que contribuyó de forma relevante al estado del arte de los recomendadores semánticos, Middleton propone un mecanismo para aliviar el problema de *cold start users* (descrito en la Sección 2.4.4) en su sistema.

Dado que en la ontología del sistema se pueden consultar las publicaciones de todos los usuarios registrados en ella, cuando un usuario nuevo se registra en Quickstep (y, por tanto, la información acerca de sus preferencias es muy limitada), Middleton explora la ontología para completar sus preferencias. En concreto, se vale de propiedades como *miembroDe* o *autorDe* para identificar a aquellos usuarios que pertenezcan a su mismo grupo de investigación y que hayan firmado juntos varias publicaciones. A continuación, asume que dichos usuarios tendrán intereses similares, por lo que utiliza las preferencias de aquellos para elaborar recomendaciones al usuario en cuestión.

Enfoque de Ziegler

Otro de los enfoques que destaca el potencial de los sistemas recomendadores semánticos es el descrito en [150]. Si bien el enfoque es suficientemente flexible para ser extendido a aplicaciones de diversa naturaleza, en [150] 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 el *sparsity problem* asociado a los enfoques colaborativos tradicionales. Tal como describimos en la Sección 2.4.5, debido a la naturaleza sintáctica de los recomendadores tradicionales, el proceso de selección de vecindario (basado en detectar solapamientos entre los productos clasificados por los usuarios comparados) requiere que dos perfiles almacenen productos identificados con *exactamente* los mismos términos para así cuantificar similitud entre ellos. Es obvio que cuanto mayor es el número de productos disponibles en el recomendador, es 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 [150] 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.

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 [150] 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, [150] 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. 2.5.

$$int(d_i) = \frac{s}{|P_1| \cdot |D_{l_1}|} \tag{2.5}$$

En esta expresión, 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. 2.6. 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 ($\#sib(d_i)$).

$$int(d_j) = k \cdot \frac{int(d_i)}{1 + \#sib(d_i)} \tag{2.6}$$

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.

Enfoque de Blanco

El sistema AVATAR propuesto por Blanco en su tesis doctoral [16] es un sistema recomendador híbrido que combina una estrategia basada en contenido y una de filtrado colaborativo para sugerir contenidos de televisión.

Además de su relevante aportación al modelado de usuarios (ver Sección 2.2.1), es especialmente interesante la explotación de la ontología de dominio hecha por Blanco. Cabe destacar la utilización de técnicas de razonamiento semántico complejas en la estrategia basada en contenido propuesta en su enfoque, que permiten incorporar gran cantidad de conocimiento útil acerca de las preferencias de los usuarios al proceso de personalización.

El enfoque de Blanco, mediante las denominadas *asociaciones semánticas complejas* [5], descubre relaciones ocultas en la ontología entre las preferencias del usuario activo y el producto (contenido de TV) objetivo de la recomendación. Para descubrir dichas asociaciones es necesario analizar las propiedades formalizadas en la ontología, más concretamente, las secuencias de propiedades existentes en ella:

- Una *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.
- Asimismo, se define una *instancia de una secuencia de propiedades*, ps, como un conjunto de instancias unidas entre sí mediante propiedades.

En particular, Blanco explota dos de los tipos de asociaciones semánticas complejas identificadas en [5]. Siendo ps_1 y ps_2 dos instancias de las secuencias de propiedades PS_1 y PS_2:

- **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 posible encontrar dos secuencias PS_1 y PS_2 unidas (i.e., PS_1 y PS_2 tienen algún nodo en común) 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.

Gracias a las asociaciones semánticas complejas identificadas entre las preferencias del usuario activo y el producto objetivo de la recomendación, AVATAR combate

el problema de *sobreespecialización* típico de las estrategias de filtrado basadas en contenido.

Por otro lado, Blanco propone una métrica de *similitud semántica* entre instancias de productos que, además de atender a la estructura jerárquica de la ontología (como las propuestas de Rada [115] y Lee [73], Ganesan *et al.* [43], Resnik [118] o Lin [78]), considera las asociaciones semánticas complejas (de longitud mínima) identificadas entre ellos.

Enfoque de Ganesan *et al.*

Además de la métrica de similitud semántica basada en relaciones jerárquicas que acabamos de mencionar, el trabajo de Ganesan *et al.* en [43], contiene una importante contribución al campo de los sistemas recomendadores colaborativos. En efecto, el enfoque propuesto en [43] permite combatir eficazmente el *sparsity problem* (ver Sección 2.4.5).

Ganesan *et al.* ponen de manifiesto cómo la naturaleza sintáctica de las técnicas de selección de vecindarios tradicionales impide que sean capaces de detectar similitudes entre dos usuarios que, a pesar de que sus perfiles no contienen exactamente los mismos productos, es obvio que comparten intereses. A modo de ilustración, consideran a dos seguidores de los Beatles, cuyos perfiles contienen varios de sus CDs. Como los CDs incluidos en uno y otro perfil tienen títulos diferentes, las estrategias tradicionales no detectan ningún solapamiento en las preferencias de estos usuarios al calcular su similitud basada en coseno (recordar Ec. 2.1).

En su enfoque, Ganesan *et al.* proponen modificar la forma de calcular el producto escalar entre los vectores de interés de los usuarios comparados, incluyendo el conocimiento semántico contenido en la jerarquía de clasificación de los productos en cuestión. La idea es comparar el nivel de interés asociado a cada producto almacenado en los perfiles de dichos usuarios, considerando la similitud semántica existente entre los productos. De esta manera es posible detectar cierto nivel de similitud entre los dos usuarios a los que les gustan los Beatles, ya que, aunque sus perfiles no contengan exactamente los mismos productos (en este caso CDs), contienen productos clasificados bajo la misma clase en la ontología.

2.5.2 Enfoques vinculados a la Web 2.0

Con la explosión de aplicaciones Web 2.0 tales como blogs, foros de debate o redes sociales y profesionales, los usuarios Web han dejado de ser meros *consumidores* de información para convertirse en *productores* de información. En efecto, los usuarios participan activamente en redes sociales, suben sus fotografías personales, comparten sus sitios web favoritos, escriben blogs y comentan acerca de la información proporcionada por otros.

Cabe destacar el enorme crecimiento de las redes sociales en los últimos años, que las ha convertido en el epítome de la Web 2.0 [146]. En efecto, el número de usuarios alcanzado en la actualidad por ciertas redes sociales deja patente su popularidad: Facebook [40] ha superado los 845 millones de usuarios desde su lanzamiento en 2004, Twitter [137] los 500 millones de usuarios desde su creación en 2006, y Google+ [49] los 100 millones de usuarios desde su lanzamiento en 2011.

A través de las redes sociales, los usuarios interactúan con sus *contactos* y compar-

ten información en línea del mismo modo que lo hacen en el mundo real. Millones de usuarios dedican un gran número de horas en estos sitios diariamente, generando gran cantidad de información útil. De ahí el gran potencial de las redes sociales como nuevas fuentes de conocimiento para los sistemas recomendadores.

Cuando un usuario se da de alta en una red social, comienza creando un perfil y, a continuación, establece conexiones con otros usuarios —en un gran número de ocasiones, gente que ya conoce (por ejemplo, amigos, familia, compañeros de trabajo, etc.)—. Los perfiles sociales de los usuarios son páginas únicas, donde pueden expresar sus pensamientos, sentimientos e intereses, mostrar su lista de contactos, publicar fotografías o vídeos y etiquetar a sus contactos en ellos. Asimismo, un usuario puede visitar los perfiles de sus contactos y publicar comentarios en ellos. Además, los usuarios pueden especificar su configuración de privacidad y limitar el acceso a cada contenido publicado en sus perfiles —permitiendo a todo el mundo, únicamente a sus contactos, o sólo a algunos de sus contactos, el acceso a según qué contenidos—.

Aparte de los ya mencionados, las redes sociales pueden incluir otros servicios, como tecnologías de mensajería instantánea, secciones de música, grupos o juegos. Más aun, ciertas redes sociales proporcionan una plataforma para desarrolladores de software que permite crear aplicaciones integradas con el entorno:

- Por un lado, es posible incorporar aplicaciones en la red social, de modo que facilitan la interacción con otros usuarios que también usen dichas aplicaciones. Por ejemplo, si hablamos de videojuegos, un usuario puede jugar con sus amigos dentro de la red social.

- Por otra parte, como resultado del tremendo crecimiento y adopción del medio social, en la actualidad es además posible incorporar capacidades sociales en sitios web externos a la red social. De esta forma, los usuarios pueden compartir en sus respectivas redes sociales, por ejemplo, las noticias que están leyendo, la música que están escuchando, entradas de blogs que les han gustado, etc.

La mayoría de las propuestas encontradas en la literatura que tratan de explotar el potencial de las redes sociales están relacionadas con el *marketing viral* [34, 64, 74, 65]. Estos enfoques ven en las redes sociales la posibilidad de aprovechar un efecto equivalente al *boca-a-boca*, en línea, para la promoción de productos y marcas comerciales. De ahí que el principal objetivo de estas propuestas sea identificar qué usuarios tienen una mayor influencia en la red, de forma que la recomendación del producto en cuestión pueda llegar a un mayor número de usuarios.

No obstante, el análisis de las redes sociales tiene multitud de aplicaciones. Destacamos la propuesta de Ding *et al.* [33], que explora las reces sociales para identificar relaciones fuertes entre usuarios —en su caso, aplicadas al dominio de justicia criminal—.

2.6 Conclusiones

En este capítulo, se han revisado los diferentes enfoques propuestos en la literatura en el campo de los sistemas recomendadores (tanto en el modelado de usuarios, como en las técnicas de filtrado de la información utilizadas), haciendo especial hincapié en los nuevos enfoques de personalización semánticos surgidos en los últimos años y

apuntando el potencial de las tecnologías Web 2.0, especialmente como nuevas fuentes de información acerca de los usuarios.

Tal como comentamos en la Sección 1.3, el objetivo de esta tesis es desarrollar una serie de estrategias que combatan las limitaciones identificadas en los sistemas recomendadores colaborativos actuales, revisadas en la Sección 2.4, de forma transparente a los usuarios —i.e., sin requerir que éstos proporcionen ningún tipo de información adicional de forma expresa—.

A tal fin, nuestro trabajo adopta los enfoques de inferencia semántica más avanzados y extiende sus capacidades. En los sucesivos capítulos, abordamos la descripción de la metodología desarrollada:

- El siguiente capítulo describe los dos elementos clave del marco de razonamiento semántico que constituye el eje central de nuestro enfoque: la **ontología de dominio** y los **perfiles de usuario**. Como cabe esperar, el modelado de usuarios empleado aprovecha el conocimiento formalizado en la ontología, adaptando el paradigma de *perfiles-ontología* —introducido en la Sección 2.2.1— a las necesidades de nuestra metodología.

- A continuación, el Capítulo 4 presenta la estrategia propuesta para combatir el problema de falsos vecindarios presentado en la Sección 2.4.1. En primer lugar, se define una nueva métrica de **similitud entre usuarios**, adaptando la métrica de similitud semántica entre instancias propuesta por Blanco (ver Sección 2.5.1) a nuestras necesidades y enriqueciendo con ella el enfoque de Ganesan para aliviar el *sparsity problem* (ver Sección 2.5.1). Acto seguido, se describe el método propuesto para la construcción de los vectores de interés de los usuarios que serán empleados en dicha métrica, aprovechando la semántica formalizada en la ontología, a fin de mitigar el problema.

- Después, el Capítulo 5 expone nuestra estrategia para la inferencia de **relaciones de confianza implícita** entre los usuarios, tomando como punto de partida el enfoque propuesto por O'Donovan (introducido en la Sección 2.4.2) y explotando la ontología de dominio para obtener relaciones de mayor calidad. Las relaciones resultantes son estructuradas en una red de confianza implícita, que será utilizada para obtener medidas de **reputación** de los usuarios del sistema.

- En el Capítulo 6 se describe la estrategia propuesta para la cuantificación de medidas de **experiencia** de los usuarios, mediante la utilización de técnicas de razonamiento semántico sobre los historiales de consumo registrados en sus perfiles, y su incorporación en el proceso de recomendación colaborativo.

- Finalmente, el Capítulo 7 presenta nuestra estrategia para combatir el problema de *cold start users* descrito en la Sección 2.4.4, mediante la obtención de **relaciones de confianza personal** a partir de la información disponible en redes sociales (ver Sección 2.5.2). Además, se detalla en él la estrategia de **selección de vecindarios híbrida**, colofón de nuestra propuesta, que combina las relaciones de confianza personal, las de similitud y confianza implícita de la forma más conveniente en función de la evolución de los usuarios en el sistema.

Parte III

Estrategias

Capítulo 3

Marco de razonamiento semántico

Este capítulo presenta el marco de razonamiento semántico de la metodología desarrollada en la tesis. Tras mostrar una visión general de la misma, se detallan las características de la ontología de dominio sobre la que se asientan nuestras estrategias y, a continuación, se describe la técnica de modelado de usuarios empleada. Dicha técnica de modelado reutiliza el conocimiento formalizado en la ontología de productos comerciales a la hora de construir los perfiles de los usuarios en el sistema.

3.1 Un sistema recomendador colaborativo semántico

Tal como apuntamos en los capítulos previos, a lo largo de la tesis describiremos diferentes estrategias para superar las limitaciones identificadas en los sistemas recomendadores colaborativos actuales, revisadas en la Sección 2.4, de forma transparente a los usuarios. A tal fin, nuestro enfoque sigue una aproximación semántica, de ahí que prácticamente todas las estrategias propuestas se asienten sobre una piedra angular común: una ontología que formaliza las descripciones semánticas de los productos disponibles en el sistema.

Esta situación se puede apreciar en la Fig. 3.1, donde se muestran las interacciones de las distintas estrategias con dicha ontología, así como con los perfiles de los usuarios —elemento clave de cualquier sistema recomendador—. Nótese que la única estrategia no semántica es la de cuantificación de relaciones de confianza personal que, tal como hemos comentado anteriormente, aprovechará la información disponible en las redes sociales para combatir el problema de *cold start users*.

En la figura se describe también a qué fase del proceso de recomendación colaborativo (ver Sección 2.3.3) contribuye cada una de las estrategias. Teniendo en cuenta que la selección de vecindarios es la etapa crucial de una estrategia basada en filtrado colaborativo, puesto que determina el éxito de las sugerencias ofrecidas, no es de extrañar que gran parte de nuestra metodología esté enfocada a enriquecer dicha fase. Ya hemos adelantado previamente que el objetivo final perseguido es combinar tales métodos en una estrategia de selección de vecindarios híbrida, adaptada a la evolución de los usuarios en el sistema.

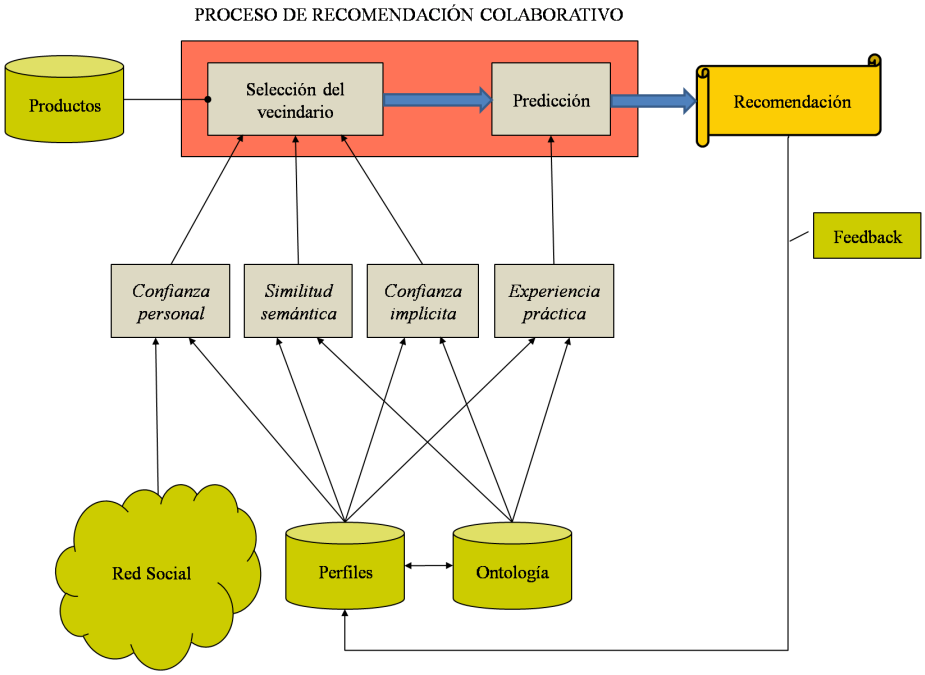


Figura 3.1: Visión general del proceso de recomendación colaborativo

Dado que el contexto elegido para ilustrar el enfoque propuesto y facilitar la comprensión del lector es el dominio del comercio electrónico, la base de conocimiento del sistema será una ontología que formalice las descripciones semánticas de los productos comerciales disponibles.

3.2 Ontología en el dominio del comercio electrónico

Si bien nuestro enfoque no está vinculado a una ontología específica —sino que es aplicable a cualquier ontología del dominio bien definida—, para ilustrar las capacidades de la metodología de razonamiento semántico desarrollada se adopta una conceptualización del dominio de comercio electrónico concreta.

La creación desde cero de dicha ontología no es necesaria, puesto que es posible extraer de forma sencilla una categorización de productos comerciales a partir de diversos estándares de clasificación de productos y servicios industriales, como UNSPSC [138], eCl@ss [36], eOTD [38] o el diccionario técnico RosettaNet [121], los cuales contienen múltiples definiciones de conceptos organizados jerárquicamente y reflejan cierto nivel de consenso en la comunidad. En nuestro caso, hemos optado por eCl@ss como punto de partida para construir la ontología de productos utilizada, por las razones de completitud, balance y mantenimiento tratadas en [55]. En concreto, hemos tomado prestados de eClassOWL (la ontología OWL de productos y servicios desarrollada por

M. Hepp [54]) algunos conceptos relativos a la categorización de productos comerciales, y hemos especificado nuevas propiedades y clases para incluir características de los productos disponibles carentes en ella —como, por ejemplo, el lugar de origen de los productos o su rango de precio—.

De esta forma, evitamos varios de los inconvenientes asociados a la creación manual de una ontología en el dominio del comercio electrónico:

- Para conseguir un alto nivel de especificación es necesario precisar una cantidad desmesurada de conceptos, a causa de la miríada de productos comerciales disponibles en el mercado.
- Es indispensable un mantenimiento temporal de la base de conocimiento. En efecto, al tratarse de un dominio caracterizado por la innovación en sus productos, precisa, con el paso del tiempo, una actualización de los conceptos que lo estructuran.
- Lleva asociado un alto coste, puesto que la construcción de este tipo de ontologías desde cero requiere significativos recursos.

Un pequeño extracto de la ontología OWL resultante se muestra en la Fig. 3.2, donde las elipses blancas representan clases de productos, los cuadrados blancos simbolizan instancias de productos, los cuadrados grises simbolizan atributos y las elipses grises representan clases de atributos.

3.2.1 Singularidades de las ontologías de productos heterogéneos

En dominios de aplicación donde existe una gran heterogeneidad de productos, como sucede en el dominio del comercio electrónico, cualquier ontología que formalice las descripciones semánticas de dichos productos posee unas características singulares, que describimos a continuación.

Múltiples jerarquías de clases de productos

La gran diversidad en la naturaleza de los productos comerciales disponibles hace que la ontología contenga múltiples jerarquías, bien diferenciadas, de clases que representan los conceptos del dominio. En la Fig. 3.2 se esbozan tres de las jerarquías de clases de productos contenidas en nuestra ontología: bajo una de ellas están clasificadas las clases correspondientes a productos alimenticios, otra contiene clases referentes a discos musicales y en la tercera se encuentran categorizados libros.

Nivel categoría

El *nivel de abstracción* de los conceptos formalizados en una ontología está relacionado con la noción de *profundidad* presentada en la Sección 2.5.1. En particular, un nivel de abstracción determinado representa a todos los conceptos (en nuestro caso, clases de productos) situados a esa profundidad concreta en la clasificación jerárquica. De este modo, el mayor nivel se corresponde con las clases hoja definidas en la ontología (las de mayor profundidad) y, por tanto, contiene a las clases que representan los conceptos más específicos del dominio. A medida que se asciende en la jerarquía, las clases representan conceptos más genéricos y el nivel disminuye, hasta alcanzar

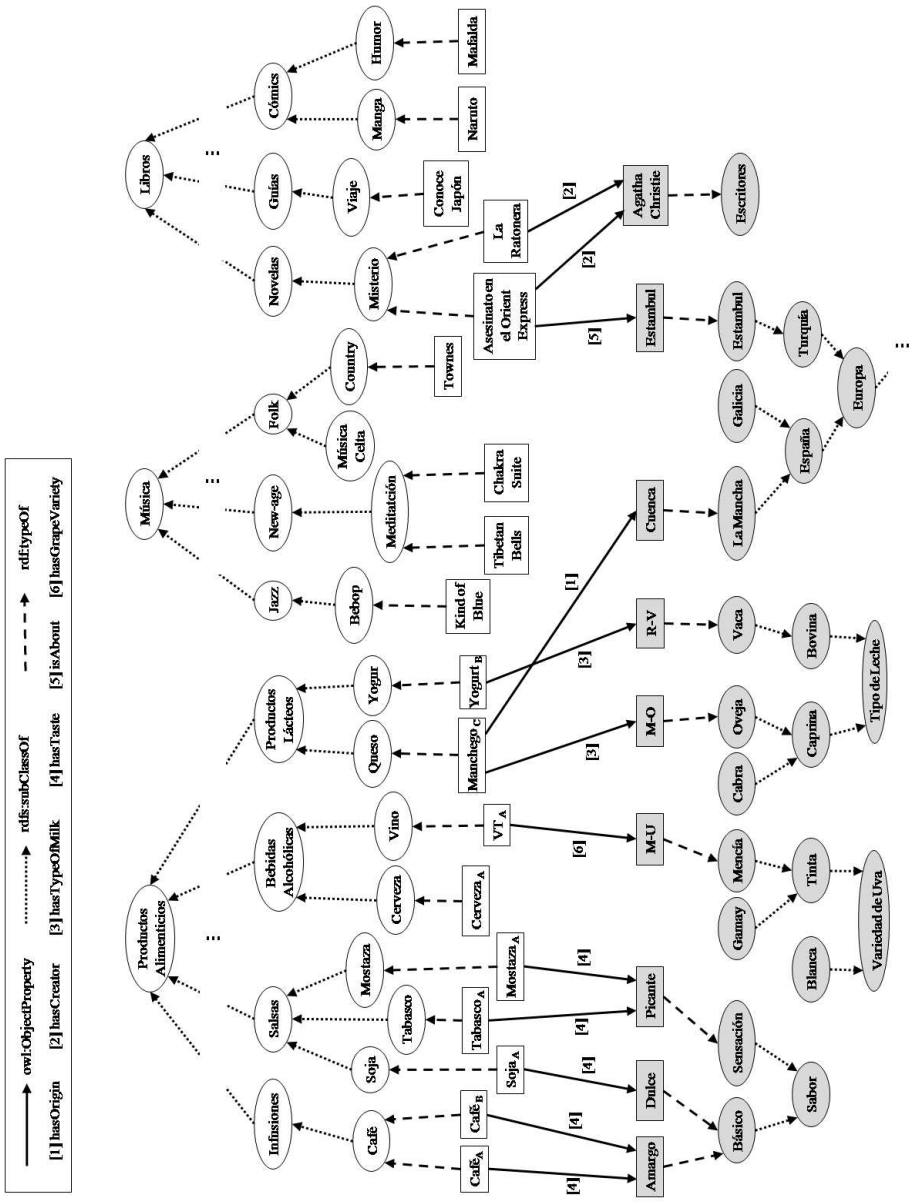


Figura 3.2: Micro extracto de nuestra ontología de productos comerciales

el nivel 0, que se corresponde con la clase raíz (*owl:Thing* —no representada en la Fig. 3.2—).

En una ontología de productos comerciales tiene especial interés el nivel 1, ya que contiene las clases que delimitan cada una de las jerarquías de clases de productos. Por ejemplo, en la Fig. 3.2, dicho nivel se corresponde con las clases *Productos Alimenticios*, *Música* y *Libros*. De ahí que, en nuestro trabajo, el nivel 1 sea denominado *nivel categoría*, puesto que incluye las clases que especifican productos de naturaleza totalmente diferente.

Atributos generales y propios

Como en toda ontología, las instancias de productos específicos (e.g., *Kind of Blue*, $Manchego_C$ o $Café_A$, en la Fig. 3.2) están relacionadas con sus respectivos atributos semánticos a través de propiedades convenientemente etiquetadas (e.g., *hasOrigin*, *hasCreator*, *hasGrapeVariety*, etc.). Asimismo, dichos atributos se representan mediante instancias clasificadas bajo diversas jerarquías de clases de atributos (e.g., *R-V*, atributo del producto $Yogur_B$ descrito en la Fig. 3.2, perteneciente a la jerarquía de atributos *Tipo de Leche*).

No obstante, dada la heterogeneidad presente en una ontología de productos comerciales, es posible diferenciar en ella dos tipos de atributos: los atributos *generales* y los atributos *propios* de una determinada jerarquía de productos. Tal como sus nombres indican, los primeros son atributos comunes a todas las categorías de productos (e.g., el lugar de origen de los productos), mientras que los segundos son característicos de un determinado tipo de productos (e.g., los atributos pertenecientes a la jerarquía *Sabor* descrita en la Fig. 3.2 únicamente estarán vinculados a productos de la categoría *Productos Alimenticios*).

3.3 Modelado de usuarios basado en perfiles-ontología

Tal como hemos adelantado previamente, la técnica de modelado adoptada en nuestro enfoque reutiliza el conocimiento formalizado en la ontología de productos comerciales al construir los perfiles de los usuarios en el sistema.

3.3.1 Perfiles-ontología

En efecto, para poder emplear técnicas de razonamiento semántico sobre las preferencias, historial de consumo y resultados de las recomendaciones previas de los usuarios —a fin de descubrir nuevo conocimiento que posibilite enriquecer el proceso de personalización automática—, la metodología desarrollada requiere una representación formal más exhaustiva que el simple almacenamiento de nombres de productos específicos y calificaciones que sobre éstos proporcionan dichos usuarios.

Con esta idea en mente, nuestro enfoque toma como punto de partida el concepto de *perfiles-ontología* presentado por Blanco en [16], introducido en la Sección 2.5.1. Dicha propuesta de modelado se basa en construir los perfiles de los usuarios a partir de la ontología de dominio considerada, aprovechando la capacidad de reutilización de conocimiento asociada a este tipo de conceptualizaciones. Se trata de mantener, junto a los niveles de interés asociados a los productos específicos registrados en el perfil de

cada usuario, referencias únicas a las instancias correspondientes a esos productos en la conceptualización OWL.

Por tanto, mediante tales referencias únicas, denominadas en nuestra metodología *item-IDs*, será posible consultar directamente en la ontología de productos comerciales todas las descripciones semánticas asociadas al producto referenciado —i.e., la clase y jerarquía bajo las que está clasificado, así como sus atributos relacionados—. Por ejemplo, consideremos el *item-ID*$_1$ registrado en los perfiles de los usuarios *A* y *B* en la Fig. 3.3. Dicho item-ID referencia al producto $Manchego_C$ y, por tanto, a una instancia de la clase *Queso* en la jerarquía de *Productos Alimenticios*, con atributos como *Cuenca* (su origen) y *M-O* (su tipo de leche).

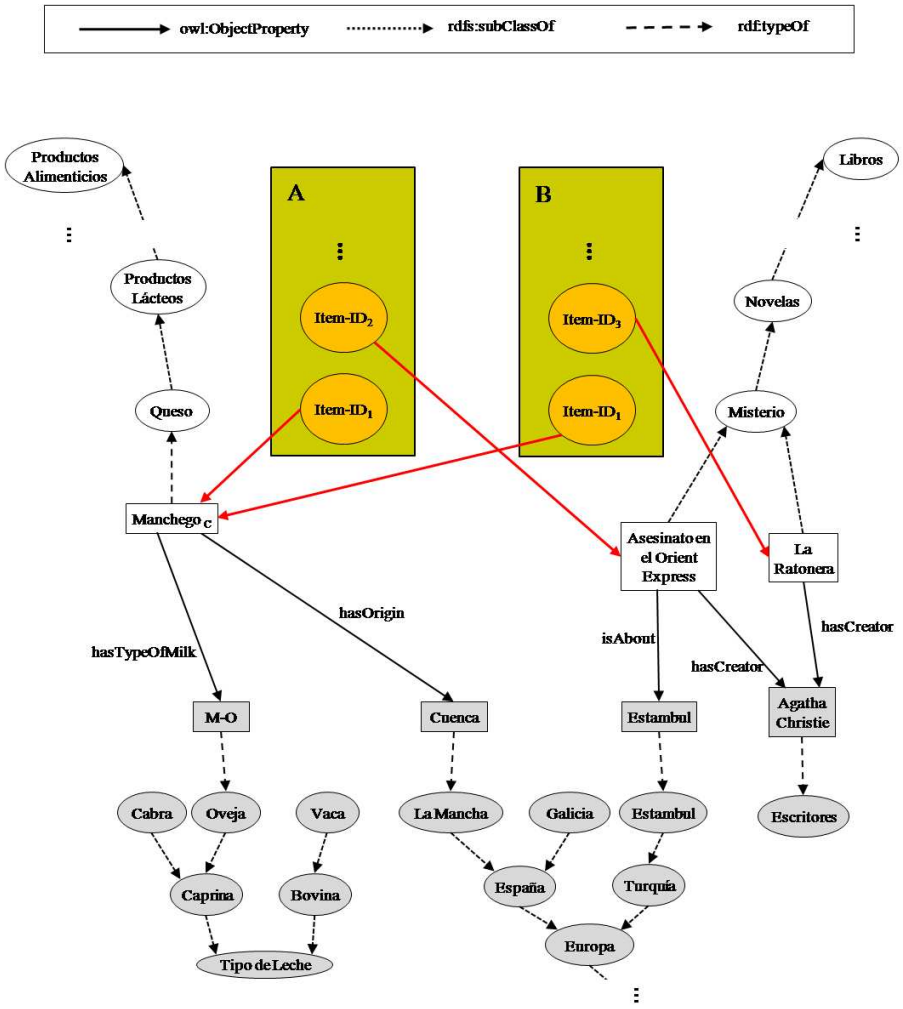


Figura 3.3: item-IDs en los perfiles de dos usuarios

3.3.2 Generación del perfil inicial

Uno de los objetivos perseguidos en nuestro trabajo es la minimización de la interacción entre el usuario y el sistema recomendador, en lo que a aporte de información explícita se refiere. Tal como describimos en la Sección 2.2.2, este propósito es incompatible con dos de los métodos de generación del perfil inicial allí revisados: la inicialización manual por parte del usuario 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 usuarios 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, cuando un usuario se da de alta en el sistema, se le ofrece la posibilidad de seleccionar algunas de sus preferencias, positivas y negativas, navegando por distintas jerarquías de productos. 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 usuarios pueden identificar fácilmente (como música, libros y sus respectivas clasificaciones más específicas). Así, por ejemplo, si a un usuario le interesan solamente los cómics, no es necesario que explore el conjunto de libros en su totalidad, sino únicamente aquellos clasificados bajo la clase *Cómics* en la jerarquía mostrada. Es obvio que este ahorro en el tiempo de búsqueda del usuario no sería posible en los métodos tradicionales basados en un conjunto de entrenamiento plano, carente de cualquier organización jerárquica.

En realidad, las jerarquías presentadas contienen únicamente el subconjunto más representativo de las clases e instancias definidas en la ontología OWL de productos comerciales, ya que la exposición de todas ellas conllevaría tiempos de navegación quizás inasumibles para muchos de los usuarios. Dicho subconjunto de productos es estratégicamente seleccionado para representar el abanico de atributos de mayor importancia —incluyendo, por ejemplo, distintos rangos de precio (en todos los tipos de productos), o diversos escritores (en la parte de libros)—. Tal selección responde a la intención de obtener la mayor cantidad de información adicional significativa acerca de las preferencias de los usuarios, tal como describiremos en la Sección 3.3.4. Asimismo, el conjunto de entrenamiento mostrado al usuario incorpora algunos atributos —generales y propios (ver Sección 3.2)— de manera que, además de productos identificados por dicho usuario como preferencias, éste pueda seleccionar de forma independiente características de su interés (por ejemplo, identificando cantantes que le agradan o desagradan, expresando su interés por productos de lujo o mostrando su afición por la comida picante).

Por otro lado, los usuarios que así lo deseen pueden inicializar sus perfiles indicando explícitamente algunos productos (que le interesen, o no), sin necesidad de explorar la jerarquía mostrada durante la fase de registro en el sistema.

De forma adicional, en caso de que el usuario utilice otras herramientas de personalización semánticas como, por ejemplo, el recomendador de contenidos de televisión AVATAR (ver Sección 2.5.1), el enfoque semántico seguido permite incorporar información útil acerca de los intereses del usuario externa a nuestro sistema —registrada en los perfiles del usuario en aquellas herramientas—. En efecto, aunque los productos clasificados en las respectivas ontologías de dominio sean distintos, pueden existir mu-

chos atributos relevantes en sendos campos, que pueden ser empleados en la generación del perfil inicial.

En cualquier caso, las preferencias sobre productos y atributos del usuario considerado se incorporan a sus perfiles, asociadas a sus correspondientes item-IDs.

3.3.3 Realimentación de relevancia

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 usuario tiene la posibilidad de aceptar o rechazar los productos seleccionados, así como de especificar, opcionalmente, calificaciones concretas en relación a cada uno de ellos. Esta realimentación de relevancia (ver Sección 2.2.3) 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 que acabamos de comentar, se añaden al perfil todos aquellos productos que el usuario haya aceptado o rechazado, asociados a sus respectivos item-IDs, así como los índices de interés correspondientes. 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 usuarios de nuestro sistema, es necesario recurrir a algún mecanismo que cuantifique el nivel de interés que tiene cada usuario en relación a los productos registrados en su perfil, distinguiendo así las preferencias positivas de las negativas. A tal fin, nuestra aproximación asigna un índice numérico a cada uno de esos productos, denominado *DOI* (*Degree Of Interest*), que será estimado, a falta de una valoración concreta especificada por el usuario, a partir de la interacción de éste con el sistema. Para ello, el sistema monitoriza las acciones del usuario posteriores a cada recomendación, observando si éste compra o no el producto sugerido, si lo pone en una *cesta de compra* o lista de artículos deseados o si busca más información acerca de ese producto (e.g., haciendo clic en un enlace proporcionado por el sistema junto a la recomendación).

3.3.4 Cálculo de los niveles de interés

Tal como acabamos de adelantar, con el propósito de asegurar los requisitos de interacción mínima entre el usuario y el sistema, nuestro enfoque no exige a los usuarios que declaren explícitamente sus niveles de interés (mediante calificaciones concretas sobre los productos recomendados), sino que en su defecto dichos valores son inferidos por el recomendador tomando como referencia las acciones que éstos llevan a cabo. 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.

Cada vez que un usuario recibe una recomendación del sistema, nuestra metodología calcula el nivel de interés del usuario (índice DOI) del producto en cuestión y lo incorpora en su perfil. Por otro lado, sirviéndose de las propiedades definidas en la ontología, la metodología propuesta también calcula los niveles de interés correspondientes a sus respectivas características semánticas.

Nivel de interés de los productos comerciales

De acuerdo con lo expuesto hasta ahora, podemos identificar dos escenarios diferentes en el cálculo del nivel de interés de los usuarios en productos:

- El primero considera la posibilidad de que los usuarios proporcionen valoraciones de productos comerciales de forma explícita —bien en la fase de registro, bien como realimentación de relevancia tras recibir recomendaciones—. En concreto, nuestro enfoque permite a los usuarios asignar valores en el rango entre -1 (máximo desagrado) y +1 (máximo agrado), o bien indicar de forma cualitativa sus preferencias, identificando si consideran dichos productos *muy interesantes*, *interesantes*, *poco interesantes* o *muy poco interesantes*. En este último caso, dichas calificaciones se traducen en los valores 1, 0.5, -0.5 y -1, respectivamente.

- En el segundo de los escenarios, 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. Para ello, definimos el índice de realimentación implícita, IF (*Implicit Feedback*), que toma un valor -0.75 si el usuario rechaza[1] el producto recomendado y 1 si éste compra el producto. Esta premisa responde a que, si un usuario compra un producto, es probable que sea porque dicho artículo le interesa. Sin embargo, si el usuario rechaza la sugerencia del sistema, no podemos asumir que el producto le desagrada completamente. En caso de que el usuario guarde el producto recomendado en una *cesta de compra* o lista de artículos deseados el valor del índice será 0.5, y si muestra interés por el producto buscando más información sobre él, 0.25.

  A fin de detectar las adquisiciones que el usuario realiza fuera del sistema recomendador (pero a raíz de las sugerencias ofrecidas por éste), se proporcionan *cupones de descuento* personalizados con las recomendaciones. De esta forma, si el usuario canjea un cupón en una tienda física, tal redención puede ser detectada por el fabricante y comunicada al sistema. Esta acción se traduce, conforme a lo expuesto anteriormente, en un índice de interés 1. Del mismo modo, el hecho de que el usuario imprima un cierto cupón indica una muestra de interés, por lo que el sistema le asigna al producto promocionado un valor 0.5 en su perfil (equivalente a guardar el producto recomendado en una *cesta de compra* o lista de artículos deseados).

Estos supuestos se recogen en la Ec. 3.1, donde $R_U(p)$ es la calificación explícita dada por el usuario U al producto p si éste valora dicho producto de forma expresa, e $IF_U(p)$ el índice de realimentación implícita inferido para U en caso contrario.

$$DOI_U(p) = \begin{cases} R_U(p) & \text{si } U \text{ valora explícitamente el producto } p \\ IF_U(p) & \text{en otro caso} \end{cases} \tag{3.1}$$

[1]Una recomendación se considera rechazada si el usuario la ignora, esto es, ni proporciona ninguna valoración explícita (de forma cuantitativa o cualitativa), ni realiza ninguna acción posterior (considerada realimentación implícita) relacionada con el producto objetivo en cuestión.

Nivel de interés de las características semánticas

La semántica, formalizada en la ontología de productos comerciales, posibilita transferir los niveles de interés de los productos, obtenidos mediante la Ec. 3.1, a sus respectivos atributos. Es decir, el DOI de cada producto almacenado en el perfil de un usuario U se propaga a través de sus propiedades. Así, por ejemplo, en la Fig. 3.2, el índice DOI calculado para *Asesinato en el Orient Express* se transfiere a sus atributos *Agatha Christie* y *Estambul*.

Teniendo en cuenta que un mismo atributo puede estar relacionado con varios productos registrados en el perfil del usuario considerado, nuestra metodología obtiene el índice DOI de un cierto atributo a como el promedio de los índices DOI de todos los productos del perfil de U que compartan dicho atributo, tal como muestra la Ec. 3.2.

$$DOI_U(a) = \frac{1}{N}\sum_{i=1}^{N} DOI_U(p_i(a)) \cdot \Delta \tag{3.2}$$

donde:

$$\Delta = \frac{N}{\#I(a)}$$

En esta expresión:

- N es el número de productos almacenados en el perfil de U que están caracterizados por el atributo a,
- $p_i(a)$ se refiere al i-ésimo de esos productos,
- y Δ es un factor cuyo valor aumenta a medida que el número de productos del perfil relacionados con ese atributo es mayor —siendo $\#I(a)$ el número total de productos caracterizados por el atributo a—.

Nótese que, de este modo, incorporamos nuevo conocimiento al perfil del usuario acerca de sus preferencias. En concreto, conforme a la Ec. 3.2, cuanto mayor sea el número de productos que contribuyen en el cálculo del DOI de un atributo, más información ofrece éste acerca de las preferencias del usuario en cuestión. Siguiendo con el ejemplo anterior, si en el perfil del usuario además de *Asesinato en el Orient Express* está registrado *La ratonera*, ambos con un índice DOI muy positivo, el DOI del atributo *Agatha Christie* (propio de la jerarquía de productos *Libros*) tomará un valor positivo, reflejando un gran interés del usuario por esa escritora —que será más significativo cuantos más libros escritos por ella almacene su perfil—.

3.3.5 Identificadores de usuario

Tal como describimos en la Sección 2.3.3, en un sistema recomendador colaborativo cualquier usuario puede adoptar dos roles diferentes, dependiendo del contexto: *usuario activo*, cuando recibe una recomendación por parte del sistema; y *vecino*, cuando es seleccionado como parte del grupo que determinará la recomendación a un cierto usuario activo.

Dado que una de las estrategias propuestas en esta tesis (detallada en el Capítulo 5) aplica técnicas de razonamiento semántico sobre los resultados obtenidos en recomendaciones previas, nuestro enfoque requiere identificar de forma única el perfil de cada usuario. Para ello, se definen las referencias *user-ID*. Los user-IDs permiten rastrear a los usuarios involucrados en cada proceso de recomendación, registrando además el rol que ha desempeñado en el mismo.

En definitiva, cuando un producto es recomendado a un cierto usuario activo, su perfil registra:

- El **item-ID** del producto en cuestión.
- Los índices **DOI** correspondientes a dicho producto y sus respectivos atributos —como valores nuevos, o actualizando los ya registrados en el perfil—.
- Los **user-IDs** de los vecinos que han participado en dicha recomendación.

3.3.6 Adaptación ante cambios en las preferencias de los usuarios

Tal como describimos en la Sección 2.2.4, 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, F) 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.

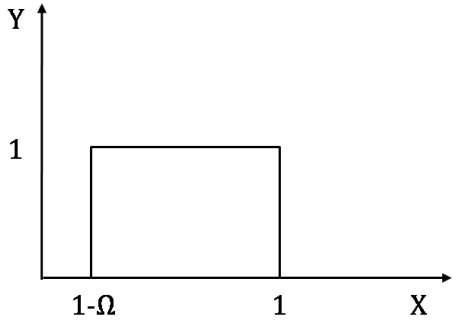
Función de olvido gradual basada en ventanas temporales

En concreto, nuestra metodología incorpora una función de olvido gradual basada en *ventanas temporales*. Es así que, durante el proceso de recomendación, el enfoque únicamente considera aquellas instancias cuyos índices de olvido F están dentro de una ventana con un ancho limitado Ω, representada en la Fig. 3.4.

El valor que toma la ventana (eje Y en la Fig. 3.4) determina las preferencias del perfil del usuario contempladas en el proceso de recomendación en un momento dado: se tienen en cuenta aquellas preferencias cuyo índice de olvido está dentro del espacio temporal considerado por la ventana (eje X en la Fig. 3.4) —i.e., desde las incorporadas en el perfil el momento actual (equivalente a índices de valor 1), hasta un cierto punto en el pasado (el de aquellas preferencias con índices de valor 1-Ω)—.

Tal como se describe en la Fig. 3.4, Ω es un parámetro ajustado dinámicamente en función de la precisión de las recomendaciones previas al usuario considerado, definida como la proporción de productos sugeridos por la herramienta y aceptados por éste.

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 F 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 recomendación.



$$\Omega = \text{Precisión} = \frac{\#\ \text{Productos aceptados}}{\#\ \text{Productos recomendados}}$$

Figura 3.4: Ventana temporal utilizada en la función de olvido gradual

Nótese que este mecanismo de redimensionamiento dinámico permite utilizar de forma eficiente el conocimiento aprendido sobre cada usuario. Es obvio que un índice de olvido poco significativo no tiene por qué identificar necesariamente preferencias no válidas para un usuario, sino únicamente aquellas almacenadas en su perfil hace tiempo. Por tanto, una ventana temporal de ancho fijo puede traducirse en pérdida de información útil, ya que todas aquellas instancias asociadas a índices reducidos son desechadas en el proceso de personalización. Gracias al proceso de ajuste dinámico utilizado, las instancias lejanas serán o no consideradas por el sistema en función de la precisión: si dichas preferencias son válidas para el usuario, la tasa de aciertos del sistema no se verá decrementada; en caso contrario, la ventana se estrechará y el recomendador trabajará únicamente sobre los intereses más recientes del usuario, adaptándose de forma flexible a su comportamiento y a sus intereses personales.

Actualización de los índices de olvido

De acuerdo con lo expuesto anteriormente, la idea es inicializar a 1 el índice de olvido de un producto nuevo que es incorporado al perfil del usuario, a la vez que se decrementan los índices asociados al resto de preferencias registradas previamente en el mismo.

Sin embargo, nuestro enfoque va un paso más allá y tiene en cuenta la heterogeneidad de productos existente en el dominio de aplicación considerado, actualizando de forma independiente las preferencias de productos de cada índole. Esta premisa es razonable, puesto que, por ejemplo, los cambios en los intereses del usuario respecto a música no tienen nada que ver con sus preferencias gastronómicas. En definitiva, la estrategia propuesta no decrementa los índices de todas las preferencias registradas en el perfil del usuario, sino que únicamente actualiza aquellas correspondientes a la categoría a la que pertenece el nuevo producto —i.e., la jerarquía de clases de la ontología

bajo la que está clasificado el mismo (ver Sección 3.2.1)—. Es decir, cuando el sistema incorpora un nuevo producto a un perfil, se inicializa a 1 el índice de olvido de dicho producto, así como el de sus características semánticas, y se decrementan los índices asociados al resto de preferencias del perfil pertenecientes a esa categoría[2].

En concreto, siendo i_j cada una de las instancias referenciadas desde el perfil de un usuario U pertenecientes a la categoría en cuestión, nuestro enfoque decrementa su índice de olvido en una proporción dependiente de la precisión del sistema, tal como muestra la Ec. 3.3.

$$F_U(i_j) = \begin{cases} F_U(i_j) - \frac{1}{1+\Omega} & \text{si } F_U(i_j) - \frac{1}{1+\Omega} \geq 0 \\ 0 & \text{en otro caso} \end{cases} \tag{3.3}$$

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.

3.4 Conclusiones

En este capítulo hemos presentado en detalle los dos elementos clave del marco de razonamiento semántico en el que se desarrollan las estrategias incluidas en nuestra metodología: la ontología del sistema, en el dominio del comercio electrónico, y la técnica de modelado de usuarios utilizada, basada en el concepto de *perfiles-ontología*.

Si bien nuestro enfoque no está vinculado a una ontología específica, se adopta en él una conceptualización del dominio de comercio electrónico concreta. Dicha base de conocimiento recoge las singularidades propias de dominios de aplicación caracterizados por la heterogeneidad de sus productos, como el que nos ocupa.

Gracias al conocimiento formalizado en la ontología del sistema, la técnica de modelado de usuarios descrita aporta descripciones semánticas muy detalladas sobre sus preferencias, que serán aprovechadas, de forma efectiva, en las estrategias presentadas en los siguientes capítulos.

[2]Conviene aclarar que si una característica semántica ya estaba definida en el perfil y, a la vez, también está asociada al producto añadido al mismo, su índice de olvido no se decrementa, sino que se inicializa a 1.

Capítulo 4

Estrategia para combatir el problema de falsos vecindarios

Este capítulo describe una estrategia de identificación de usuarios con preferencias similares, para ser empleada en sistemas recomendadores colaborativos en cuyos dominios de aplicación existe una gran diversidad de productos. En tales dominios, los perfiles de los usuarios contienen preferencias sobre productos de diferente naturaleza, lo que puede llevar a las técnicas de formación de vecindarios convencionales a realizar una selección contraproducente de usuarios afines. La metodología que proponemos explota técnicas de razonamiento semántico para tratar de aliviar el problema de falsos vecindarios, a fin de mejorar el proceso de selección de usuarios con intereses similares y, en consecuencia, el resultado de las recomendaciones.

4.1 Introducción

Tal como describimos en la Sección 2.4, en dominios de aplicación donde los productos objetivo son de muy distinta naturaleza (como es el caso del comercio electrónico, en el que el producto a recomendar puede ser desde un libro a un producto alimenticio), las estrategias colaborativas tradicionales pueden llevar a la formación de *falsos vecindarios* [83, 84]. En concreto, cuando miden la similitud entre usuarios, dichas estrategias tienen en cuenta los índices de interés asignados a *todos* los productos incluidos en sus perfiles. Esto es, se forma un solo vecindario para el usuario activo independientemente del producto objetivo que se esté considerando en ese momento.

El problema surge cuando los usuarios tienen preferencias similares en gran cantidad de productos (por lo cual son considerados vecinos) pero difieren en los del tipo del producto objetivo. Dicho de otro modo, la gran diversidad en la naturaleza de los productos hace que esas técnicas sean propensas a una selección contraproducente de vecinos, dado que dos usuarios pueden ser afines en un tipo de productos pero no en otros. Por ejemplo, dos usuarios que tienen los mismos gustos en libros, coches, equipamiento deportivo y música, pero no comparten gustos gastronómicos, serán falsos vecinos siempre que el producto objetivo sea un producto alimenticio.

El considerar todos los productos incluidos en los perfiles de los usuarios al cuan-

tificar la similitud existente entre el usuario activo y los candidatos a vecinos tiene dos consecuencias negativas en la selección del vecindario:

- Por un lado, pueden formar parte del vecindario usuarios cuyas preferencias difieren de las del usuario activo respecto al producto objetivo: los denominados falsos vecinos. Estos son aquellos usuarios que tienen gustos similares a los del usuario activo en gran cantidad de productos, pero no en los del tipo del producto considerado en la recomendación actual.
- Por otro lado, dado que el tamaño del vecindario es limitado —generalmente, a los n candidatos a vecinos más afines (ver Sección 2.3.3)—, se desecharán candidatos que, pese a no coincidir en la mayoría de productos, tienen intereses comunes en relación al producto objetivo. En otras palabras, pueden no ser incluidos en el vecindario usuarios con un gran potencial para obtener buenos resultados en el proceso de recomendación en curso.

Es por todo esto que el fundamento de la estrategia propuesta en este capítulo radica en seleccionar el vecindario *dinámicamente*, considerando para el usuario activo el vecindario más apropiado en cada caso —en función del tipo de producto objetivo contemplado en la recomendación—.

Aunque este problema ya ha sido identificado en la literatura por varios autores (ver Sección 2.4), la naturaleza sintáctica de los enfoques revisados limita en gran medida la calidad de las soluciones propuestas. En efecto, al prescindir de la semántica de los productos en el proceso, únicamente son capaces de detectar similitudes entre usuarios que contienen en sus perfiles preferencias identificadas *exactamente* con los mismos términos. Por ejemplo, no detectarán una preferencia común entre dos usuarios a los que les gusta el café si en sus perfiles tienen registradas valoraciones referentes a distintas marcas de café. Precisamente, el tratar de seleccionar diferentes vecindarios para cada usuario activo empleando aproximaciones sintácticas, agrava el *sparsity problem* asociado a los sistemas colaborativos, tal como veremos en la Sección 4.3.1.

Nuestro enfoque no sólo tiene en cuenta la semántica asociada a los productos contenidos en los perfiles (formalizada en la ontología de productos comerciales) a la hora de estimar la similitud entre dos usuarios dados, sino que además explota técnicas de razonamiento sobre dicha ontología a fin de incorporar nuevo conocimiento en el proceso. En particular, adaptaremos la métrica de *similitud semántica* definida por Blanco en [16] (introducida en la Sección 2.5.1) para descubrir relaciones semánticas significativas, ocultas en la base de conocimiento, existentes entre los productos almacenados en el perfil del usuario activo y los registrados en los perfiles de los candidatos a vecinos. Tal como describiremos en la Sección 4.2.2, el partir de una métrica empleada en la fase basada en contenido de un enfoque híbrido, con el objetivo de aprovechar sus beneficios en un enfoque puramente colaborativo, requiere redefinir parte de la misma.

Con la métrica resultante, enriqueceremos la propuesta de Ganesan *et al.* [43] de redefinir el producto escalar en el cálculo de la correlación entre las preferencias de dos usuarios (descrita en la Sección 2.5.1) y la adaptaremos para satisfacer las necesidades de nuestra estrategia —esto es, tratar de aliviar el problema de falsos vecindarios—. Además, aprovechando las posibilidades que ofrece la técnica de modelado de usuarios utilizada en nuestro enfoque (ver Sección 3.3.4), la estrategia propuesta incorpora en el proceso una nueva métrica de similitud entre usuarios que atiende a los niveles de interés en las características semánticas registradas en sus perfiles, explotando la información adicional que éstos proporcionan acerca de las preferencias de los usuarios.

4.2 Similitud semántica entre productos

La medida de similitud semántica definida en AVATAR (introducida en la Sección 2.5.1) es utilizada por dicho sistema en la fase basada en contenido de su estrategia de personalización híbrida, esto es, para identificar productos (en su caso, programas de televisión) parecidos a aquellos por los que el usuario activo haya mostrado interés en el pasado. En otras palabras, su métrica es empleada para estimar la similitud entre el producto objetivo de la recomendación y las preferencias contenidas en el perfil del usuario activo en cuestión.

Tal como hemos adelantado en la sección previa, en nuestra estrategia adaptaremos esta métrica para inferir la similitud entre cada uno de los productos almacenados en el perfil del usuario activo y cada uno de los registrados en el perfil de un candidato a vecino dado. El objetivo último perseguido, que será abordado en la Sección 4.3, es incorporar la información resultante a la medida de correlación entre las preferencias de estos usuarios, combatiendo el problema de falsos vecinos.

La similitud semántica entre dos productos específicos depende de las relaciones semánticas inferidas entre ambos en la ontología de productos comerciales. Dado que nuestra metodología se apoya en 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 descritas en la Sección 2.5.1). Esto último es especialmente importante en nuestro enfoque, puesto 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 fiables.

La métrica propuesta por Blanco 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. Para este propósito, se definen en ella dos componentes: la *similitud semántica jerárquica* ($SemSim_{Hie}$) y la *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. 4.1, en la que se cuantifica la similitud semántica entre los productos i_i e i_j.

$$SemSim(i_i, i_j) = \alpha \cdot SemSim_{Inf}(i_i, i_j) + (1-\alpha) \cdot SemSim_{Hie}(i_i, i_j) \tag{4.1}$$

Las expresiones matemáticas de ambas componentes son presentadas a continuación e ilustradas mediante diferentes ejemplos.

4.2.1 Similitud semántica jerárquica

El valor de la similitud jerárquica entre i_i e i_j depende de la posición que ocupan en la ontología las clases a las que pertenecen dichas instancias, C_{i_i} y C_{i_j} respectivamente. Su definición se basa en los conceptos de *profundidad* y *antecesor común más cercano (LCA)* presentados en la Sección 2.5.1. La expresión resultante se muestra en la Ec. 4.2, de la que es posible extraer algunas conclusiones interesantes acerca de los valores cuantificados por esta métrica.

$$SemSim_{Hie}(i_i, i_j) = \frac{depth(LCA_{i_i,i_j})}{\max(depth(i_i), depth(i_j))} \tag{4.2}$$

Para esclarecer dichas conclusiones, detalladas a continuación, consideraremos a modo de ilustración el escenario descrito en la Fig. 4.1, donde se muestra un pequeño extracto de la ontología de productos comerciales del sistema (en concreto, parte de la jerarquía de la categoría *Productos Alimenticios* y parte de la jerarquía de la categoría *Libros*) y, rodeados en ella, los *LCA*s correspondientes a ciertos pares de productos (instancias de la ontología).

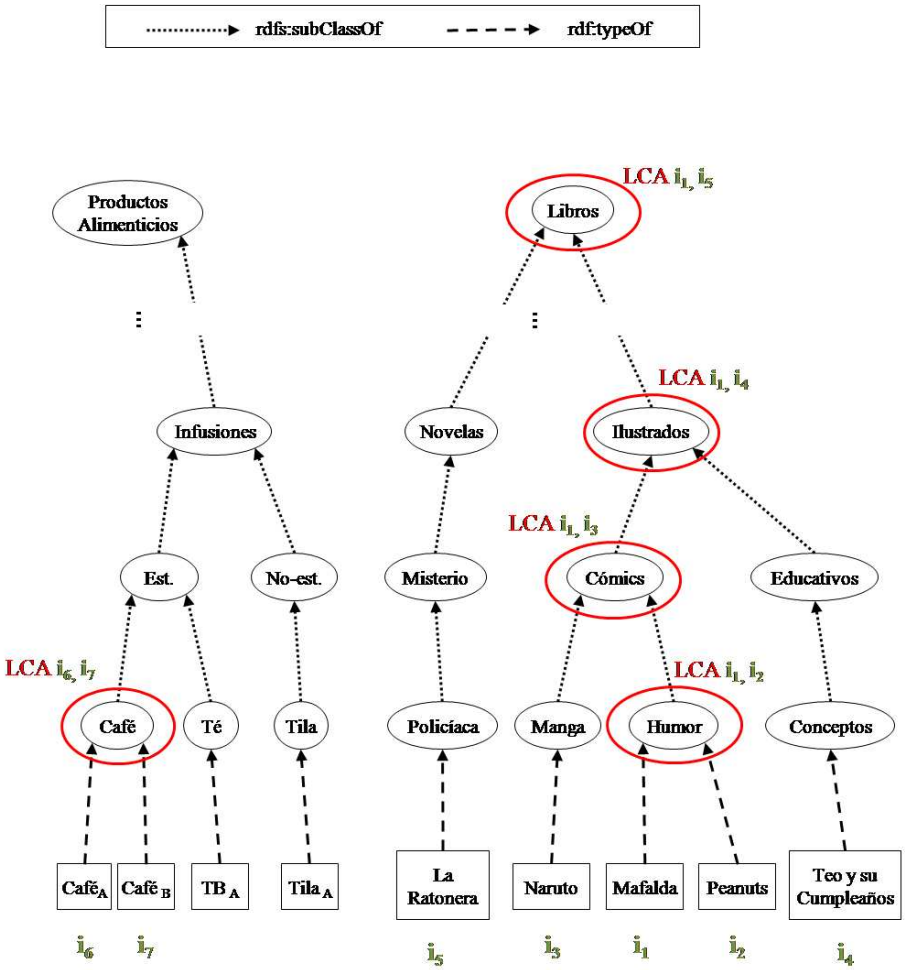


Figura 4.1: Similitud semántica jerárquica entre productos

- La similitud jerárquica $SemSim_{Hie}(i_i, i_j)$ se anula cuando el único antecesor común que comparten las clases C_{i_i} y C_{i_j} es el nodo raíz de la ontología (*owl:Thing*), 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.

  Así, por ejemplo, en el escenario mostrado en la Fig. 4.1, la similitud jerárquica entre la instancia $Café_A$ y la instancia *Mafalda* será nula, ya que dichos productos pertenecen a distintas jerarquías de productos —$Café_A$ a la jerarquía de *Productos Alimenticios*, y *Mafalda* a la jerarquía de *Libros*—. Precisamente, el *LCA* de las clases *Café* y *Humor* es el nodo raíz de la ontología, *owl:Thing* —no representado en la figura—.

- En caso de que exista un *LCA* diferente de dicha clase raíz, la similitud entre i_i e i_j 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, delimita en mayor medida la relación existente entre las dos instancias comparadas. En efecto, cuanto más profundo es el *LCA*, más concreto es el tipo de productos acotado por la subcategoría bajo la que están clasificadas ambas instancias y, por tanto, mayor será su relación de similitud.

  Para ilustrar esta idea, consideremos la instancia *Mafalda* en la Fig. 4.1. La similitud semántica jerárquica entre dicha instancia y *La Ratonera* será menos significativa que la existente entre *Mafalda* y la instancia *Teo y su Cumpleaños*, puesto que la profundidad del *LCA* de éstas (la clase *Ilustrados*) es mayor que la del *LCA* correspondiente a *Mafalda* y *La Ratonera* (*Libros*). Nótese que es coherente que *Mafalda* y *Teo y su Cumpleaños* obtengan mayor valor de similitud, dado que ambos son libros ilustrados, mientras que lo único que tienen en común *Mafalda* y *La Ratonera* (en cuanto a sus descripciones jerárquicas se refiere) es que ambos son libros. Igualmente, es posible observar en la figura cómo la similitud jerárquica entre *Mafalda* y la instancia *Naruto* será aún mayor, ya que el *LCA* entre ambos (*Cómics*) tiene una profundidad superior a los anteriores —i.e., *Mafalda* y *Naruto* no sólo son libros ilustrados, sino que además son ambos cómics—.

- Por último, la Ec. 4.2 conduce a valores de similitud más elevados cuanto más próximo esté el antecesor común LCA_{i_i,i_j} de las clases a las que pertenecen dichas instancias, debido a que, en este caso, la relación entre i_i e i_j es más significativa —i.e., la distancia que separa a una instancia de la otra en la jerarquía es menor—. Precisamente el valor máximo de similitud jerárquica se alcanza cuando las dos instancias comparadas pertenecen a la misma clase (i.e., $C_{i_i} = C_{i_j}$). En ese caso, el antecesor común de ambas es la propia clase y, en consecuencia, $SemSim_{Hie}(i_i, i_j)$ toma el valor 1.

  Por ejemplo, en la Fig. 4.1, *Mafalda* y la instancia *Peanuts* pertenecen a la misma clase, *Humor* (que es a su vez el *LCA*) y, por tanto, la similitud semántica jerárquica obtenida ente ellas es máxima. Asimismo, la similitud semántica jerárquica entre las instancias $Café_A$ y $Café_B$ también será 1 —en efecto, hablamos de dos marcas de café, instancias de la clase *Café*—.

4.2.2 Similitud semántica inferencial

La similitud semántica inferencial está basada en descubrir relaciones implícitas entre i_i e i_j, utilizando para ello las propiedades definidas en la ontología. Para su definición, Blanco adapta dos de las asociaciones semánticas identificadas por Anyanwu y Sheth en el proyecto SemDis [5], tal como adelantamos en la Sección 2.5.1. En concreto, conforme a su propuesta, la similitud inferencial aplicada a nuestra estrategia descubre asociaciones ρ-*path* y ρ-*join* de longitud mínima entre los dos productos comparados. Ambos tipos de asociaciones comparten una misma intuición, que establece que dos productos están relacionados cuando comparten características semánticas (e.g., atributos específicos en las jerarquías *Lugar de Origen*, *Rango de Precio*, *Autor*, *Variedad de Uva*, *Tipo de Leche*, *Sabor*, etc.), tal como se aprecia en las definiciones presentadas a continuación:

- **Asociación ρ-*path***(i_i,i_j): Este tipo de asociaciones se infieren entre dos productos que están conectados a través de un *atributo común a* (i.e., *a* es una característica semántica presente tanto en i_i como en i_j), tal como se muestra en la Fig. 4.2.

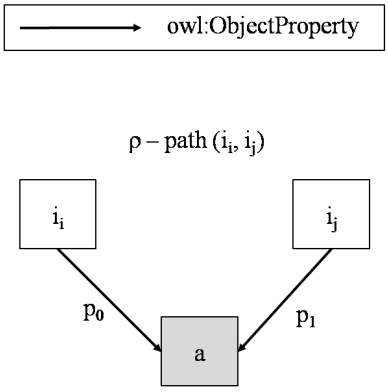


Figura 4.2: Asociación ρ-*path* entre los productos i_i e i_j

Ejemplo: Consideremos el extracto de la ontología de productos representado en la Fig. 4.3. En este escenario, es posible identificar una relación implícita ([A] en la figura) entre el libro *Kim* y el titulado *Siddhartha Gautama, el Buda*. Tal como muestra la Fig. 4.3, la inferencia de esta asociación está basada en que ambos libros tienen un tema en común: *India*[1].
Por el contrario, un enfoque puramente jerárquico no descubriría relación alguna entre los dos productos mencionados, debido a que no existe ninguna clase (diferente del nodo raíz de esa categoría, *Libros*) que sea antecesor común de sus respectivas clases. Como consecuencia de ello, este tipo de propuestas cuantificarían una similitud mínima entre los dos libros considerados, a pesar de estar implícitamente relacionados a través de un atributo común, tal como infiere la métrica considerada en nuestra estrategia.

[1]Conviene destacar que la similitud inferencial considera un gran número de características semánticas en su métrica, de forma que dos productos serán considerados similares únicamente cuando compartan un número significativo de atributos. En este ejemplo, proponemos un escenario muy simple para ilustrar el tipo de asociaciones inferidas en dicho enfoque.

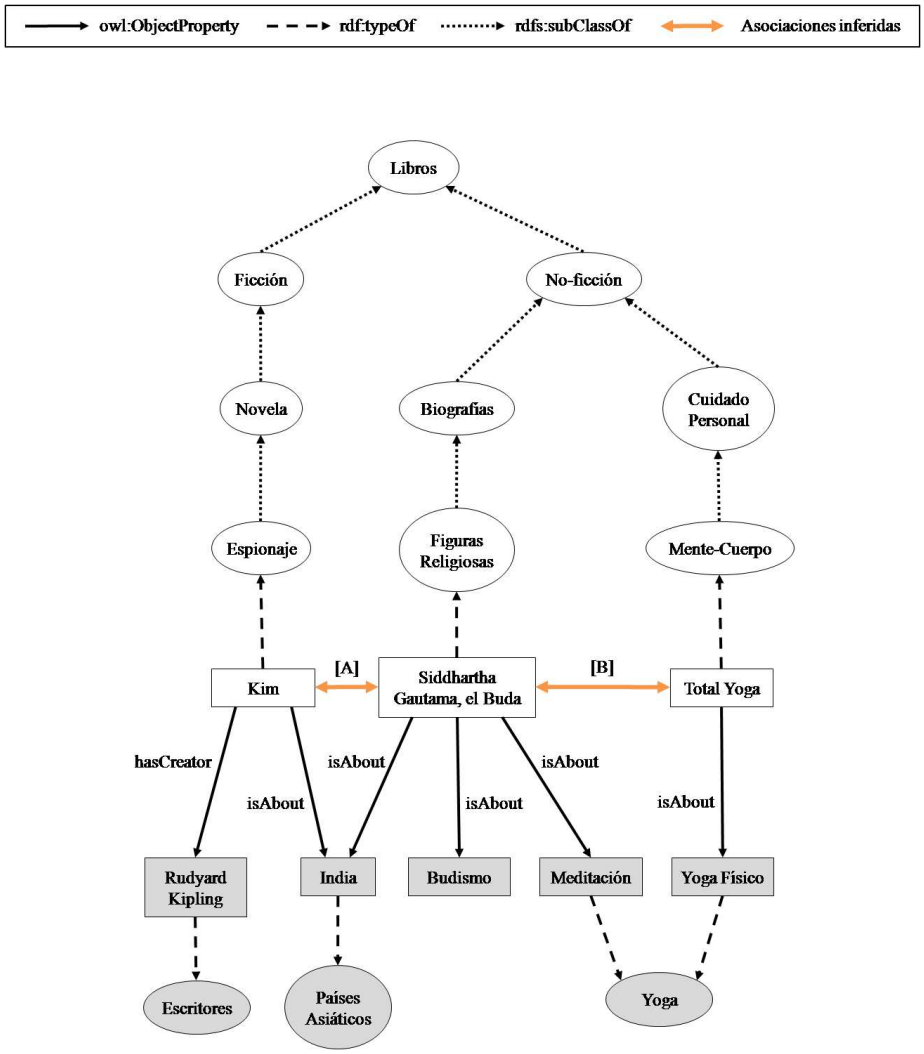


Figura 4.3: Algunas asociaciones semánticas utilizadas en el cálculo de la similitud semántica inferencial

- **Asociación** ρ-***join***(i_i,i_j): Esta asociación, representada en la Fig. 4.4, es inferida entre dos productos que están conectados a través de dos *atributos hermanos* a_1 y a_2 (i.e., atributos que pertenecen a una misma clase A, la denominada *clase de unión* —de ahí la asociación ρ-*join*—).

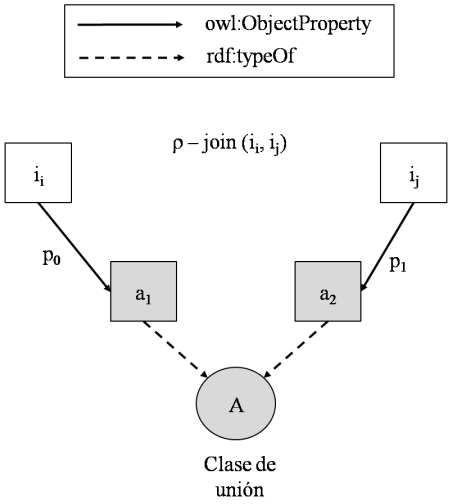


Figura 4.4: Asociación ρ-*join* entre los productos i_i e i_j

Ejemplo: En la Fig. 4.3 es posible identificar una asociación implícita entre el libro *Siddhartha Gautama, el Buda* y el titulado *Total Yoga* (representada como [B]). Ésta es debida a que ambos productos tienen contenidos relacionados con la práctica del yoga. A pesar de que ambos productos tienen definidos diferentes atributos (*Meditación* y *Yoga Físico*, respectivamente), nuestro enfoque es capaz de relacionarlos semánticamente mediante la clase de unión *Yoga*.

Tomando como base estas asociaciones, redefinimos la expresión matemática de la similitud inferencial presentada en AVATAR [16] para adaptarla a las necesidades de nuestra estrategia. La redefinición de la expresión propuesta por Blanco responde a la diferente aplicación de la misma en cada enfoque. Blanco hace uso de la similitud inferencial en la fase basada en contenido de su estrategia de recomendación híbrida, para determinar la intensidad de la relación semántica entre el producto objetivo y las preferencias del usuario activo. Por tanto, es razonable que en aquella expresión se consideren los índices de interés del usuario activo respecto a los atributos involucrados en las asociaciones semánticas descubiertas, puesto que su propósito es inferir la idoneidad de recomendar un producto para ese usuario en concreto. Sin embargo, el objetivo buscado en nuestro enfoque es determinar la similitud existente entre dos productos de forma inherente, para incluir dichas medidas absolutas en el cálculo de la similitud entre las preferencias de dos usuarios (donde se considerarán los niveles de interés de los usuarios en los respectivos productos), tal como describiremos en la Sección 4.3.2.

En consecuencia, son dos las premisas que hemos considerado para dar forma a dicha expresión en nuestro enfoque:

1. En primer lugar, el valor de similitud resultante debe ser más elevado cuanto mayor sea el número de asociaciones semánticas identificadas entre el producto i_i y el producto i_j.

2. Por otra parte, las asociaciones ρ-*path* son más significativas que las asociaciones ρ-*join*. En el primer caso, los productos comparados comparten exactamente los mismos atributos, mientras que en el segundo escenario estas características semánticas, si bien pertenecen a una misma clase, son diferentes. Consecuentemente, las asociaciones ρ-*path* han de tener un mayor peso en el cálculo de la similitud inferencial.

Para poder incluir estas premisas en el cálculo, utilizamos los siguientes factores:

- *Número máximo de asociaciones* ($\#\rho_{MAX}$). De acuerdo a la primera premisa, se tendrá en cuenta el número de asociaciones semánticas identificadas entre dos productos, respecto al máximo alcanzable. Como su nombre indica, este factor refleja el número máximo de posibles asociaciones semánticas (tanto ρ-*path* como ρ-*join*) entre dos instancias.

- *Factor de penalización* (f_p). Con arreglo a la segunda premisa, el objetivo de este factor es penalizar los valores de similitud inferencial obtenidos a partir de asociaciones ρ-*join*.

Como resultado, obtenemos la expresión mostrada en la Ec. 4.3, donde $\#\rho(i_i, i_j)$ es el número de asociaciones semánticas inferidas entre i_i e i_j, ρ_k es la k-ésima de esas asociaciones, $\#\rho_{MAX}(i_i, i_j)$ es el número máximo de posibles asociaciones entre ambos productos —determinado por el mínimo entre el número de atributos vinculados a i_i, $\#A(i_i)$, y el número de atributos vinculados a i_j, $\#A(i_j)$— y $f_p(\rho_k)$ es el factor de penalización asociado a la asociación ρ_k.

$$SemSim_{Inf}(i_i, i_j) = \frac{1}{\#\rho_{MAX}(i_i, i_j)} \sum_{k=1}^{\#\rho(i_i, i_j)} f_p(\rho_k) \tag{4.3}$$

donde:

$$\#\rho_{MAX}(i_i, i_j) = min(\#A(i_i), \#A(i_j))$$

$$f_p(\rho_k) = \begin{cases} 1 & \text{si } \rho_k \text{ es asociación } \rho\text{-}path \\ 0{,}85 & \text{si } \rho_k \text{ es asociación } \rho\text{-}join \end{cases}$$

4.3 Similitud entre usuarios

Se detalla en esta sección cómo nuestra estrategia modifica el proceso de identificación de usuarios afines para tratar de aliviar el problema de falsos vecindarios, especificando una nueva métrica de similitud entre usuarios —en la que se incorpora la métrica de similitud semántica entre productos presentada en la sección previa—. En primer lugar, se recuerdan las expresiones tradicionales de similitud entre usuarios, ya que éstas serán tomadas como punto de partida en nuestra métrica. A continuación, se exponen los beneficios que aporta el emplear un enfoque semántico en el proceso y se define la expresión de la similitud semántica entre usuarios propuesta. Finalmente, se describe el procedimiento seguido a fin de mitigar el problema de falsos vecindarios en base a dicha métrica.

4.3.1 Medidas convencionales

En la Sección 2.3.3 describimos las dos métricas utilizadas de forma generalizada en los sistemas recomendadores colaborativos para calcular la similitud entre las preferencias de dos usuarios. En ambos casos, las preferencias de los usuarios se representan mediante sendos *vectores de interés*, cuyas componentes son las valoraciones dadas por cada uno de ellos a los productos incluidos en sus correspondientes perfiles. En concreto, se obtiene la similitud entre un cierto usuario activo U y cada candidato a vecino CN_i, cuyos vectores de interés son v_U^p y $v_{CN_i}^p$, respectivamente, mediante una de las siguientes expresiones:

- **Similitud del coseno**: Esta medida define la similitud entre los dos vectores como el coseno del ángulo que forman. Su expresión, que se deriva de la fórmula del *producto escalar* de dichos vectores, está formalizada en la Ec. 4.4, donde $\left|\overrightarrow{v_U^p}\right|$ y $\left|\overrightarrow{v_{CN_i}^p}\right|$ son los módulos de los vectores v_U^p y $v_{CN_i}^p$, respectivamente.

$$Sim_{Cos}(v_U^p, v_{CN_i}^p) = \frac{\overrightarrow{v_U^p} \cdot \overrightarrow{v_{CN_i}^p}}{\left|\overrightarrow{v_U^p}\right| \cdot \left|\overrightarrow{v_{CN_i}^p}\right|} \tag{4.4}$$

- **Correlación Pearson-r**: Formalizada en la Ec. 4.5, esta expresión es equivalente a restar la media de los elementos de cada vector ($\overline{v_U^p}$ y $\overline{v_{CN_i}^p}$) antes de calcular su similitud del coseno. El motivo de esta modificación es compensar el efecto de *datos no centrados* en los vectores, esto es, hacer que la media aritmética de sus componentes sea cero. De esta forma, la medida de similitud obtenida no se ve distorsionada cuando la media de las componentes incluidas en los vectores (i.e., los índices de interés registrados en los perfiles de los usuarios) no se anula.

$$corr(v_U^p, v_{CN_i}^p) = \frac{\sum_k (v_U^p[k] - \overline{v_U^p})(v_{CN_i}^p[k] - \overline{v_{CN_i}^p})}{\sqrt{\sum_k (v_U^p[k] - \overline{v_U^p})^2 \cdot \sum_k (v_{CN_i}^p[k] - \overline{v_{CN_i}^p})^2}} \tag{4.5}$$

Ambas expresiones permiten cuantificar tanto valores positivos como negativos de similitud entre las preferencias de los usuarios U y CN_i. De hecho, las medidas oscilan

entre -1 y 1, obteniendo valores positivos si U ha clasificado con niveles de interés similares a los de CN_i los mismos productos (i.e., los vectores v_U^p y $v_{CN_i}^p$ contienen valores parecidos en las mismas componentes), y negativos cuando las valoraciones de ambos usuarios son muy diferentes.

El mayor problema asociado a la utilización de estas medidas de similitud en los sistemas colaborativos tradicionales es el ya mencionado *sparsity problem*. En efecto, para que el valor resultante nos ofrezca información acerca de la similitud entre dos usuarios (i.e., no sea cero), se requiere que exista solapamiento entre los productos clasificados por los usuarios comparados. Dada la naturaleza sintáctica de los enfoques tradicionales, esta condición hace que sea necesario que los dos perfiles considerados almacenen productos identificados exactamente con los mismos términos. Es obvio que cuanto mayor sea el número de productos disponibles en el sistema recomendador, será menos probable que dos usuarios clasifiquen los mismos productos, dificultando la identificación de usuarios similares a un usuario activo dado. De ahí que este problema sea especialmente acusado en dominios de aplicación como el comercio electrónico, debido a la miríada de productos disponibles en el mercado.

4.3.2 Un enfoque semántico

La incorporación de la semántica en el proceso, formalizada en una ontología de dominio, facilita la comparación de las preferencias de los usuarios atendiendo no sólo a los términos que identifican a los productos registrados en sus perfiles, sino también al significado de dichos términos. En otras palabras, es posible incluir gran cantidad de información útil acerca de las preferencias de los usuarios, no disponible en los enfoques sintácticos tradicionales.

Nuestra propuesta explota la semántica definida en la ontología de productos, en el proceso de selección de vecindarios, desde dos perspectivas:

- Por un lado, las descripciones semánticas de los productos posibilitan una comparación flexible y significativa de los niveles de interés asociados a dichos productos en los perfiles de los usuarios. En efecto, tal como hemos comentado anteriormente, explotando la semántica formalizada en una ontología de dominio es posible inferir relaciones de similitud entre productos que, si bien no están identificados con los mismos términos, son semejantes. Como resultado, al poder construirse vecindarios sobre productos similares —pero no necesariamente los mismos—, se logra mitigar de forma eficiente los efectos adversos del *sparsity problem*.

- Por otra parte, nuestra estrategia va un paso más allá en el cálculo de similitud entre las preferencias de dos usuarios, aprovechando la información que ofrecen acerca de los intereses de dichos usuarios las características semánticas de los productos en sí mismas.

Reuniendo estas ideas, se propone una nueva métrica de similitud entre usuarios, que incluye dos componentes: una medida de similitud en base a productos, y una medida de similitud en base a atributos. Describimos a continuación cada una de ellas, así como la expresión resultante de la combinación de ambas.

Similitud semántica entre usuarios en base a productos

Tal como introdujimos brevemente en la Sección 2.5.1, con el objetivo de aliviar los efectos adversos del *sparsity ploblem*, Ganesan *et al.* proponen modificar la forma de calcular el producto escalar entre los vectores de interés de los usuarios comparados, incluyendo el conocimiento semántico contenido en la jerarquía de clasificación de los productos en cuestión.

La idea es comparar el nivel de interés asociado a cada producto almacenado en el perfil de U con cada uno de los del perfil de CN_i (i.e., cada componente del vector v_U^p con cada una del vector $v_{CN_i}^p$), considerando la similitud semántica existente entre dichos productos. De esta manera, por ejemplo, es posible detectar cierto nivel de similitud entre dos usuarios que tienen gran afición por las novelas de misterio aunque los títulos de novelas de misterio contenidas en sus respectivos perfiles sean diferentes.

En vista de que la propuesta de Ganesan *et al.* se limita a incorporar una medida de similitud jerárquica en el proceso, nuestro enfoque enriquece su metodología haciendo uso de la medida de similitud semántica entre productos descrita en la sección previa. Por consiguiente, el producto escalar de los vectores de interés v_U^p y $v_{CN_i}^p$ será obtenido mediante la Ec. 4.6, donde $v_U^p[j]$ es la componente j-ésima del vector v_U^p —cuyo valor es, por tanto, el índice DOI registrado en el perfil de U asociado al producto que representa—, $v_{CN_i}^p[k]$ la componente k-ésima del vector $v_{CN_i}^p$ y $SemSim(i_j, i_k)$ la similitud semántica entre los productos correspondientes a dichas componentes (ver Ec. 4.1).

$$\overrightarrow{v_U^p} \cdot \overrightarrow{v_{CN_i}^p} = \sum_j \sum_k v_U^p[j] \cdot v_{CN_i}^p[k] \cdot SemSim(i_j, i_k) \tag{4.6}$$

Nótese que, a diferencia del producto escalar tradicional, esta aproximación no requiere que los vectores tengan el mismo tamaño ni exista solapamiento entre sus componentes, ya que se comparan todas las componentes de uno con todas las del otro.

Además, al emplear la métrica de similitud semántica entre productos formalizada en la Ec. 4.1, se logra optimizar aún más la detección de similitudes entre usuarios en presencia de *sparsity*, identificando relaciones que pasarían desapercibidas si únicamente consideramos medidas de similitud semántica jerárquica entre los productos clasificados en sus perfiles. Recordemos, por ejemplo, las asociaciones semánticas inferidas entre el libro *Kim*, el titulado *Siddhartha Gautama, el Buda* y el titulado *Total Yoga*, en el escenario mostrado en la Fig. 4.3.

Reuniendo las ideas presentadas hasta este punto, nuestra estrategia obtiene la denominada *similitud semántica en base a productos* entre dos usuarios como la correlación Pearson-r entre sus respectivos vectores de interés, sobre la que se aplica la modificación del producto escalar formalizada en Ec. 4.6. Recordemos que la expresión de la correlación Pearson-r es equivalente a restar la media de los elementos de cada vector y, a continuación, calcular su similitud del coseno, mediante la Ec. 4.4.

En definitiva, definimos la similitud semántica en base a productos entre el usuario activo U y el candidato a vecino CN_i mediante la Ec. 4.7.

$$sim_{U,CN_i}^p = \frac{\sum_j \sum_k v_{\phi_U}^p[j] \cdot v_{\phi_{CN_i}}^p[k] \cdot SemSim(i_j, i_k)}{\left|\overrightarrow{v_{\phi_U}^p}\right| \cdot \left|\overrightarrow{v_{\phi_{CN_i}}^p}\right|} \tag{4.7}$$

donde:

$$v_{\phi U}^{p} = (v_U^p[1] - \overline{v_U^p}, v_U^p[2] - \overline{v_U^p}, \ldots, v_U^p[j] - \overline{v_U^p})$$

$$v_{\phi CN_i}^{p} = (v_{CN_i}^p[1] - \overline{v_{CN_i}^p}, v_{CN_i}^p[2] - \overline{v_{CN_i}^p}, \ldots, v_{CN_i}^p[k] - \overline{v_{CN_i}^p})$$

En esta expresión, $v_{\phi U}^{p}$ y $v_{\phi CN_i}^{p}$ son los *vectores centrados* de los vectores de interés de los usuarios —v_U^p y $v_{CN_i}^p$, respectivamente—. Esto es, dichos vectores son el resultado de restar la media de los elementos de cada vector a cada una de sus componentes ($\overline{v_U^p}$ a cada componente de v_U^p y $\overline{v_{CN_i}^p}$ a cada componente de $v_{CN_i}^p$).

Similitud semántica entre usuarios en base a atributos

La estrategia de modelado utilizada en nuestro enfoque asigna niveles de interés (índices DOI) individuales a las características semánticas de los productos almacenados en el perfil del usuario, tal como describimos en la Sección 3.3.4. Dichos índices DOI proporcionan información adicional acerca de las preferencias del usuario, ya que reflejan su interés en según qué atributos de los productos.

Más concretamente, en función del tipo de características semánticas consideradas, generales o propias de una determinada jerarquía de productos (ver Sección 3.2), los correspondientes niveles de interés representan:

- **Atributos generales**: particularidades en los intereses generales y hábitos de consumo del usuario en cuestión como, por ejemplo, su predilección por artículos de lujo o un especial apego a productos exóticos.

- **Atributos propios**: información adicional sobre las preferencias del usuario respecto a productos de una naturaleza concreta como, por ejemplo, un particular gusto por la comida picante o su interés por un cierto escritor o grupo de música.

A fin de aprovechar esta importante información en el proceso de selección de vecindarios, nuestra metodología construye *vectores de interés de atributos* para el usuario activo y cada uno de los candidatos a vecinos (v_U^a y $v_{CN_i}^a$, respectivamente), de manera semejante a la formación de los *vectores de interés de productos* utilizados hasta el momento (v_U^p y $v_{CN_i}^p$). Es decir, teniendo en cuenta que los vectores de interés de productos contienen los índices DOI correspondientes a los productos registrados en los perfiles de los usuarios, los vectores de interés de atributos tendrán como componentes los índices DOI asociados a las características semánticas incluidas en dichos perfiles.

Luego, la *similitud semántica en base a atributos* entre un usuario activo U y un candidato a vecino CN_i, se obtiene mediante el cálculo de la correlación Pearson-r entre sus vectores de interés de atributos, v_U^a y $v_{CN_i}^a$, tal como muestra la Ec. 4.8.

$$sim_{U,CN_i}^{a} = corr(v_U^a, v_{CN_i}^a) = \frac{\sum_k (v_U^a[k] - \overline{v_U^a})(v_{CN_i}^a[k] - \overline{v_{CN_i}^a})}{\sqrt{\sum_k (v_U^a[k] - \overline{v_U^a})^2 \cdot \sum_k (v_{CN_i}^a[k] - \overline{v_{CN_i}^a})^2}} \tag{4.8}$$

Nótese que, a diferencia de la métrica de similitud semántica entre productos, en este caso se emplea la expresión tradicional, en la que las componentes k-ésimas de dichos vectores representan lo mismo —i.e., los índices DOI asociados al atributo correspondiente, registrados en los perfiles de U y CN_i—.

Similitud semántica entre usuarios

Reuniendo las ideas y expresiones presentadas, nuestra estrategia define la similitud semántica entre un usuario activo U y un candidato a vecino CN_i a partir de su similitud semántica en base a productos, sim^p_{U,CN_i}, y su similitud semántica en base a atributos, sim^a_{U,CN_i}, tal como describe la Ec. 4.9. En esta expresión, θ es un parámetro de combinación que modula el peso de cada componente[1].

$$sim_{U,CN_i} = \theta \cdot sim^p_{U,CN_i} + (1-\theta) \cdot sim^a_{U,CN_i} \tag{4.9}$$

En suma, a la hora de identificar usuarios afines, el enfoque propuesto considera los niveles de interés asignados a productos *similares* (semánticamente) en sus respectivos perfiles y los niveles de interés de dichos usuarios en relación a las características semánticas registradas en ellos. De este modo, es capaz de detectar similitudes significativas entre las preferencias de los usuarios que pasarían desapercibidas en los enfoques adoptados hasta la fecha, combatiendo eficazmente el *sparsity problem* típico de los sistemas recomendadores colaborativos.

- Por un lado, nuestra métrica es capaz de detectar similitudes entre las preferencias de los usuarios más allá de las relaciones jerárquicas existentes entre las correspondientes instancias de productos en la ontología, descubriendo relaciones ocultas en la base de conocimiento —mediante la métrica de similitud semántica entre productos presentada en la Sección 4.2—. De este modo, se requiere un menor solapamiento entre los productos clasificados en los perfiles para identificar afinidades relevantes entre los intereses de los usuarios.
- Por otra parte, al aprovechar la información adicional reflejada en los niveles de interés de los atributos registrados en los perfiles de los usuarios, es posible detectar afinidades entre dichos usuarios respecto a características semánticas concretas (e.g., una común predilección por artículos de lujo).

4.3.3 Mitigando el problema de falsos vecindarios

Para combatir el problema de falsos vecindarios, presente en sistemas recomendadores cuyo dominio de aplicación cuenta con productos de muy distinta naturaleza, no basta con la modificación de la métrica de similitud entre usuarios que acabamos de describir. Tal como adelantamos en la Sección 4.1, con ese fin nuestro enfoque realiza una selección del vecindario de forma dinámica, identificando para el usuario activo un grupo u otro de vecinos en función del tipo de producto objetivo considerado.

Con esta idea en mente, la estrategia propuesta sigue los siguientes pasos, ilustrados en la Fig. 4.5, que serán detallados a lo largo de esta sección:

[1] Dicho parámetro tomará un valor elevado (superior a 0.5), a fin de dar mayor consideración a las medidas de similitud semántica en base a productos —puesto que éstas son las que ofrecen valores de similitud entre usuarios más precisos—.

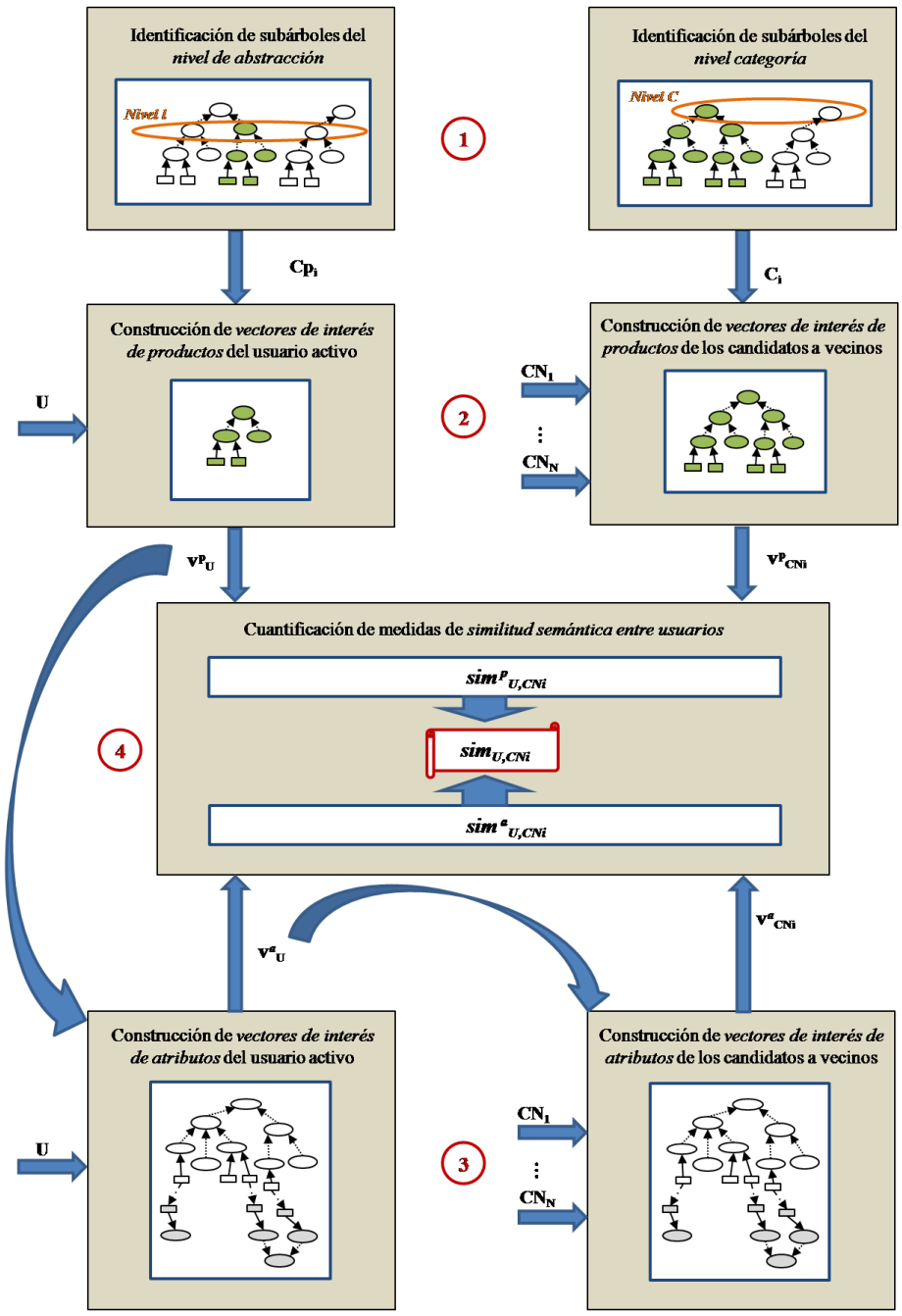


Figura 4.5: Similitud entre usuarios a fin de mitigar el problema de falsos vecindarios

1. Selección del *nivel de abstracción* en la ontología. Identificación de los subárboles delimitados por las clases de dicho nivel y los propios del *nivel categoría*.
2. Construcción de los *vectores de interés de productos* del usuario activo y los candidatos a vecinos.
3. Construcción de los *vectores de interés de atributos* del usuario activo y los candidatos a vecinos.
4. Cuantificación de medidas de *similitud semántica entre usuarios*. Formación de vecindarios para el usuario activo.

Dado que este procedimiento puede realizarse *offline*, durante el proceso de recomendación bastará con seleccionar, para el usuario activo, el vecindario más apropiado según el producto objetivo que esté siendo considerado.

Selección del nivel de abstracción

El punto de partida lo constituye la selección del *nivel de abstracción* en la ontología (ver Sección 3.2). Ésta es una decisión determinante para todo el proceso y de ella depende el número de vecindarios distintos que identificaremos para el usuario activo U. En efecto, las clases de ese nivel, C_{p_i}, serán los nodos raíz de los subárboles que delimitarán los productos a incluir como componentes en el vector de interés de productos de U, v_U^p, en cada caso. Para ilustrar esta idea, consideremos la simplificación de la estructura jerárquica de la ontología de productos comerciales mostrada en la Fig. 4.6. En la figura se muestran, rodeados por líneas discontinuas, los subárboles propios del *nivel categoría* (i.e., del nivel 1) y, rodeados por líneas continuas, los propios de un nivel intermedio l.

Cabe destacar que la selección del nivel de abstracción en la ontología viene marcada por una relación de compromiso entre precisión y *sparsity*. Por un lado, cuanto mayor sea el nivel considerado, más dispersos serán los datos que obtendremos —los vectores de interés de U contendrán un menor número de componentes—. Por otra parte, a medida que descendemos en la jerarquía las clases representan conceptos más específicos, lo que conducirá a descubrimientos más precisos.

Construcción de los vectores de interés de productos

Una vez seleccionado el nivel de abstracción, se construyen los vectores de interés de productos de U, incluyendo como componentes los índices DOI asociados a los productos almacenados en su perfil que correspondan.

- Si el nivel considerado es el nivel categoría, el vector v_U^p contendrá: en un caso los productos clasificados en la ontología bajo el subárbol de la clase *Productos Alimenticios*, en otro caso los productos bajo el subárbol de *Música*, en otro los clasificados bajo el subárbol de *Libros* y así sucesivamente. Dicho de otro modo, el objetivo buscado será identificar usuarios con preferencias similares a las de U respecto a gustos gastronómicos, intereses musicales, libros, etc., de forma independiente.
- De manera análoga, si el nivel considerado es el nivel intermedio marcado en la Fig. 4.6, los índices DOI incluidos en los vectores serán los correspondientes a

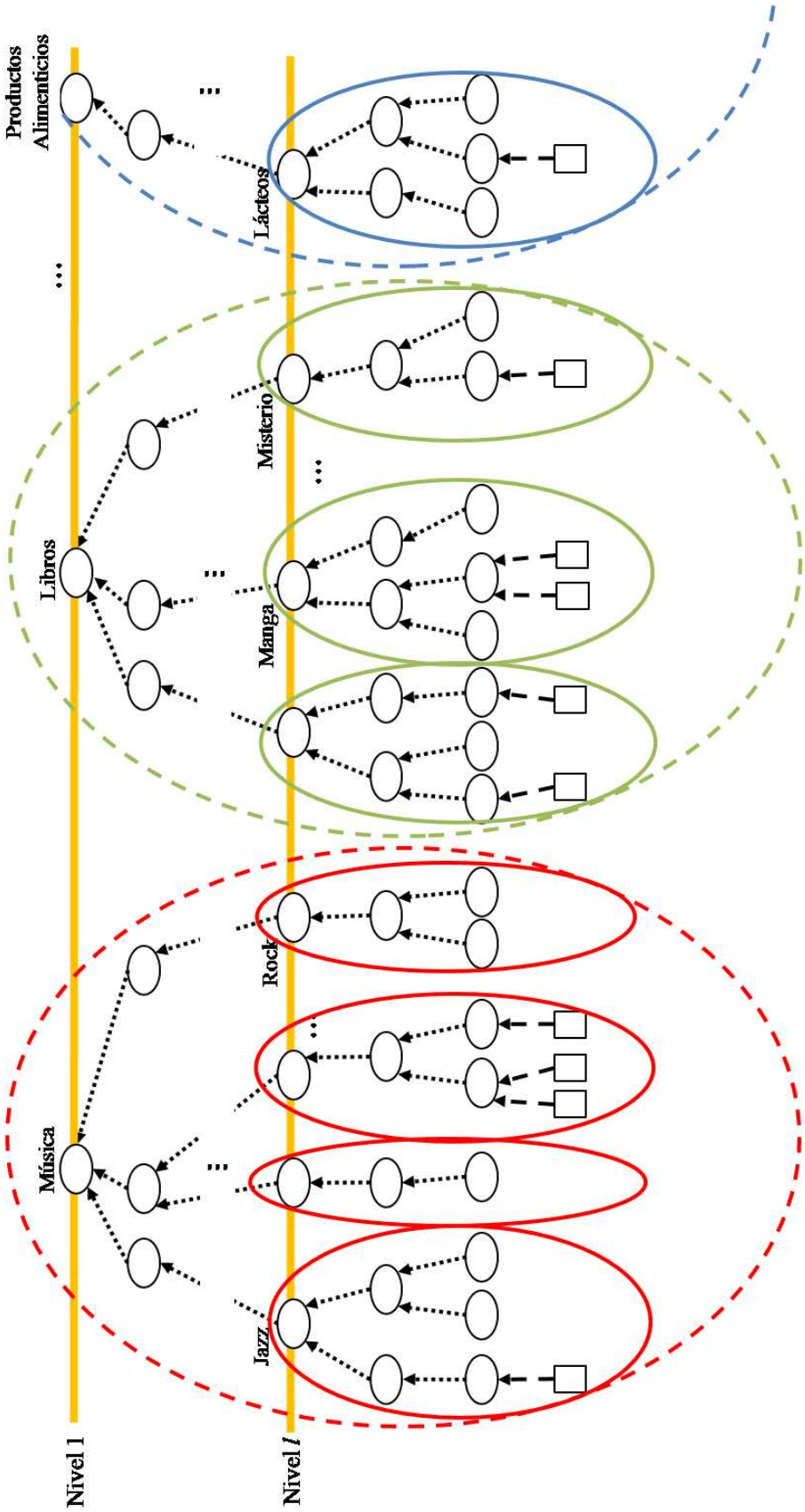


Figura 4.6: Subárboles en función del nivel de abstracción

los productos delimitados por los subárboles propios de las clases de ese nivel en la ontología, más específicas. Por tanto, el objetivo ya no será buscar usuarios con preferencias similares en libros, sino, por ejemplo, respecto a novelas de misterio (productos bajo el subárbol de raíz *Misterio*) o cómics manga (productos bajo el subárbol de *Manga*). Igualmente, no se tratará de identificar candidatos a vecinos afines a los gustos musicales del usuario activo en general, sino, por ejemplo, con preferencias semejantes en música rock (productos clasificados bajo la clase *Rock*) o jazz (productos clasificados bajo la clase *Jazz*).

Los vectores de interés de productos de los candidatos a vecinos, $v^p_{CN_i}$, se obtienen del mismo modo, pero considerando en todo caso el nivel categoría. De esta forma, es posible aprovechar todo el potencial de la estrategia descrita en la Sección 4.3.2 e identificar similitudes entre las preferencias de los usuarios ocultas en la base de conocimiento del sistema, pero sin incurrir en el problema de falsos vecindarios. Esto es, se detectarán similitudes entre los productos registrados en el perfil del usuario activo clasificados bajo el subárbol considerado y todos aquellos productos de los perfiles de los candidatos que pertenezcan a la categoría en cuestión —dotando de una mayor flexibilidad al proceso—.

Construcción de los vectores de interés de atributos

En cuanto a la construcción de los vectores de interés de atributos, el procedimiento cambia respecto al anterior, puesto que, en este caso, la construcción de los vectores de los candidatos depende de los vectores obtenidos para el usuario activo U.

En particular, se diferencian dos etapas en el proceso:

1. En primer lugar, se crean los vectores de interés de atributos de U, atendiendo a los vectores de interés de productos obtenidos para U en el paso previo. Dado un vector v^p_U, que incluye los índices DOI de los productos del perfil de U clasificados bajo una cierta clase C_{p_i} (una de las clases del nivel de abstracción considerado), el vector v^a_U correspondiente tendrá como componentes los niveles de interés de los atributos (propios y generales) asociados a los productos contenidos en v^p_U. Así, si dicho vector contiene productos alimenticios, una de las componentes del vector v^a_U puede ser, por ejemplo, el nivel de interés de ese usuario en el atributo *Picante*. De manera semejante, si el vector de partida, v^p_U, ha sido construido respecto a la clase *Misterio* y el usuario tiene registrados en su perfil libros de de la escritora *Agatha Christie*, el correspondiente vector de atributos incluirá el índice DOI asociado a ese atributo.

2. Una vez construidos los vectores v^a_U, éstos se usan como patrones para construir los vectores de los candidatos. Esto es, dado un vector v^a_U, el vector de interés de atributos de un candidato a vecino CN_i, $v^a_{CN_i}$, contendrá los índices DOI de dicho usuario respecto a cada uno de los atributos considerados previamente para la construcción del vector de interés de atributos del usuario activo[2]. Siguiendo con el ejemplo anterior, en el primer caso se tendrá en cuenta, entre otras características semánticas, si al candidato a vecino en cuestión le gusta la comida picante, y en el segundo supuesto se considerará si también tiene interés por la mencionada escritora.

[2]En caso de que el candidato no tenga registrado esa característica semántica en su perfil, la componente correspondiente será 0.

Cuantificación de medidas de similitud semántica entre usuarios

Por último, para cuantificar las medidas de similitud semántica entre U y cada uno de los usuarios candidatos a vecinos CN_i, basta aplicar la Ec. 4.9 a los vectores de interés construidos para cada una de las clases C_{p_i} del nivel de abstracción considerado, tal como muestra la Ec. 4.10. Por tanto, en esta expresión, $sim^p_{U,CN_i}(C_{p_i})$ se refiere a la similitud semántica en base a productos entre U y CN_i obtenida según el método descrito en la Sección 4.3.2, a partir de los vectores de interés calculados para dichos usuarios respecto a la clase C_{p_i}. Igualmente, $sim^a_{U,CN_i}(C_{p_i})$ se refiere a la similitud semántica en base a atributos entre U y CN_i obtenida según el método descrito en la Sección 4.3.2, utilizando los vectores de interés de atributos construidos en el paso previo en función de la clase C_{p_i}.

$$sim_{U,CN_i}(C_{p_i}) = \theta \cdot sim^p_{U,CN_i}(C_{p_i}) + (1-\theta) \cdot sim^a_{U,CN_i}(C_{p_i}) \tag{4.10}$$

Así, en un cierto proceso de recomendación, el vecindario estará formado por aquellos usuarios que han obtenido mayores valores de similitud con el usuario activo respecto a productos clasificados bajo el mismo subárbol en la ontología que el producto objetivo, y respecto a sus correspondientes atributos.

4.4 Ejemplo de aplicación

Para clarificar lo presentado hasta el momento, en esta sección se ilustra la estrategia propuesta mediante un ejemplo. Por claridad, el escenario de aplicación planteado asume ciertas simplificaciones. Éstas no impedirán, sin embargo, mostrar los beneficios que aporta el razonamiento semántico en dicha estrategia para combatir el problema de falsos vecindarios en un proceso de recomendación colaborativo.

- Por un lado, los perfiles de los usuarios considerados contendrán un número reducido de productos. A modo de referencia, la Tabla 4.1 describe una parte de los perfiles de tres usuarios: el usuario activo del proceso de recomendación en curso, U, y dos candidatos a vecinos, CN_1 y CN_2. En concreto, dicha tabla recoge los productos registrados en los perfiles de esos usuarios clasificados en la ontología bajo la jerarquía *Productos Alimenticios* (en negro) y bajo las jerarquías *Música* y *Libros* (en gris). Dado que asumiremos que el producto objetivo de la recomendación pertenece a la jerarquía *Productos Alimenticios*, únicamente se incluyen en la Tabla 4.1 los atributos relacionados con los productos de esta categoría (en la zona sombreada).
- Por otra parte, no desarrollaremos todo el proceso de forma cuantitativa, sino que nos centraremos en los aspectos más significativos de la estrategia de forma cualitativa. De ahí que en la Tabla 4.1 no se representen los valores numéricos de los índices DOI asociados a los productos y atributos registrados en los perfiles[3].
- La tercera de las simplificaciones está relacionada con la base de conocimiento sobre la que se aplica el razonamiento semántico. Considerando la jerarquía de

[3] En su lugar, se utilizará la siguiente simbología para expresar el nivel de interés de los usuarios: $++$ indica gran interés (índice DOI muy positivo), $+$ interés moderado (índice DOI positivo), $-$ desinterés moderado (índice DOI negativo) y $--$ gran desinterés (índice DOI muy negativo).

U	CN_1	CN_2
Tetilla$_B$ ++	*Tetilla*$_B$ ++	*Tetilla*$_B$ - -
Albariño$_C$ ++	*Albariño*$_A$ ++	*Zanahorias*$_A$ ++
Tabasco$_A$ ++	*Mostaza*$_A$ ++	*Mostaza*$_A$ - -
Cayena$_C$ ++	*Cayena*$_A$ ++	*Cayena*$_A$ - -
Endivias$_B$ - -	*Aceitunas*$_B$ - -	*Endivias*$_B$ ++
Galicia ++	*Galicia* ++	*Galicia* -
T-V +	*T-V* +	*T-V* -
A-V +	*A-V* +	*Dulce* +
Picante ++	*Picante* ++	*Picante* - -
Amargo -	*Amargo* -	*Amargo* +
Tibetan Bells ++	*Tibetan Bells* - -	*Tibetan Bells* ++
Kind of Blue - -	*Groovin'High* ++	*Kind of Blue* - -
Townes - -	*Bouncing with Bud* +	*Townes* -
Kim ++	*Kim* -	*Kim* ++
Total Yoga ++	*Asesinato en el O. E.* ++	*Total Yoga* ++
La Ratonera -	*La Ratonera* ++	*Peanuts* ++
Mafalda ++	*Mafalda* -	*Mafalda* ++
Naruto -	*Naruto* +	*Bola de Dragón* -

Tabla 4.1: Perfiles parciales de tres usuarios

Productos alimenticios, atenderemos al extracto de la misma representado en la Fig. 4.7, que contiene únicamente parte de las características semánticas de los productos incluidos en los perfiles de los usuarios clasificados en la ontología bajo esa jerarquía.

Apoyándose en la Tabla 4.1 y en la Fig. 4.7, este ejemplo ilustra el problema de falsos vecindarios y cómo la estrategia presentada en este capítulo se vale de técnicas de razonamiento semántico a fin de combatirlo eficazmente.

4.4.1 Problema de falsos vecindarios

A la vista de la Tabla 4.1, puede deducirse fácilmente que una estrategia de selección de vecindarios tradicional consideraría a CN_2 un mejor candidato a vecino para U que CN_1. Recordemos que las estrategias convencionales obtienen un único vecindario para cada usuario activo, teniendo en cuenta —a la hora de estimar la similitud entre dos individuos— los índices de interés asignados a *todos* los productos registrados en sus perfiles. Por tanto, dado que U comparte más preferencias (i.e., tiene similares índices DOI de productos) con CN_2 que con CN_1, CN_2 sería seleccionado para formar parte del vecindario de U frente a CN_1.

Sin embargo, si se examina más detenidamente la Tabla 4.1 es posible observar que, pese a compartir intereses con U en el resto de productos, los índices DOI de CN_2 asociados a productos de la categoría *Productos Alimenticios* (resaltados en negro en la tabla) son contrarios a los de aquel. En efecto, las preferencias de CN_2 difieren completamente de las de U en lo que se refiere a gustos gastronómicos, por lo que será falso vecino de U siempre que el producto objetivo de la recomendación sea un producto alimenticio. Por tanto, CN_2 sería falso vecino en el ejemplo que nos ocupa.

En el escenario considerado, la utilización de una estrategia tradicional no sólo llevaría a la selección de CN_2 —siendo éste falso vecino—, además podría no ser incluido en el vecindario CN_1, quien, pese a diferir en el resto de preferencias con U, tiene gustos gastronómicos muy similares a los suyos.

4.4.2 Estrategia semántica propuesta

Explotando la semántica formalizada en una ontología de dominio, tal como describimos en la Sección 4.3.3, es posible agrupar las preferencias registradas en los perfiles de los usuarios en función de la naturaleza de los productos y, de este modo, realizar una selección de vecindarios dinámica, atendiendo en cada caso a las preferencias concernientes a aquellos productos que son del mismo tipo que el objetivo de la recomendación en curso.

Puesto que en este ejemplo el nivel de abstracción considerado es el nivel categoría[4], basta con identificar la jerarquía bajo la que están clasificados en la ontología los productos almacenados en los perfiles de los usuarios. En particular, nos centraremos en aquellos pertenecientes a la jerarquía *Productos Alimenticios* (ver Fig. 4.7 y productos resaltados en negro en la Tabla 4.1) dado que, tal como comentamos anteriormente, es la categoría de productos en cuestión.

[4]Recordemos que en la selección del nivel de abstracción existe un compromiso entre precisión y *sparsity* (ver Sección 4.3.3). Como los perfiles a examinar contienen un número reducido de preferencias, en este caso, considerar un nivel más concreto agravaría el ya existente *sparsity problem*.

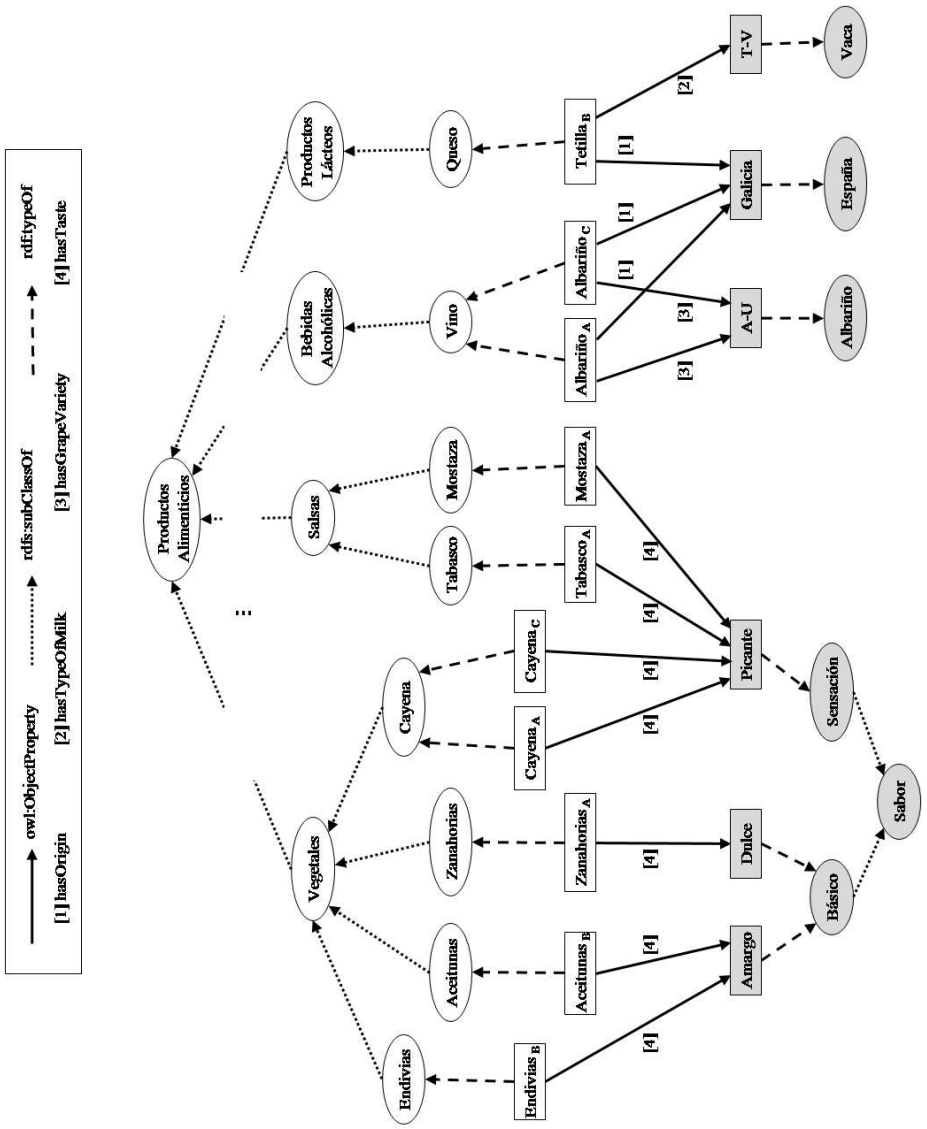


Figura 4.7: Extracto de la ontología considerado en el ejemplo de aplicación

Similitud semántica entre usuarios en base a productos

A fin de obtener la similitud semántica en base a productos entre el usuario activo U y los candidatos a vecinos, respecto a la categoría *Productos Alimenticios*, se construyen los *vectores de interés de productos* de los usuarios incluyendo los índices DOI asociados a los productos pertenecientes a dicha categoría. Así, en base a los perfiles de U, CN_1 y CN_2 representados en la Tabla 4.1, sus respectivos vectores de interés de productos —v_U^p, $v_{CN_1}^p$ y $v_{CN_2}^p$— se construyen como sigue:

$$v_U^p = (DOI_U(Tetilla_B), DOI_U(Albariño_C), DOI_U(Tabasco_A), DOI_U(Cayena_C), DOI_U(Endivias_B))$$

$$v_{CN_1}^p = (DOI_{CN_1}(Tetilla_B), DOI_{CN_1}(Albariño_A), DOI_{CN_1}(Mostaza_A), DOI_{CN_1}(Cayena_A), DOI_{CN_1}(Aceitunas_B))$$

$$v_{CN_2}^p = (DOI_{CN_2}(Tetilla_B), DOI_{CN_2}(Zanahorias_A), DOI_{CN_2}(Mostaza_A), DOI_{CN_2}(Cayena_A), DOI_{CN_2}(Endivias_B))$$

Tal como comentamos en la Sección 4.3.2, el tamaño de estos vectores y el orden de sus componentes no tiene que ser necesariamente el mismo, ya que la métrica de similitud propuesta permite comparar cada componente de v_U^p con cada una de las componentes de los vectores de los candidatos (ver Ec. 4.7).

En concreto, nuestra estrategia cuantifica la similitud entre cada par de índices DOI comparados (uno del vector v_U^p y uno del vector del candidato considerado) teniendo en cuenta la similitud semántica —jerárquica e inferencial (ver Sección 4.2)— existente entre los productos correspondientes. Para ilustrar en qué medida la estrategia semántica propuesta permite enriquecer el proceso, examinemos la Fig. 4.7:

- Frente a un enfoque sintáctico, que únicamente permitiría detectar una común preferencia entre U y CN_1 ($Tetilla_B$) y dos preferencias contrarias entre U y CN_2 ($Tetilla_B$ y $Endivias_B$), la explotación de la semántica asociada a los productos comerciales posibilita la comparación de productos que, aun estando identificados con distintos términos (e.g., distintas marcas comerciales), son del mismo tipo. Este es el caso de los vinos $Albariño_C$ y $Albariño_A$, o los vegetales $Cayena_C$ y $Cayena_A$. Nuestra estrategia descubre estas relaciones gracias a la similitud semántica jerárquica (ver Sección 4.2.1). Precisamente, los ejemplos anteriores obtienen un valor máximo como resultado de la Ec. 4.2.

- Además de considerar la similitud semántica jerárquica de los productos, la similitud inferencial posibilita descubrir que existe una relación más significativa entre, por ejemplo, los productos $Endivias_B$ y $Aceitunas_B$ que entre éstos y $Zanahorias_A$, puesto que aquéllos comparten el atributo *Amargo*[5] —que no es una característica semántica de $Zanahorias_A$—. Nótese que atendiendo únicamente a la similitud jerárquica los tres productos obtendrían los mismos valores de si-

[5]En este punto, conviene recordar la simplicidad del ejemplo que nos ocupa, ya que en un escenario complejo se considerarían un gran número de características semánticas en el cálculo de similitud inferencial entre dos productos.

militud, ya que se encuentran al mismo nivel de profundidad y el *LCA* común a todos ellos es *Vegetales*.

- Por otro lado, tal como detallamos en la Sección 4.2.2, la similitud semántica inferencial permite identificar relaciones entre los productos ocultas en la base de conocimiento. Por ejemplo, es posible detectar una mayor relación entre las salsas $Tabasco_A$ y $Mostaza_A$ con el vegetal $Cayena_A$ (o $Cayena_C$) que con cualquier otro vegetal de los representados en el escenario de la Fig. 4.7, puesto que éstos comparten el atributo *Picante*.

Es razonable concluir que la estrategia propuesta permite aliviar el *sparsity problem*, puesto que facilita la comparación de las preferencias de los usuarios de forma significativa aun cuando el solapamiento existente entre los vectores de interés de los usuarios comparados es mínimo.

Similitud semántica entre usuarios en base a atributos

Tal como describimos en la Sección 4.3.2, nuestro enfoque no sólo se vale de las características semánticas de los artículos comerciales a la hora de medir la similitud inferencial entre dos productos, sino que aprovecha la información adicional que éstas proporcionan acerca de las preferencias de los usuarios para obtener una nueva medida de similitud entre ellos. En efecto, el que un cierto atributo aparezca repetidamente ligado a diversos productos registrados en el perfil de un usuario refleja intereses no considerados anteriormente en el proceso de selección de vecindarios.

Por ejemplo, nuestra métrica tendrá en cuenta que los atributos más relevantes para U en cuanto a sus preferencias gastronómicas son *Galicia* y *Picante*. Precisamente, U ha mostrado un gran interés por varios productos de origen gallego ($Tetilla_B$ y $Albariño_C$) y por alimentos que provocan una sensación picante en el sentido del gusto ($Tabasco_A$ y $Cayena_C$). Esto permite descubrir una mayor similitud entre U y CN_1, puesto que el perfil de CN_1 también refleja un gran interés de dicho usuario por productos de origen gallego ($Tetilla_B$ y $Albariño_A$) y por el picante ($Mostaza_A$ y $Cayena_A$).

4.5 Beneficios de la estrategia

La estrategia presentada en este capítulo modifica el proceso de selección de vecindarios tradicional —basado en identificar aquellos usuarios cuyas preferencias sean lo más similares a las del usuario activo— a fin de combatir el problema de falsos vecindarios. Construyendo distintos vecindarios para cada usuario activo, en función de los diferentes tipos de productos disponibles, identificamos qué candidatos a vecinos son los más apropiados en cada recomendación concreta.

Tal como adelantábamos en la Sección 4.1, con nuestro enfoque no sólo tratamos de evitar incluir en el vecindario candidatos que resultan contraproducentes para la recomendación de según qué productos (i.e., aquellos que comparten gran cantidad de intereses con el usuario activo, salvo los de ese tipo), sino que también identificamos candidatos con un gran potencial para obtener buenos resultados en la recomendación de ciertos productos que pasarían desapercibidos en las técnicas tradicionales (porque aunque sus preferencias sean muy similares a las del usuario activo en productos de una índole, difieren en el resto).

La aproximación semántica utilizada en nuestra metodología aporta grandes beneficios a la estrategia propuesta:

- Por su naturaleza, la estructura jerárquica definida en la ontología permite clasificar de forma significativa los productos a considerar en cada selección de vecindario. Más aun, permite dotar de gran flexibilidad al proceso, pudiendo realizar dicha clasificación no sólo por categorías de productos (por ejemplo, *Música*, *Libros*, *Equipamiento Deportivo*, etc.), también por subcategorías más concretas (por ejemplo, en lugar de considerar todos los productos de la categoría *Música*, considerar únicamente los clasificados bajo su subclase *Clásica*) —eligiendo un nivel de abstracción de la ontología u otro—.

- Tal como expusimos en la Sección 4.3.2, su potencialidad para mitigar el *sparsity problem* es evidente, ya que permite detectar similitudes entre las preferencias de los usuarios aunque éstos no tengan clasificados en sus perfiles exactamente los mismos productos. Asimismo, la técnica de modelado de usuarios utilizada en nuestro enfoque posibilita la incorporación de información adicional acerca de los intereses de los usuarios en el proceso, detectando afinidades en sus preferencias respecto a características semánticas de los productos.

- Además, explotando técnicas de razonamiento sobre la ontología, es posible descubrir conocimiento útil sobre las preferencias de los usuarios. La identificación de asociaciones semánticas complejas entre los productos almacenados en el perfil del usuario activo y los perfiles de los candidatos a vecinos permite inferir similitudes entre ellos ocultas en la base de conocimiento, mejorando notablemente el proceso de selección de vecindarios.

Capítulo 5

Estrategia para la inclusión de un modelo de confianza implícita

La estrategia presentada en este capítulo se sirve de la semántica formalizada en la ontología de dominio para construir redes de confianza implícita entre los usuarios de nuestro sistema recomendador colaborativo. Dicha estrategia persigue incorporar en el proceso de selección de vecindarios un modelo de confianza que no requiera información explícita por parte de los usuarios, a fin de mejorar la precisión y fiabilidad de las recomendaciones de forma trasparente a los mismos. En primer lugar, nuestro enfoque semántico permite inferir relaciones de confianza significativas entre pares de individuos, condicionadas a clases de productos, a partir de resultados obtenidos en recomendaciones previas. A continuación, las relaciones inferidas son organizadas formando redes de confianza implícita entre todos los usuarios del sistema. Explorando las estructuras resultantes, la metodología propuesta descubre medidas de reputación individuales. Finalmente, los valores de confianza y reputación son incorporados en el proceso de formación de vecindarios.

5.1 Introducción

Desde la aparición de los sistemas recomendadores colaborativos, las estrategias empleadas en el proceso de selección de vecindarios se han fundamentado en la búsqueda de individuos con preferencias similares a las de un cierto usuario activo.

Sin embargo, tal como adelantamos en la Sección 2.4, al atender únicamente a relaciones de similitud, los enfoques tradicionales son incapaces de detectar qué información proviene de usuarios confiables y cuál no, por lo que los resultados de las recomendaciones pueden verse distorsionados por la influencia de usuarios maliciosos. Además de combatir este problema, el incorporar información acerca de la fiabilidad de los individuos como un factor más a la hora de identificar usuarios afines permite enriquecer notablemente el proceso de formación de vecindarios. En efecto, si se dispone de las relaciones de confianza existentes entre los usuarios, es posible tener en

cuenta en el proceso de selección qué candidatos son considerados fidedignos según el criterio del usuario activo en cuestión.

Por este motivo, en los últimos años ha crecido el interés en la comunidad investigadora por el desarrollo de estrategias de selección de vecindarios *trust-aware*. Estas estrategias incorporan modelos de confianza en el proceso, a fin de complementar (o incluso sustituir) la búsqueda de vecindarios convencional, basada en similitud, con mecanismos de selección que atienden a las relaciones de confianza existentes entre el usuario activo y los candidatos a vecinos. La idea es, por tanto, considerar no sólo si los candidatos tienen preferencias similares a las del usuario activo en cuestión, también si son o no fiables. Generalmente, los modelos de confianza incluidos en estas estrategias consisten en una red en la que los nodos representan a los usuarios del sistema y los arcos simbolizan las relaciones de confianza existentes entre ellos.

No obstante, la mayoría de los planteamientos propuestos hasta la fecha dependen de la posibilidad de formación de redes de confianza *explícita* entre los usuarios, es decir, requieren que los usuarios especifiquen expresamente sus relaciones de confianza —indicando en quién confían, en quién no y cuánto—. En consecuencia, la aplicación de estos enfoques se ve seriamente limitada:

- Por un lado, necesitan información que los usuarios pueden no estar dispuestos a proporcionar, bien por pereza —puesto que valorar una a una las relaciones de confianza con otros individuos es una tarea tediosa—, bien por no querer facilitar ese tipo de información, de carácter personal.

- Por otra parte, no son aptos para ser utilizados en sistemas abiertos donde los individuos, por norma general, no se conocen los unos a los otros.

La estrategia que presentamos en este capítulo, descrita en [86], infiere relaciones de confianza *implícita* entre los usuarios del sistema recomendador a partir de los resultados obtenidos en recomendaciones previas. El objetivo buscado es incorporar medidas de confianza significativas en el proceso de formación de vecindarios de forma transparente a los usuarios, sin requerir que estos proporcionen ningún tipo de información adicional.

A tal fin, nuestra estrategia toma como punto de partida el enfoque formulado por O'Donovan y Smyth en [106] (introducido en la Sección 2.4.2), basado en derivar relaciones de confianza atendiendo a los nexos establecidos de forma automática en cada proceso de recomendación entre el usuario activo y sus respectivos vecinos, y lo enriquece explotando la semántica de los productos que han sido objeto de tales recomendaciones. Precisamente el citado enfoque, si bien destaca por su cualidad implícita a la hora de obtener relaciones de confianza entre usuarios de un sistema recomendador colaborativo, está limitado por su naturaleza sintáctica.

La exploración de la clasificación jerárquica bajo la que están categorizados los productos comerciales en la ontología provee a nuestra metodología de nuevo conocimiento, que será incorporado al proceso de inferencia para obtener relaciones de confianza implícita condicionadas a clases de productos. En particular, la posibilidad de considerar diferentes *niveles de abstracción* (ver Sección 3.2) a la hora de examinar los resultados obtenidos en recomendaciones previas permite dotar a nuestro enfoque de una gran flexibilidad.

Más aun, la estrategia propuesta no se limita a obtener relaciones de confianza entre pares de individuos, sino que reúne dichas relaciones y construye redes de confianza

implícita entre todos los usuarios del sistema, también condicionadas a clases de productos —del nivel de abstracción considerado en el proceso de inferencia—. Las redes resultantes facilitan la estimación de medidas individuales de reputación, mediante la propagación de los valores de confianza recibidos por cada usuario. Estas medidas globales de fiabilidad pueden ser empleadas, a falta de las pertinentes relaciones de confianza entre el usuario activo en cuestión y los candidatos a vecinos, para enriquecer la fase de selección de vecindarios de nuestro sistema recomendador colaborativo.

5.2 Cuantificación de relaciones de confianza entre pares de usuarios

En esta sección detallamos cómo nuestra estrategia se vale de la ontología de productos para inferir relaciones de confianza entre pares de usuarios a cualquier nivel de abstracción de la misma. Dado que dichas relaciones serán estimadas a partir de los resultados obtenidos en recomendaciones anteriores, definimos previamente qué entendemos por éxito o fracaso en una recomendación y exponemos la motivación de emplear un enfoque semántico en el proceso.

5.2.1 Relaciones de confianza

Desde que fueron introducidos en la Sección 2.3.3, los términos *usuario activo* y *vecino* han sido utilizados para designar a los distintos participantes de un sistema recomendador colaborativo. Llegados a este punto cabe puntualizar que, en todo momento, nos referimos a un determinado subconjunto de la misma comunidad de individuos (los usuarios del sistema), ejerciendo una u otra función en el proceso de recomendación en curso. Esto es, un usuario de un sistema recomendador colaborativo puede desempeñar dos roles: usuario activo, cuando es quien recibe las sugerencias del sistema en un momento dado; y vecino, cuando es seleccionado como parte del grupo que determina si un producto objetivo se recomienda a un cierto usuario activo o no.

Con esta idea en mente, la fiabilidad de un usuario será siempre inferida de sus contribuciones como vecino. En otras palabras, cuantificaremos positivamente o negativamente una relación de confianza entre dos usuarios dependiendo de si uno ha contribuido (como vecino) en recomendaciones que han sido o no del agrado del otro (en su rol de usuario activo).

Gráficamente, representamos la relación de confianza entre dos usuarios A y B con un arco dirigido de A a B, cuyo valor asociado indica cuánto A confía o desconfía de B —como resultado de las contribuciones de B en recomendaciones a A (ver Fig. 5.1)—.

5.2.2 Definición de éxito

Tal como describimos en la Sección 2.3.3, en una recomendación colaborativa el sistema sugiere (o no) un determinado producto objetivo al usuario activo dependiendo de si encaja con las preferencias del grupo de usuarios afines que han sido seleccionados para formar su vecindario —i.e., en función de las valoraciones registradas en los perfiles de dichos vecinos en relación al producto en cuestión—.

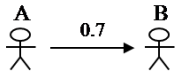


Figura 5.1: Representación gráfica de una relación de confianza

Para la definición de *éxito*, adaptamos la expresión dada por O'Donovan y Smyth en [106] para su utilización con los *niveles de interés* (ver Sección 3.3.4) empleados en nuestro enfoque. En concreto, consideramos la contribución de un vecino N en la recomendación del producto objetivo i al usuario activo U un éxito si la diferencia entre el nivel de interés mostrado por U —después de la recomendación en cuestión, como realimentación de relevancia explícita o implícita (ver Sección 3.3.3)— y el registrado en el perfil de N es menor que un límite de error dado, de acuerdo a la expresión mostrada en la Ec. 5.1.

$$E_N i, U = |DOI_N(i) - DOI_U(i)| = \varepsilon \in [0,2] \tag{5.1}$$

Cabe destacar que esta expresión aísla la contribución de N del resultado global de la recomendación, cuantificando si su contribución concreta ha tenido un efecto positivo o negativo en el resultado final independientemente de los otros vecinos. De este modo, es posible detectar éxitos de N no sólo cuando ha favorecido (con un índice DOI positivo) en recomendaciones que han sido aceptadas por U, sino también cuando ha ido en contra de una recomendación (con un índice DOI negativo) que finalmente ha sido rechazada por el usuario activo —la recomendación se hizo debido a las contribuciones del resto de vecinos—.

5.2.3 Enfoque semántico vs. enfoque sintáctico

Las medidas de éxito o fracaso obtenidas como resultado de la Ec. 5.1 sirven de base en el trabajo de O'Donovan y Smyth [106] para obtener automáticamente relaciones de confianza entre pares de usuarios. Sin embargo, tal como comentamos en la sección Sección 2.4.2, la naturaleza sintáctica de dicho enfoque hace que esté limitado a construir tales relaciones atendiendo únicamente a las palabras que identifican al producto recomendado. En consecuencia, teniendo en cuenta que las recomendaciones suelen tener como objeto un producto concreto, solamente es posible construir dos tipos de relaciones de confianza:

- *Condicionadas a ítem*: Directamente de los resultados de las recomendaciones pueden inferirse relaciones como, por ejemplo, "*A* confía en *B* para recomendar el disco titulado *Kind of Blue*". Es obvio que este tipo de relaciones conducen a un problema de *sparsity*, ya que tratar de seleccionar vecinos confiables para recomendar un producto específico (de entre la miríada de productos disponibles) puede ser como buscar una aguja en un pajar.

- *Condicionadas a usuario*: Reuniendo todas las contribuciones de un usuario en recomendaciones a otro pueden obtenerse relaciones del tipo "*A* confía en *B*", pero estas relaciones suponen una *sobregeneralización*. En efecto, en dominios de aplicación donde existe una gran diversidad en la naturaleza de los productos que pueden ser objeto de una recomendación, es muy improbable que un usuario sea considerado fiable recomendando todo tipo de productos.

Es por estos motivos que nuestra estrategia lleva el enfoque de O'Donovan y Smyth un paso más allá, incorporando en el proceso la información semántica disponible en la ontología de productos, a fin de obtener relaciones de confianza *condicionadas a clase* en cualquier nivel de la jerarquía de productos considerada —i.e., inferir cuánto confía un usuario en otro para recomendar productos de una clase determinada—.

Como hemos mencionado anteriormente, la semántica formalizada en una base de conocimiento posibilita el razonamiento sobre los conceptos que encierran las palabras. En el caso que nos ocupa, esto se traduce en poder identificar cada producto no sólo por los vocablos con los que está registrado en el sistema, también por el tipo de producto que representa (ver Sección 3.2). De este modo, por ejemplo, el producto *Kind of Blue* no será únicamente identificado por esas tres palabras, sino que lo reconoceremos como un producto musical, en concreto, un disco de jazz bebop titulado *Kind of Blue*. Tal conocimiento es fácilmente deducido del extracto de la jerarquía *Música* representado en la Fig. 5.2, donde el producto *Kind of Blue* se corresponde con una instancia de la clase hoja *Bebop* y, por tanto, también de su superclase *Jazz*.

5.2.4 Inferencia de relaciones de confianza condicionadas a clase

Con el propósito de definir relaciones de confianza implícita entre dos usuarios condicionadas a una clase de productos dada, nuestra estrategia explota la estructura jerárquica bajo la que se encuentran clasificados los productos que han sido objeto de recomendaciones anteriores en las que han participado ambos usuarios —uno como usuario activo y el otro como vecino—.

En primer lugar, nuestra aproximación obtiene valores de confianza en el *nivel hoja*, a partir de las recomendaciones previas de productos clasificados bajo cualquier clase hoja. Por ejemplo, será posible inferir que "*A* confía en *B* para recomendar discos de bebop" a partir de los éxitos de *B* en las recomendaciones a *A* de productos que pertenecen a la clase *Bebop* —por ejemplo, el disco *Kind of Blue*, el denominado *Groovin'High* o el titulado *Bouncing with Bud* (ver Fig. 5.2)—.

En concreto, cuantificamos el valor de una relación de confianza de un usuario activo U a un vecino N respecto a la clase hoja C_L como la media de los éxitos y fracasos obtenidos por N en sus contribuciones en recomendaciones de productos (clasificados en la ontología bajo esa clase) hechas a U. Esta idea aparece reflejada en la Ec. 5.2, donde $E_N i, U$ es la medida del éxito o fracaso de N en su contribución a la recomendación del producto i a U (ver Ec. 5.1) y M el número de recomendaciones de productos de la clase C_L hechas al usuario U en las que N ha participado como vecino.

$$IT_{U,N}(C_L) = \frac{1}{M}\sum_{i=1}^{M}(1 - E_N i, U) \tag{5.2}$$

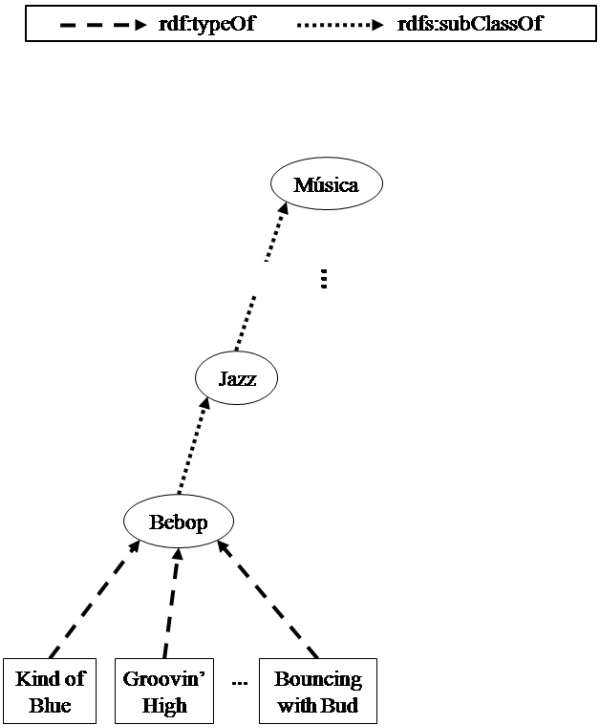


Figura 5.2: Clasificación jerárquica de la clase *Bebop*

De acuerdo a la Ec. 5.2, el valor de las relaciones de confianza inferidas oscila entre -1 y 1, siendo -1 el valor que se atribuye a una relación de desconfianza máxima y 1 el que simboliza una relación de confianza máxima.

Una vez inferidos los valores de confianza en las clases hoja que correspondan, nuestra estrategia obtiene relaciones de confianza en cualquier otro nivel mediante la propagación de dichos valores a través de la jerarquía. Siguiendo con el ejemplo anterior, el nivel previo se corresponde con la clase *Jazz*, por lo que el valor resultante reflejará las contribuciones de B en todas las recomendaciones a A de discos de cualquier género de jazz.

Para ello, nuestra metodología adopta el enfoque propuesto por Ziegler en [149] (descrito en la Sección 2.5.1), aplicado en este caso a la propagación de valores de confianza en lugar de niveles de interés. En particular, a la hora de propagar un valor de una clase a su respectiva superclase, se tiene en cuenta el nivel de dicha superclase y el número de hermanos que tenga en la jerarquía la clase desde la que se propaga el valor. Tal como muestra la Ec. 5.3, los mayores valores corresponden a aquellas superclases más específicas de la jerarquía, es decir, más cercanas a las clases hoja de la misma. Asimismo, el valor de confianza condicionado a la superclase C_{p+1} será más significativo cuanto mayor sea el valor propagado por la clase C_p —subclase de C_{p+1}—, y cuanto menor sea el número de hermanos de ésta.

$$IT_{U,N}(C_{p+1}) = \frac{IT_{U,N}(C_p)}{1 + \#sib(C_p)} \tag{5.3}$$

La Ec. 5.3 sólo permite cuantificar el valor 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 el valor final. En nuestro enfoque, el valor de confianza condicionado a una superclase se calcula sumando los valores que propagan (mediante la Ec. 5.3) cada una de las clases que dependen directamente de ella en la jerarquía de productos. De esta forma, cuanto mayor sea la proporción de subclases que propaguen un valor no nulo (i.e., tengan asociada una relación de confianza implícita), más significativo será el valor que recibe la superclase en cuestión.

El menor nivel de abstracción contemplado es el *nivel categoría* (nivel 1), a fin de evitar el problema de sobregeneralización antes mencionado. En el ejemplo que nos ocupa la clase categoría es *Música*, de modo que, cuando el nivel considerado por el sistema sea el nivel categoría, se obtendrán relaciones del tipo "*A* confía en *B* para recomendar música".

5.3 Construcción de redes de confianza implícita

A partir de las relaciones de confianza implícita obtenidas entre pares de usuarios mediante el método descrito en la sección previa, nuestra estrategia construye redes de confianza entre los usuarios del sistema. En esta sección nos centramos en el proceso de formación de dichas redes, cuyo número variará en función del nivel de abstracción considerado en el proceso de inferencia de relaciones.

Independientemente del tipo de relaciones de confianza que las conformen, los nodos de las redes resultantes simbolizan en cualquier caso a los usuarios dados de alta en el sistema recomendador. Por consiguiente, para identificar los valores asociados a los roles que un usuario ha representado en recomendaciones previas (como usuario activo, vecino, o ambos), basta con observar las direcciones de los arcos conectados a su nodo en una red dada: los arcos dirigidos a dicho nodo se corresponden con valoraciones de su rol como vecino, mientras que los arcos que tienen su origen en él son debidos a su rol de usuario activo.

Sin embargo, los arcos que conectan a los nodos serán diferentes en cada red, puesto que, en función de la clase de productos considerada, se habrán inferido unas u otras relaciones de confianza entre distintos pares de usuarios.

Además, tal como hemos adelantado, el número de redes de confianza derivadas será diferente dependiendo del nivel seleccionado en la ontología de dominio. Precisamente, se construirán tantas redes como clases haya definidas en la ontología en dicho nivel de abstracción.

- Si el nivel tomado como punto de partida es el nivel hoja, se formarán tantas redes como clases hoja tenga la ontología de productos. En cada red se incluyen, por tanto, las relaciones establecidas entre los usuarios condicionadas a una cierta clase hoja.

  En la Fig. 5.3 se ilustra cómo cada clase hoja (lc_1 a lc_M) se corresponde con una red (en la parte central inferior de la figura). Un pequeño extracto de una de estas

redes se muestra en la parte derecha de la Fig. 5.3, cuyos arcos representan, por ejemplo, valores de confianza implícita condicionados a la clase *Bebop*.

- En el extremo opuesto, si el nivel considerado es el nivel categoría, habrá tantas redes como categorías (i.e., clases de nivel 1) estén definidas en la ontología (cc_1 a cc_N en la Fig. 5.3). Por consiguiente, cada red contiene relaciones de confianza implícita condicionadas a una determinada categoría.

  Por ejemplo, una red puede incluir arcos que representan valores de confianza condicionados a la clase *Música*, otra arcos que representan valores condicionados a la clase *Libros*, etc.

Al igual que en nuestra estrategia para mitigar el problema de falsos vecinos, descrita en el capítulo previo, la selección del nivel de abstracción en la ontología establece una relación de compromiso entre precisión y *sparsity*. Por un lado, cuanto mayor sea el nivel considerado, mayor será el número de redes que se formen ($N \ll M$, en la Fig. 5.3) y habrá una mayor dispersión de datos. Por otra parte, a medida que descendemos en la jerarquía las clases representan conceptos más específicos, lo que conducirá a descubrimientos más precisos.

5.4 Cuantificación de medidas de reputación

Entendiendo *confianza* como "lo fiable que un usuario considera a otro", el concepto de *reputación* puede ser definido como "lo fiable que un usuario es considerado en general". Es decir, si bien las relaciones de confianza son establecidas de forma independiente entre pares de individuos, las medidas de reputación reflejan la confianza global de un grupo en un individuo.

Con esta idea en mente, una vez formadas las redes de confianza implícita condicionadas a clases de productos de un nivel de abstracción dado, nuestra metodología explora dichas redes para obtener medidas de reputación —i.e., para cuantificar cómo de fiable es valorado cada usuario por el conjunto de la comunidad—. Al igual que para obtener valores de confianza, la reputación de un usuario está ligada a su rol de vecino, puesto que el objetivo buscado es estimar la fiabilidad de dicho usuario a la hora de recomendar productos de un cierto tipo.

Para ello, nuestra estrategia toma como punto de partida una cierta red de confianza, a fin de determinar medidas de reputación condicionadas a su respectiva clase de productos. A modo de ejemplo, consideremos la red correspondiente a la clase *Bebop* representada en la Fig. 5.3 y el usuario A. Cada arco dirigido a A representa cuánto confía el usuario del que parte dicho arco (en este escenario, D, E o F) en A para recomendar discos de jazz bebop. En consecuencia, la medida de reputación de A obtenida por medio de la exploración de esa red expresará cómo de fiable es considerado A, en general, para recomendar discos de jazz bebop.

En concreto, dado un usuario (en su rol de vecino) N y una clase de productos C_p, nuestro enfoque cuantifica la reputación de N condicionada a C_p mediante la Ec. 5.4. En esta expresión, L es el número de arcos dirigidos a N, y U_i el usuario del que parte el i-ésimo arco.

$$IR_N(C_p) = 0{,}5(1 + \frac{1}{L}\sum_{i=1}^{L} IT_{U_i,N}(C_p) \cdot IR_{U_i}(C_p)) \tag{5.4}$$

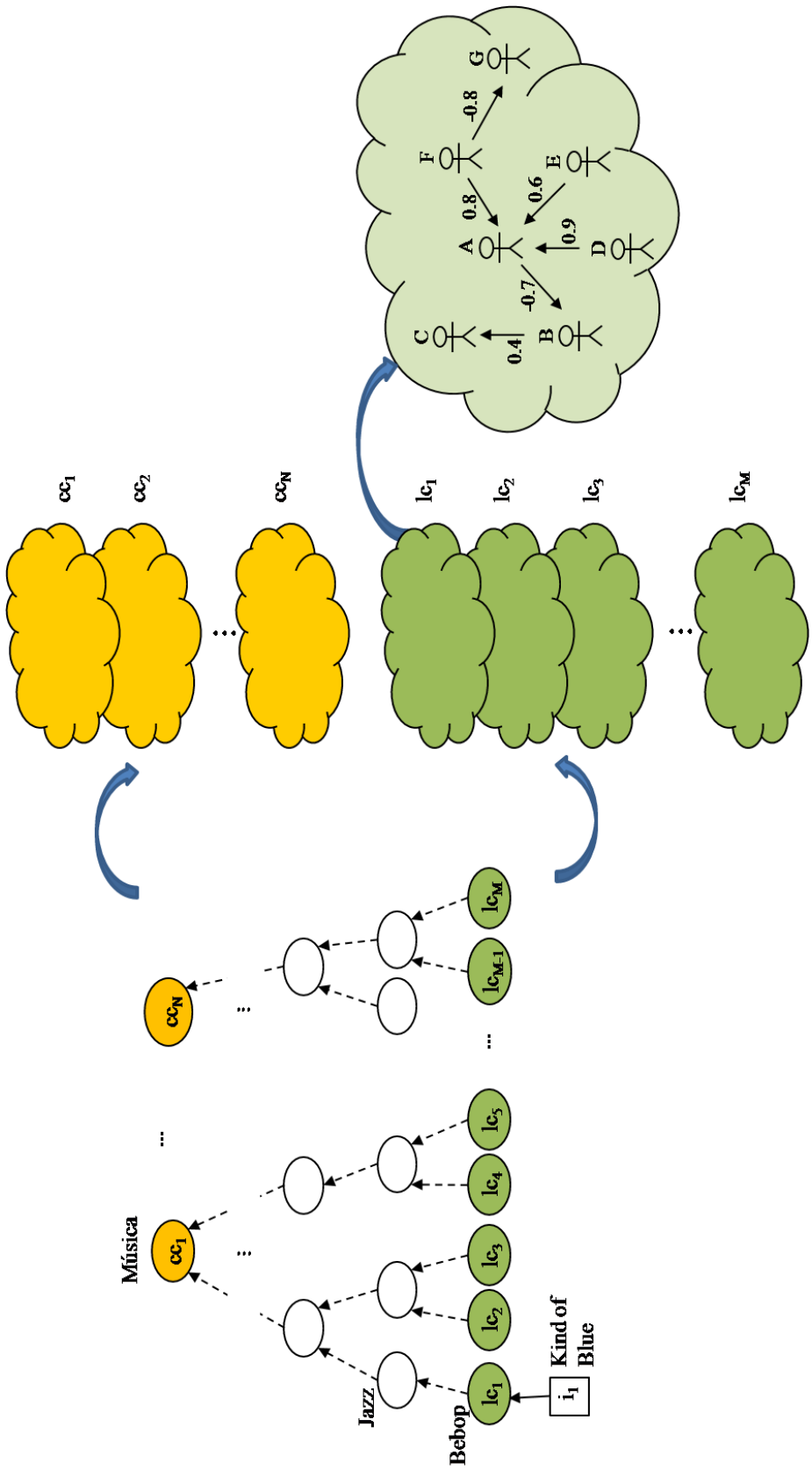


Figura 5.3: Construcción de redes de confianza implícita

De acuerdo a la Ec. 5.4, la métrica de reputación propuesta reúne las valoraciones de confianza que un usuario recibe de otros, teniendo en cuenta a su vez la reputación de estos últimos. Precisamente, en ella se pondera cada valor de confianza respecto a la reputación del usuario que lo origina. De esta forma, las valoraciones dadas por usuarios confiables tendrán un mayor peso en el resultado final.

Por lo tanto, cuanto más positivos sean los valores de las relaciones de confianza hacia un usuario y cuanto más fiables sean los usuarios de los que parten dichas relaciones, mayor será la medida de reputación obtenida por aquel. Para ilustrar esta idea, consideremos el usuario B en el escenario mostrado en la parte derecha de la Fig. 5.3. A fin de obtener la reputación de B hemos de centrarnos en los arcos dirigidos a él. En este caso, encontramos solamente uno, proveniente de A, cuyo valor es negativo (i.e., A desconfía de B). Dicho valor ha de ser ponderado respecto a la reputación de A. Nótese que el usuario A recibe tres valores de confianza, de D, E y F, todos ellos positivos (i.e., todos confían en A, en mayor o menor medida). En consecuencia, B obtiene una medida de reputación muy mala.

Cabe destacar que cada vez que un valor de reputación es renovado, dicho cambio se propaga a través de toda la red, actualizando los valores de reputación correspondientes. Siguiendo con el ejemplo anterior, si la reputación de F cambia (debido a nuevas contribuciones como vecino en recomendaciones de productos de ese tipo) provoca un cambio en los valores de reputación de A y G. A su vez, el valor actualizado de A provoca la modificación del valor de reputación asociado a B, y así sucesivamente.

5.5 Selección de vecindario *trust-aware*

Nuestro enfoque incorpora en el proceso de formación de vecindario los valores de confianza y reputación obtenidos según los métodos presentados en este capítulo, complementando las medidas de similitud inferidas mediante nuestra estrategia para combatir el problema de falsos vecindarios, descrita en el Capítulo 4. En particular, ahora se tendrá en cuenta la fiabilidad de los candidatos a vecinos que han sido seleccionados con aquella estrategia.

Recordemos que mediante el proceso detallado en la Sección 4.3.3, una vez elegido un determinado nivel de abstracción en la ontología, se obtienen medidas de similitud semántica entre dos usuarios respecto a cada una de las clases de productos definidas en dicho nivel. De esta forma, es posible disponer de diferentes vecindarios para un usuario activo dado, y seleccionar el más apropiado en cada proceso de recomendación en función del producto objetivo considerado.

A fin de incorporar en la fase de selección de vecindarios la fiabilidad de los candidatos a vecinos, el enfoque propuesto atiende tanto al valor de similitud inferido entre el usuario activo y cada candidato (respecto a una cierta clase de productos) como a la relación de confianza existente entre ellos o, en su defecto, la reputación de este último —en cualquier caso, valores condicionados a la susodicha clase—. De este modo, en un determinado proceso de recomendación, se tendrán en cuenta las medidas de similitud, confianza y reputación de los usuarios en relación a la clase (del nivel de abstracción considerado) bajo la que esté clasificado el producto objetivo.

Por tanto, dado un usuario activo U y una clase de productos C_p:

- Si el candidato a vecino considerado, CN_i, ha participado como vecino en recomendaciones previas a U de productos pertenecientes a la clase en cuestión,

es posible utilizar el valor de confianza implícita inferido entre ellos. Basta con localizar el arco dirigido de U a CN_i en la red de confianza correspondiente.

En este escenario, la idoneidad de dicho candidato a formar parte del vecindario de U —$S_U^{C_p}(CN_i)$— se calcula mediante la Ec. 5.5, donde $IT_{U,CN_i}(C_p)$ es el valor de confianza implícita de U a CN_i, $Sim_{U,CN_i}(C_p)$ es el valor de similitud entre ellos respecto a la clase de productos considerada y φ es un parámetro de combinación que modula el peso de cada componente[1].

$$S_U^{C_p}(CN_i) = \varphi \cdot Sim_{U,CN_i}(C_p) + (1-\varphi) \cdot IT_{U,CN_i}(C_p) \tag{5.5}$$

- Si el candidato a vecino CN_i nunca ha participado en recomendaciones a U de productos clasificados bajo C_p en la ontología, no existe un valor de confianza implícita inferido entre ellos. En definitiva, no existe un arco que los una en la red de confianza pertinente. En este caso, se pondera el valor de similitud entre U y CN_i por la medida de reputación de CN_i —obtenida mediante el método presentado en la sección previa—.

  Teniendo en cuenta que las medidas de reputación ofrecen una información menos significativa para el proceso de selección que las relaciones de confianza, ya que no reflejan la experiencia previa particular del usuario con ese candidato (sino el de toda la comunidad), nuestro enfoque penaliza su contribución mediante el factor f_{IR} —que tomará valores menores a 0.8—.

  La expresión que formaliza el cálculo de la idoneidad del candidato CN_i en este escenario se formaliza en la Ec. 5.6, donde $IR_{CN_i}(C_p)$ es la medida de reputación de dicho candidato en productos de la clase C_p.

$$S_U^{C_p}(CN_i) = \varphi \cdot Sim_{U,CN_i}(C_p) + (1-\varphi) \cdot (2 \cdot IR_{CN_i}(C_p) - 1) \cdot f_{IR} \tag{5.6}$$

En consecuencia, serán seleccionados como vecinos de U (en recomendaciones de productos clasificados bajo la clase C_p) los candidatos que obtienen los valores más altos en el par de expresiones Ec. 5.5-Ec. 5.6. De este modo, el vecindario estará formado por usuarios cuyas preferencias (en relación a productos de la clase del producto objetivo) son las más similares a las del usuario activo y además tienen el historial más exitoso de contribuciones en recomendaciones de productos de ese tipo.

5.6 Ejemplo de aplicación

A fin de ilustrar la estrategia propuesta con un sencillo ejemplo, consideremos el extracto de la red de confianza implícita condicionada a la clase *Rock* —perteneciente a la jerarquía *Música*— representado en la Fig. 5.4. En este escenario, se muestran resaltados el usuario activo de la recomendación en curso, U, y cuatro candidatos a formar parte de su vecindario —CN_1, CN_2, CN_3 y CN_4—, preseleccionados mediante el proceso descrito en la Sección 4.3.3 en función de la similitud entre sus preferencias y las de U respecto a dicha clase.

[1] En el Capítulo 7 se definirán dos parámetros para determinar de forma significativa el peso de sendas componentes: la *completitud del perfil* de un usuario y la *madurez* de las relaciones de confianza implícita, respectivamente.

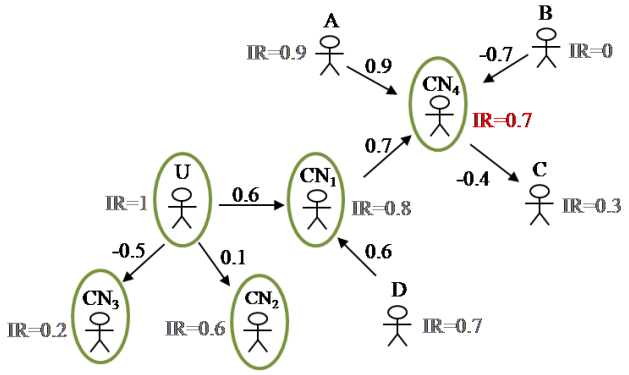


Figura 5.4: Extracto de la red de confianza implícita condicionada a *Rock*

En primer lugar, nos centraremos en el proceso de inferencia de las relaciones de confianza implícita condicionadas a la clase *Rock*. Más concretamente, con el objetivo de mostrar los aspectos más significativos de la estrategia con suficiente claridad, estableceremos una comparación entre las relaciones inferidas para CN_1 y CN_2 —i.e., aquellos candidatos que obtienen valores positivos de confianza por parte de U—. A continuación, desarrollaremos el proceso de cuantificación de la medida de reputación de CN_4 y concluiremos la sección analizando de forma cualitativa las implicaciones de nuestra selección de vecindario *trust-aware* en el escenario de la Fig. 5.4.

5.6.1 Relaciones de confianza condicionadas a la clase *Rock*

Tal como acabamos de comentar, nos centraremos en dos de las relaciones de confianza implícita inferidas para U: la establecida entre éste y el usuario CN_1, y la que lo vincula con CN_2 (ver Fig. 5.4).

Recordemos que el punto de partida de nuestra estrategia son las medidas de éxito o fracaso obtenidas en base a los resultados de las recomendaciones previas en las que han participado los dos usuarios considerados —uno en su rol de usuario activo y el otro como vecino—. Por tanto, el primer paso para inferir la relación de confianza condicionada a la clase *Rock* entre U y CN_1 es identificar en el perfil de U las recomendaciones de productos, clasificados bajo la clase *Rock* en la ontología del sistema, en las que CN_1 ha participado como vecino. Dicho de otro modo, de entre los productos pertenecientes a la clase *Rock* registrados en el perfil de U, se tendrán en cuenta aquellos cuyo *user-ID* se corresponda con el usuario CN_1 (ver Sección 3.3.5). De igual modo, se identifican las recomendaciones de productos de esa clase en las que CN_2 ha participado como vecino de U.

Supongamos que en la Fig. 5.5 están representadas las instancias de los productos objetivo de las recomendaciones previas hechas a U en las que ha participado como vecino CN_1 (a la izquierda de la figura) y aquellas en las que CN_2 ha formado parte de su vecindario (a la derecha). Dichas instancias —i_1, i_2, i_3, i_4, i_5 en el caso de CN_1, y j_1, j_2 en el caso de CN_2— se muestran clasificadas bajo sus respectivas clases hoja, en el subárbol delimitado por la clase *Rock*. Asimismo, la Fig. 5.5 incluye junto a cada

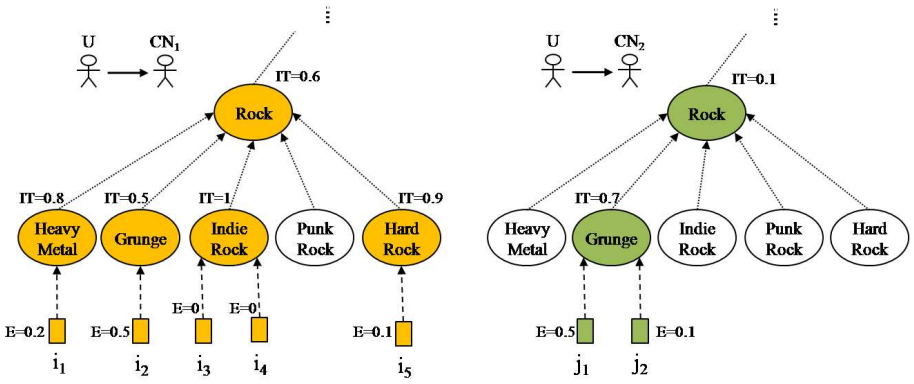


Figura 5.5: Contribuciones de CN_1 y CN_2 en recomendaciones a U —clase *Rock*—

instancia las medidas de éxito o fracaso obtenidas como resultado de la Ec. 5.1 para el usuario en cuestión y, junto a cada clase, el valor de confianza implícita correspondiente (algunos de cuyos cálculos se muestran a continuación).

Tal como detallamos en la Sección 5.2.4, la estrategia propuesta cuantifica el valor de una relación de confianza implícita entre dos usuarios respecto a una clase hoja mediante la Ec. 5.2, en función de las medidas de éxito o fracaso obtenidas en recomendaciones previas de productos de dicha clase. De modo que, por ejemplo, el valor concerniente a la clase *Heavy Metal* para CN_1 y el de la clase *Grunge* para CN_2 se obtienen como sigue:

$$IT_{U,CN_1}(Heavy\ Metal) = 1 - E_{CN_1} i_1, U = 1 - 0{,}2 = 0{,}8$$

$$IT_{U,CN_2}(Grunge) = \frac{1}{2}(1 - E_{CN_2} j_1, U + 1 - E_{CN_2} j_2, U) = \frac{1}{2}(1 - 0{,}5 + 1 - 0{,}1) = 0{,}7$$

En cuanto a la propagación de los valores de las clases hoja a la clase del nivel de abstracción considerado en este ejemplo —i.e., la clase *Rock*—, se describe a continuación cómo se realiza (a partir de la Ec. 5.3) en cada uno de los escenarios representados en la Fig. 5.5:

$$IT_{U,CN_1}(Rock) = \frac{IT_{U,CN_1}(Grunge)}{1 + \#sib(Grunge)} = \frac{0{,}7}{5} = 0{,}14$$

$$IT_{U,CN_2}(Rock) = \frac{IT_{U,CN_2}(Heavy\ Metal)}{1 + \#sib(Heavy\ Metal)} + \frac{IT_{U,CN_2}(Grunge)}{1 + \#sib(Grunge)} + \frac{IT_{U,CN_2}(Indie\ Rock)}{1 + \#sib(Indie\ Rock)}$$

$$+ \frac{IT_{U,CN_2}(Hard\ Rock)}{1 + \#sib(Hard\ Rock)} = \frac{0{,}8 + 0{,}5 + 1 + 0{,}9}{5} = 0{,}64$$

Nótese que, pese a tratarse de dos usuarios con un historial de éxitos en sus contribuciones como vecinos de U, el valor de confianza respecto a la clase *Rock* inferido para CN_1 es mucho más significativo que el inferido para CN_2. Esta diferencia se debe a que hay una proporción muy superior de subclases que propagan un valor de confianza no nulo en el caso de CN_1 (ver Fig. 5.5). Tal como comentamos en la Sección 5.2.4, es razonable dar una mayor consideración a la relación de confianza condicionada a la clase *Rock* inferida entre U y CN_1 dado que su valor proviene de un historial de éxitos en recomendaciones de discos de diversas subcategorías de música rock (en concreto, cuatro de las cinco definidas en la ontología), mientras que la inferida entre U y CN_2 únicamente proviene de recomendaciones de discos de la subcategoría *Grunge*.

5.6.2 Reputación de CN_4

Puesto que uno de los candidatos a vecino preseleccionados, CN_4, no ha participado previamente como vecino en recomendaciones a U de productos de la clase *Rock*, se tendrá en cuenta en el proceso de selección su reputación respecto a dicha clase.

Para ilustrar el cálculo de la medida de reputación de CN_4, condicionada a la clase *Rock*, consideraremos el extracto de la red de confianza representada en la Fig. 5.4, que incluye junto a cada usuario su medida de reputación, obtenida mediante la Ec. 5.4. Tal como describimos en la Sección 5.4, la estrategia propuesta reúne las valoraciones de confianza implícita que el usuario en cuestión recibe de otros, teniendo en cuenta la reputación de estos últimos —medidas que dependerán a su vez de las relaciones de confianza que éstos reciben de otros, atendiendo a la reputación de aquellos, y así sucesivamente—.

Por tanto, a la vista de la Fig. 5.4, se obtiene la reputación de CN_4 aplicando la Ec. 5.4 como sigue:

$$IR_{CN_4}(Rock) = 0{,}5(1 + \frac{1}{3}(IT_{CN_1,CN_4}(Rock) \cdot IR_{CN_1}(Rock) + IT_{A,CN_4}(Rock) \cdot IR_A(Rock) +$$

$$IT_{B,CN_4}(Rock) \cdot IR_B(Rock))) = 0{,}5(1 + \frac{1}{3}(0{,}9 \cdot 0{,}9 + 0{,}7 \cdot 0{,}8 - 0{,}7 \cdot 0))$$

$$= 0{,}73$$

Nótese que, a pesar de que CN_4 recibe una valoración de -0.7 por parte del usuario B, ésta no influye negativamente en la medida de reputación de CN_4. Esto es debido a que B ha obtenido la peor medida de reputación como resultado de sus contribuciones como vecino a otros usuarios (no representados en la Fig. 5.4) y, en consecuencia, su valoración a CN_4 se anula en la Ec. 5.4.

5.6.3 Selección de vecindario *trust-aware*

En base a lo comentado en la Sección 5.5, es posible extraer varias conclusiones en el presente ejemplo:

- Si las medidas de similitud obtenidas entre U y los candidatos CN_1 y CN_2 son semejantes, CN_1 obtendrá un mayor valor de idoneidad (ver Ec. 5.5) y, por tanto, podría ser seleccionado, frente a CN_2, para formar parte del vecindario de U.

- De los candidatos considerados, CN_3 obtendrá el menor valor de idoneidad, salvo que el valor de similitud inferido entre U y CN_3 sea muy elevado y éste tenga un mayor peso en el proceso de selección —i.e., $\varphi > 0{,}5$ en la Ec. 5.5—. Esta situación sería conveniente, por ejemplo, en caso de que el historial de recomendaciones previas entre los usuarios sea escaso, por lo que las relaciones de confianza implícita inferidas puedan no ser lo suficientemente maduras.

- La estrategia propuesta prima a aquellos candidatos con los que U tiene relaciones de confianza implícita positivas sobre los que tienen buenos valores de reputación, dado que aquellas reflejan la experiencia previa particular del usuario activo. Por tanto, la idoneidad de CN_4 es obtenida mediante la Ec. 5.6 y, previsiblemente, será menor que la de CN_1.

5.7 Beneficios de la estrategia

La estrategia presentada en este capítulo posibilita la incorporación de modelos de confianza en el proceso de selección de vecindarios de forma transparente a los usuarios, obteniendo relaciones de confianza a partir de los resultados obtenidos en recomendaciones previas. Más aun, nuestro enfoque emplea técnicas de razonamiento semántico sobre dichos resultados (explorando la ontología de productos comerciales), lo que permite cuantificar relaciones de confianza implícita condicionadas a clases de productos. Precisamente, la posibilidad de considerar diferentes niveles de abstracción, a la hora de inferir las relaciones de confianza que conformarán las redes implícitas, provee de una gran flexibilidad al proceso.

Las relaciones obtenidas entre pares de usuarios son, por tanto, estructuradas en redes de confianza implícita entre todos los usuarios del sistema. Frente a los modelos de confianza revisados en la literatura (que disponen de una única red, con independencia del producto objetivo considerado en la recomendación), nuestro enfoque construye diferentes redes, cada una de ellas formada con relaciones de confianza respecto a una determinada clase de productos. La flexibilidad que aporta la selección del nivel de abstracción en la ontología permite establecer una relación de compromiso entre precisión y *sparsity*, tal como comentamos en la Sección 5.3, pudiendo construir un tipo u otro de redes en función de las necesidades del sistema.

Por otro lado, a través de la exploración de las redes resultantes, nuestro enfoque cuantifica medidas de reputación de cada usuario. Esto permite disponer de información adicional sobre la fiabilidad de los candidatos a vecinos (en particular, lo fiables que son considerados por la comunidad en conjunto) en el proceso de selección de vecindarios.

Cabe destacar que esta estrategia está estrechamente vinculada a la presentada en el capítulo anterior, por lo que los beneficios adquiridos con aquella permanecen. En efecto, tanto las medidas de confianza implícita como los valores de reputación están condicionados a la misma clase que considere aquella estrategia para evitar el problema de falsos vecinos.

Capítulo 6

Estrategia para la cuantificación de la experiencia práctica de los usuarios

Este capítulo describe una nueva métrica de experiencia en el campo del comercio electrónico, obtenida en base a los productos que los usuarios han probado. La estrategia desarrollada no requiere que los usuarios proporcionen ningún tipo de información adicional, sino que explota los historiales de consumo registrados en sus perfiles —información disponible en cualquier sistema recomendador—. Para estimar la experiencia práctica de los usuarios en productos de un cierto tipo, la estrategia propuesta aplica técnicas de razonamiento semántico sobre la información de consumo de los mismos, inspeccionando las descripciones semánticas de los productos comerciales de ese tipo. Las medidas resultantes son incorporadas en la etapa de predicción del proceso de recomendación colaborativo, ponderando las contribuciones de los usuarios que forman el vecindario pertinente en función de su experiencia acerca del producto objetivo, con el fin de mejorar los resultados de las recomendaciones.

6.1 Introducción

La *experiencia*, entendida como "la aptitud o conocimiento que una persona tiene en un cierto campo", es un concepto que puede resultar de gran utilidad en un amplio rango de aplicaciones, desde la búsqueda del empleado con mayor experiencia para resolver un cierto problema en una empresa [90], a la incorporación de medidas de experiencia en un sistema recomendador colaborativo. En este último contexto, que es el que nos ocupa, se trata de valorar la experiencia de los usuarios que forman parte del vecindario de un cierto usuario activo, a fin de mejorar la calidad de las recomendaciones ofrecidas.

Bajo el razonable supuesto de que los individuos valoran en mayor medida las opiniones de aquellos con mayor experiencia, a la hora de pedir consejo sobre la elección de un producto, encontramos en la literatura recientes publicaciones de trabajos que siguen esta línea de investigación, tal como comentamos en la Sección 2.4.3. La idea es, por tanto, valorar la experiencia de los vecinos respecto al producto objetivo de la

recomendación, teniendo en mayor consideración las opiniones de los usuarios más expertos en esa clase de productos, en el cálculo de predicción del interés del usuario activo (ver Sección 2.3.3).

Sin embargo, el mayor problema de este tipo de enfoques es la falta de conocimiento acerca de las aptitudes de los usuarios en un sistema recomendador, puesto que la información disponible suele estar limitada al historial de consumo registrado en sus perfiles. Es por este motivo que la mayoría de dichos trabajos dependen de información explícita adicional para estimar la experiencia de los usuarios. Por ejemplo, el trabajo presentado en [22] depende de las calificaciones explícitas y subjetivas que dan los usuarios del sitio web Epinions (ver Sección 2.4.2) a las reseñas sobre productos escritas por otros usuarios, para poder inferir la experiencia de estos últimos. Por su parte, los autores de [143] basan su métrica de experiencia en la calidad de los ítems valorados por los usuarios en sus perfiles, considerando a un usuario más experto cuantos más ítems de calidad conoce. La mayor limitación de este enfoque radica en cómo determinar la mencionada calidad. En cuanto a la incorporación de conocimiento del dominio de aplicación en la cuantificación de la experiencia de los usuarios, destacamos el trabajo de Dima *et al.* [32]. Si bien está limitado a ser utilizado en un recomendador de películas, su enfoque define la experiencia de un usuario en base al conocimiento que éste tenga acerca de ciertos géneros, actores o directores.

La métrica de experiencia que proponemos en este capítulo, publicada en [85, 87], es aplicable a todo el dominio del comercio electrónico y no requiere más información que el historial de consumo del usuario en cuestión —disponible en cualquier sistema recomendador—. Se trata de una medida de la experiencia *práctica* de los usuarios, obtenida por medio de técnicas de razonamiento semántico sobre los productos registrados en sus perfiles. Mediante el análisis de los atributos semánticos asociados a los productos comerciales que un usuario ha comprado (y, presumiblemente, probado), nuestra estrategia infiere la experiencia práctica de dicho usuario en un tipo de productos dado. En concreto, la experiencia de un usuario en una clase de productos dependerá de la *variabilidad* de los productos específicos almacenados en su perfil clasificados bajo esa clase en la ontología. De este modo, por ejemplo, cuanto mayor sea la diversidad de vinos que un usuario ha probado (de diferentes marcas, regiones, variedades de uva, calidades, etc.) mayor será la experiencia práctica que dicho usuario ha podido ganar sobre el vino.

El objetivo último de la estrategia propuesta es incorporar la información acerca de la experiencia práctica de los usuarios en la fase de predicción del proceso de recomendación de nuestro sistema colaborativo. Recordemos que en un sistema recomendador colaborativo, una vez seleccionado el vecindario para el usuario activo en cuestión, se infiere el nivel de interés potencial de dicho usuario en el producto objetivo de la recomendación en función de las valoraciones registradas en los perfiles de sus vecinos sobre ese producto (tal como se describe en la Sección 2.3.3). Nuestro enfoque enriquece esta etapa del proceso de personalización, condicionando el peso de las contribuciones de cada uno de los vecinos (seleccionados mediante la estrategia detallada en la Sección 5.5) a la experiencia práctica que hayan podido adquirir en el tipo de productos considerado en la recomendación.

6.2 Experiencia práctica

Tal como acabamos de adelantar, el objetivo perseguido es disponer de información acerca de la experiencia de los usuarios respecto al tipo de productos considerado en cada recomendación. Con ese propósito en mente, la métrica desarrollada cuantifica la experiencia práctica de los usuarios sobre una determinada clase de la ontología de productos del sistema.

En particular, nuestra estrategia se limita a inferir medidas de experiencia sobre *clases hoja* de la ontología (ver Sección 3.2). El motivo de esta restricción se debe a la consideración de que las clases hoja son las que categorizan de forma más precisa las instancias de los productos que los usuarios han probado. Nótese que, a medida que se asciende por la jerarquía y las clases representan conceptos más generales, es probable que se incluyan subcategorías de productos no incluidos en el perfil del usuario considerado. Por tanto, para determinar la experiencia que proviene del aprendizaje práctico de los usuarios de forma significativa, las clases hoja son precisamente las que recogen la información buscada.

Más concretamente, para definir la experiencia de un usuario en una cierta clase hoja de productos, nuestro enfoque se basa en las siguientes premisas:

- Por un lado, cuanta mayor variedad de productos clasificados bajo esa clase haya valorado el usuario, más experto será considerado en dicha clase de productos. Dicho de otro modo, cuanto mayor sea el número de productos diferentes (pero instancias de la misma clase hoja en la ontología) contenidos en el perfil de un usuario, mayor conocimiento práctico puede haber adquirido sobre ese tipo de productos. Por ejemplo, si la clase considerada es *Vino*, cuanto mayor número de marcas de vino distintas haya probado, mayor experiencia puede haber ganado en cuanto a vino.

- Por otra parte, la experiencia práctica de un usuario en una determinada clase de productos dependerá de la variedad de atributos de los productos registrados en su perfil que pertenezcan a esa clase. Cuanto más diversos sean los productos que el usuario ha probado, en lo que respecta a sus características semánticas, mayor será su experiencia práctica en dicha clase de productos. Siguiendo con el ejemplo anterior, cuantos más tipos diferentes de vino (procedentes de distintos lugares de origen, producidos con diversas variedades de uva, de diferentes cualidades, etc.) ha probado el usuario, mejor.

Conforme a estas premisas, la expresión matemática de la medida de experiencia práctica propuesta se compone de dos factores, denominados *variabilidad cuantitativa* y *variabilidad cualitativa*, respectivamente. Antes de centrarnos en la descripción de cada una de estas componentes, detalladas a continuación, la Ec. 6.1 muestra la expresión resultante, donde $QnVar_U(C_L)$ es la variabilidad cuantitativa del usuario U respecto a la clase hoja C_L, $QlVar_U(C_L)$ su variabilidad cualitativa en dicha clase y γ un parámetro de combinación configurable que modula el peso de cada componente.

$$Exp_U(C_L) = \gamma \cdot QnVar_U(C_L) + (1-\gamma) \cdot QlVar_U(C_L) \tag{6.1}$$

6.2.1 Variabilidad cuantitativa

La variabilidad cuantitativa de un usuario respecto a una cierta clase hoja dependerá del número de instancias diferentes contenidas en su perfil que estén clasificadas bajo esa clase en la ontología. No obstante, dado que nuestro enfoque pretende no estar ligado a una ontología específica —y teniendo además en cuenta que las ontologías son estructuras dinámicas que evolucionan y son pobladas a lo largo del tiempo—, la variabilidad cuantitativa se define en base a dos factores, tal como muestra la Ec. 6.2. El primer factor, $V_U(C_L)$, denominado *variabilidad*, depende del usuario considerado. El segundo, $f_p(C_L)$, denominado *factor de instanciación*, depende de la naturaleza de los productos y de cómo estos están formalizados en la ontología.

$$QnVar_U(C_L) = V_U(C_L) \cdot f_p(C_L) \tag{6.2}$$

Variabilidad

Esta componente cuantifica la proporción de instancias diferentes de C_L registradas en el perfil del usuario, respecto al total de instancias definidas bajo dicha clase en la ontología. Por ejemplo, si las instancias se corresponden con marcas comerciales y la clase considerada es *Queso*, el cálculo de variabilidad considera cuántas marcas de queso distintas ha probado el usuario. A modo de ejemplo, consideremos la columna izquierda de la Tabla 6.1, donde se muestran las instancias pertenecientes a la clase *Queso* almacenadas en los perfiles de dos usuarios: N_1 y N_2 —en la parte superior e inferior de dicha columna, respectivamente—. Es inmediato deducir que la variabilidad de N_1 será mayor que la de N_2, puesto que N_1 ha probado más marcas de queso (instancias de la clase *Queso* en la ontología de productos) que N_2.

De acuerdo a esta idea, nuestra estrategia obtiene la variabilidad de un usuario U respecto a una clase hoja C_L por medio de la Ec. 6.3, donde $\#I_U(C_L)$ es el número de instancias diferentes de la clase C_L registradas en el perfil del usuario, y $\#I(C_L)$ el número total de instancias clasificadas en la ontología bajo esa clase.

$$V_U(C_L) = \frac{\#I_U(C_L)}{\#I(C_L)} \tag{6.3}$$

Factor de instanciación

Tal como adelantamos previamente, el propósito del factor de instanciación es evitar que la estimación de la experiencia de un usuario en una clase se vea distorsionada por la ontología que está siendo utilizada por el sistema. La idea es penalizar a las clases que sean menos representativas de la variedad existente en el mercado, es decir, aquellas de las que podemos extraer poca información acerca de la experiencia práctica de los usuarios en ese tipo de productos. En concreto, nuestra estrategia considera que una clase no es lo suficientemente representativa del mercado cuando, habiendo en él un gran número de productos clasificables bajo dicha clase, ha sido instanciada con muy pocos productos. Por otro lado, es necesario tener en cuenta aquellos productos cuya oferta en el mercado es tal que, llegados a un punto, el número de instancias de la clase correspondiente en la ontología no debe influir en los resultados.

A tal fin, la Ec. 6.4 define el factor de instanciación de una clase C_L como el número

INSTANCIAS (*Queso*)	ATRIBUTOS				
	Región	*Tipo de Leche*	*Curación*	*Textura*	*Presencia de Hongo*
N_1					
$Mozzarella_A$	*Roma (Lacio)*	*M-B (Búfalo)*	F_1 *(Fresco)*	*S-*s_1 *(Semisuave)*	–
$Manchego_C$	*Cuenca (La Mancha)*	*M-O (Oveja)*	C_3 *(Curado)*	*C-*c_7 *(Compacto)*	–
$Manchego_B$	*Toledo (La Mancha)*	*M-O (Oveja)*	C_3 *(Curado)*	*C-*c_7 *(Compacto)*	–
$Roquefort_B$	*R-s-S (Aveyron)*	*Lac-O (Oveja)*	*S-*c_6 *(Semicurado)*	*S-*s_5 *(Semisuave)*	*P-R (Azul)*
$Gruyere_A$	*Berna (Berna)*	*G-V (Vaca)*	C_2 *(Curado)*	*C-*c_1 *(Compacto)*	–
$Camembert_A$	*Orne (Normandía)*	*C-V (Vaca)*	*S-*c_3 *(Semicurado)*	*C-*r_1 *(Cremoso)*	*C-M (Corteza)*
N_2					
$Gruyere_A$	*Berna (Berna)*	*G-V (Vaca)*	C_2 *(Curado)*	*C-*c_1 *(Compacto)*	–
$Parmesano_A$	*Parma (Parma)*	*P-V (Vaca)*	C_5 *(Curado)*	G_3 *(Granuloso)*	–

Tabla 6.1: Productos de la clase *Queso* en los perfiles de dos usuarios

de instancias definidas bajo esa clase en la ontología, $\#I(C_L)$, respecto al número de instancias que marca el umbral (a partir del cual es irrelevante que dicha clase esté más o menos instanciada) para los productos disponibles en el mercado que pueden ser categorizados bajo C_L, $\#I_M(C_L)$.

$$f_p(C_L) = \begin{cases} 1 & \text{si } \#I(C_L) \geq \#I_M(C_L) \\ \frac{\#I(C_L)}{\#I_M(C_L)} & \text{si } \#I(C_L) < \#I_M(C_L) \end{cases} \tag{6.4}$$

Ejemplo: Para ilustrar mejor la necesidad de este factor, consideremos dos ontologías de productos comerciales O y O', pudiendo ser éstas dos ontologías diferentes desde un principio o la misma ontología que ha evolucionado a lo largo del tiempo. Dado un producto que es ofrecido en el mercado por 200 marcas comerciales distintas, cuyo umbral $\#I_M(C_L)$ se ha fijado a 100, supongamos que su correspondiente clase, C_L, está instanciada en O solamente con 10 instancias, mientras que en O' hay clasificadas 100. Si nuestra estrategia únicamente considera la proporción de instancias que contiene el perfil del usuario en cuestión (i.e., su variabilidad), asumirá erróneamente el mismo nivel de experiencia: (i) cuando el usuario ha probado 5 productos (teniendo a su disposición la ontología O), y (ii) cuando el usuario ha probado 50 (teniendo a su disposición la ontología O').

6.2.2 Variabilidad cualitativa

Cada instancia de productos comerciales formalizada en la ontología está vinculada a través de propiedades convenientemente etiquetadas a sus correspondientes atributos semánticos, que están a su vez clasificados bajo diversas jerarquías de atributos. Por ejemplo, el escenario representado en la Fig. 6.1 muestra los atributos relacionados con las instancias $Mozzarella_A$ y $Roquefort_B$ de la clase *Queso*, categorizados bajo sus pertinentes jerarquías de atributos: *Roma* y *R-s-S* bajo la jerarquía *Región*, los atributos *M-B* y *Lac-O* bajo *Tipo de Leche*, los atributos F_1 y $S\text{-}c_6$ bajo *Curación*, los atributos $S\text{-}s_1$ y $S\text{-}s_5$ bajo *Textura* y, únicamente en el caso de *Roquefort*, el atributo *P-R* bajo la jerarquía *Presencia de Hongo*.

En base a esta información, se define la variabilidad cualitativa de un usuario en una cierta clase de productos atendiendo a la disparidad de atributos concernientes a los productos —instancias de esa clase— almacenados en su perfil. Cuanto más diversas sean las descripciones semánticas de tales productos, mayor será la variabilidad cualitativa de ese usuario en la clase considerada. Por ejemplo, si la clase de productos examinada es *Queso*, cuanto mayor rango de *Regiones*, *Tipos de Leche*, *Curación*, *Texturas*, etc., haya probado el usuario, mejor.

Para reflejar esta idea, nuestra estrategia explora la ontología a fin de obtener la *disimilitud semántica entre instancias* de los productos del perfil del usuario que pertenecen a la clase en cuestión. Dado que la disimilitud semántica entre dos instancias dependerá a su vez de la *disimilitud semántica entre atributos* inferida entre las correspondientes descripciones semánticas asociados a esos productos, describimos previamente este parámetro. Reuniendo ambos conceptos, definiremos finalmente la expresión global de la variabilidad cualitativa.

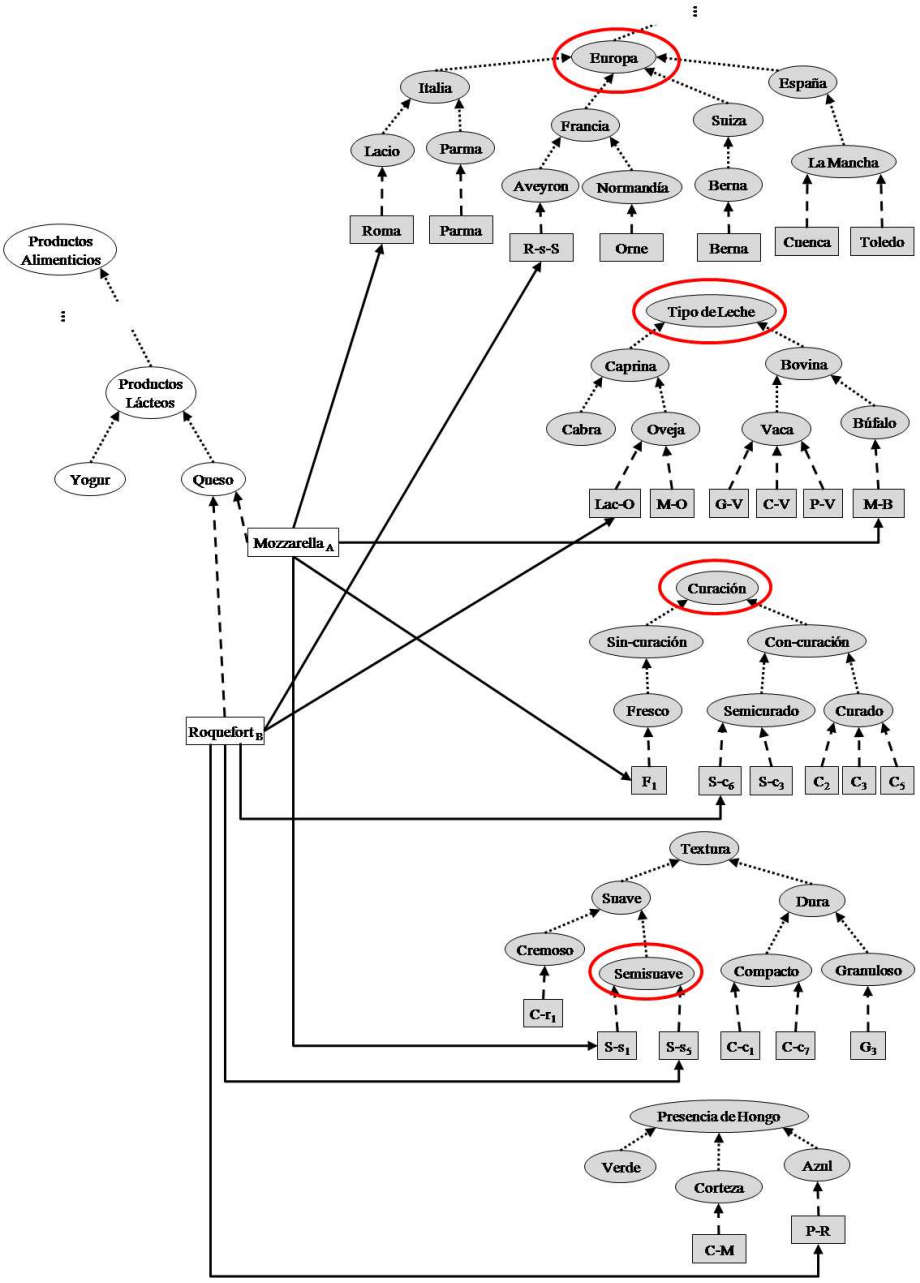


Figura 6.1: Disimilitud semántica entre dos instancias

Disimilitud semántica entre atributos

Con el propósito de cuantificar la disimilitud semántica entre dos atributos pertenecientes a la misma jerarquía de atributos, nuestro enfoque estudia la posición que ocupan en dicha jerarquía sus respectivas clases. En particular, la estrategia propuesta define la disimilitud semántica entre dos atributos a_x y a_y como el opuesto de la similitud semántica jerárquica (descrita en la Sección 4.2.1) entre ellos, aplicada aquí a atributos en lugar de a instancias de clases. Tal como muestra la Ec. 6.5, únicamente los atributos clasificados bajo la misma jerarquía de atributos ($hie(a_x) = hie(a_y)$) son comparables.

$$SemDis(a_x, a_y) = \begin{cases} 1 - SemSim_{Hie}(a_x, a_y) & \text{si } hie(a_x) = hie(a_y) \\ \nexists & \text{en otro caso} \end{cases} \tag{6.5}$$

donde:

$$SemSim_{Hie}(a_x, a_y) = \frac{depth(LCA_{a_x,a_y})}{max(depth(a_x), depth(a_y))}$$

Esta expresión, por tanto, recurre de modo opuesto a los conceptos de *profundidad* y *antecesor común más cercano (LCA)*, presentados en la Sección 3.2. En consecuencia, cuanto más lejano esté el antecesor común LCA_{a_x,a_y} de las clases a las que pertenecen los atributos considerados y cuanto menor sea la profundidad de dicho *LCA* (i.e., cuanto más próximo esté a la clase raíz de la jerarquía), mayor será su disimilitud semántica. De manera que, si dos atributos pertenecen a la misma clase, su disimilitud semántica se anula.

En el ejemplo mostrado en la Fig. 6.1 pueden verse, rodeados, los *LCAs* identificados entre los atributos relacionados a las instancias $Mozzarella_A$ y $Roquefort_B$. A la vista de la figura, es sencillo inferir que la disimilitud semántica entre los atributos *Lac-O* y *M-B* (pertenecientes a la jerarquía *Tipo de Leche*) es mayor que la existente entre los atributos $S\text{-}s_1$ y $S\text{-}s_5$ (clasificados bajo la jerarquía *Textura*).

Disimilitud semántica entre instancias

A partir de la disimilitud semántica entre atributos, se construye la expresión de la disimilitud semántica entre instancias. En concreto, para obtener la disimilitud semántica entre dos instancias i_x e i_y, nuestra estrategia cuantifica la disimilitud semántica existente entre cada par de atributos (uno vinculado a i_x y el otro a i_y) que pertenecen a la misma jerarquía, tal como indica la Ec. 6.6.

$$SemDis(i_x, i_y) = \frac{1}{\mid A_{xy} \mid} \sum_{j=1}^{A(i_x)} \sum_{k=1}^{A(i_y)} SemDis(a_j(i_x), a_k(i_y)) \tag{6.6}$$

En esta expresión, $A(i_x)$ es el número de atributos vinculados a la instancia i_x, $A(i_y)$ el número de atributos relacionados con la instancia i_y, y A_{xy} el conjunto de pares de atributos de i_x e i_y comparables, esto es, aquellos que pertenecen a la misma jerarquía de atributos, tal como muestra la Ec. 6.7.

$$A_{xy} = \left\{ (a_j(i_x), a_k(i_y)) \mid hie(a_j(i_x)) = hie(a_k(i_y)) \right\} \tag{6.7}$$

En el escenario representado en la Fig. 6.1, los atributos comparables entre las instancias $Mozzarella_A$ y $Roquefort_B$ son los clasificados bajo las jerarquías *Región*, *Tipo de Leche*, *Curación* y *Textura*. Por el contrario, el atributo *P-R* de $Roquefort_B$ no será considerado en el cálculo de la disimilitud semántica entre $Mozzarella_A$ y $Roquefort_B$, ya que $Mozzarella_A$ no ha sido caracterizada con ningún atributo perteneciente a la jerarquía *Presencia de Hongo*.

Expresión matemática de la variabilidad cualitativa

Reuniendo las ideas y los parámetros presentados hasta este punto, nuestro enfoque define la variabilidad cualitativa de un usuario en una clase hoja C_L como el promedio de las disimilitudes semánticas existentes entre las instancias registradas en su perfil que pertenezcan a dicha clase ($\#I_U(C_L)$). A tal fin, mediante la Ec. 6.8, se obtienen los valores de disimilitud semántica de todas las posibles combinaciones distintas, dos a dos, entre instancias[1].

$$\begin{aligned} QlVar_U(C_L) &= \frac{1}{\binom{\#I_U(C_L)}{2}} \sum_{f=1}^{\#I_U(C_L)-1} \sum_{g=f+1}^{\#I_U(C_L)} SemDis(i_f, i_g) \\ &= \frac{2! \cdot (\#I_U(C_L)-2)!}{\#I_U(C_L)!} \sum_{f=1}^{\#I_U(C_L)-1} \sum_{g=f+1}^{\#I_U(C_L)} SemDis(i_f, i_g) \end{aligned} \tag{6.8}$$

A modo de ejemplo, consideremos de nuevo la Tabla 6.1, que recoge las instancias pertenecientes a la clase *Queso* en los perfiles de los usuarios N_1 y N_2. En la columna derecha de la tabla se muestran los atributos vinculados a cada instancia (organizados bajo sus correspondientes jerarquías) y, entre paréntesis, las clases a las que pertenecen. Mediante la exploración del fragmento de la ontología mostrado en la Fig. 6.1 es posible inferir que N_1 es superior a N_2 en relación a su variabilidad cualitativa. En efecto, el perfil de N_1 contiene quesos que conducen a valores altos de disimilitud semántica, cuyas clases de atributos varían desde *Oveja* a *Búfalo* o *Vaca* (en la jerarquía *Tipo de Leche*), o desde *Fresco* a *Semicurado* o *Curado* (en la jerarquía *Curación*). En cambio, N_2 únicamente ha probado quesos *Curados* hechos con leche de *Vaca* que tienen una textura *Dura* (*LCA* entre sus atributos de las clases *Compacto* y *Granuloso*).

[1] Teniendo en cuenta que $SemDis(i_x, i_y) = SemDis(i_y, i_x)$.

6.3 Incorporación de medidas de experiencia en el proceso de recomendación

Tal como adelantamos al comienzo del capítulo, el objetivo perseguido es incorporar las medidas de experiencia práctica de los usuarios en la fase de predicción del proceso de recomendación colaborativo. Es decir, una vez se ha seleccionado el vecindario apropiado para el usuario activo (de acuerdo a la estrategia *trust-aware* descrita en la Sección 5.5 del capítulo previo), nuestro enfoque tiene en cuenta la experiencia práctica de cada vecino, en el producto objetivo de la recomendación, a la hora de estimar el nivel de interés potencial del usuario activo en dicho producto.

Dado que la expresión matemática tomada como punto de partida difiere de la utilizada tradicionalmente, descrita en la Sección 2.3.3, detallamos en primer lugar cómo nuestro enfoque permite enriquecer aquélla. Para concluir, describiremos el criterio final de recomendación seguido.

6.3.1 Métrica de predicción tradicional

La expresión tradicional utilizada para el cuantificación del nivel de interés potencial de un cierto usuario activo U en relación al producto objetivo de la recomendación, a, se muestra en la Ec. 6.9, donde M es el tamaño del vecindario seleccionado para U, N_k es el k-ésimo vecino de U, $R_{N_k}(a)$ es la valoración dada por el vecino N_k en relación al producto a, y $sim(U,N_k)$ es la similitud entre las preferencias de los usuarios U y N_k —obtenida mediante la *correlación Pearson-r* o la *similitud del coseno*, tal como describimos en la Sección 4.3.1—.

$$Pred(a,U) = \frac{1}{M}\sum_{k=1}^{M} R_{N_k}(a) \cdot sim_{U,N_k} \tag{6.9}$$

La metodología presentada hasta este punto en la tesis permite enriquecer la expresión descrita en la Ec. 6.9 en dos aspectos:

- Por un lado, gracias a la técnica de modelado de usuarios utilizada, nuestra métrica de predicción no requerirá necesariamente valoraciones explícitas dadas por los usuarios. En efecto, los índices DOI obtenidos mediante el método detallado en la Sección 3.3.4 recogen no sólo los niveles de interés declarados expresamente por los usuarios, también los inferidos a partir de la realimentación de relevancia implícita adquirida por el sistema.

- Por otra parte, la métrica de similitud desarrollada dentro de nuestra estrategia para combatir el problema de falsos vecindarios, presentada en la Sección 4.3, permite disponer de medidas de similitud entre usuarios más fiables, relativas al tipo de producto considerado en la recomendación. En efecto, recordemos que las métricas de similitud entre usuarios tradicionales consideran todas las preferencias incluidas en los perfiles de los usuarios al cuantificar su afinidad, derivando en resultados contraproducentes cuando existe una gran diversidad de productos disponibles en el sistema recomendador.

6.3.2 Métrica de predicción propuesta

Tomando como punto de partida la expresión de predicción tradicional enriquecida con las propuestas que acabamos de comentar, nuestra estrategia incorpora las medidas de experiencia práctica de los usuarios que forman parte del vecindario.

En particular, se ponderará la contribución de cada vecino dependiendo de su nivel de experiencia en la clase hoja del producto considerado en la recomendación. Volviendo al ejemplo de la Tabla 6.1, si N_1 y N_2 forman parte del vecindario de un cierto usuario activo, y el producto que está siendo considerado para una posible recomendación pertenece a la clase *Queso*, las opiniones de N_1 tendrán más peso que las de N_2 en la predicción —puesto que el nivel de experiencia de N_1 en *Queso* es mayor, de acuerdo a lo expuesto en la Sección 6.2—.

Para ser más específicos, el enfoque propuesto cuantifica el nivel de interés potencial de un usuario activo U respecto al producto objetivo a mediante la Ec. 6.10, donde N_k es el k-ésimo vecino de U, $DOI_{N_k}(a)$ el nivel de interés de N_k en a, $sim_{U,N_k}(C_p(a))$ la similitud entre las preferencias de U y N_k en productos del tipo de a —en el nivel de abstracción seleccionado (ver Sección 4.3.3)—, M el tamaño del vecindario considerado, y $Exp_{N_k}(C_L(a))$ el nivel de experiencia de N_k en la clase hoja a la que pertenece el producto a.

$$Pred(a,U) = \frac{1}{M}\sum_{k=1}^{M} DOI_{N_k}(a) \cdot sim_{U,N_k}(C_p(a)) \cdot Exp_{N_k}(C_L(a)) \tag{6.10}$$

Como consecuencia, cuanto más positivas sean las opiniones de los vecinos de U acerca del producto objetivo a (teniendo en cuenta la experiencia práctica de cada uno de ellos en ese tipo de productos) y cuanto más similares sean las preferencias de estos usuarios respecto a las de U, más apropiado será considerado a para ser recomendado a dicho usuario activo.

6.3.3 Criterio de recomendación: rango del umbral $\beta^{C_p}_{Pred_U}$

En concreto, una vez cuantificado $Pred(a,U)$ mediante la Ec. 6.10, nuestra estrategia sugiere el producto a al usuario activo U si dicho valor de predicción supera un umbral $\beta^{C_p}_{Pred_U}$ previamente establecido.

Nótese que, para cada usuario, se define un umbral respecto a cada una de las clases C_p del nivel de abstracción considerado en el proceso de recomendación. El motivo de esta distinción guarda relación con la estrategia de selección de vecindarios empleada (ver Sección 5.5). En efecto, dado que el resultado de la predicción está estrechamente vinculado con el vecindario seleccionado, y nuestro enfoque dispone de tantos vecindarios para el usuario activo como clases C_p, se asigna un umbral diferente para cada vecindario.

Este umbral $\beta^{C_p}_{Pred_U}$ no permanece invariable a lo largo del tiempo, sino que oscila en el rango $[\omega_3, \omega_2]$, siendo ω_3 y ω_2 valores elevados (en torno a 0.8 y 0.9, respectivamente). Nuestro enfoque modifica el valor correspondiente a un cierto usuario a medida que éste interacciona con el sistema. En particular, el parámetro que modula este proceso es la precisión que el usuario ha obtenido en las últimas sugerencias ofrecidas por el recomendador, en productos del tipo C_p. Así, el umbral de recomendación

es decrementado (hasta alcanzar el mínimo ω_3) cuando se advierte un incremento en la precisión del sistema y, por el contrario, es aumentado (hasta ω_2) cuando la precisión de las sugerencias de productos C_p se reduce. En caso de que la precisión de la herramienta permanezca invariable entre dos recomendaciones consecutivas, el umbral también conserva su valor.

Recordemos que la precisión de las recomendaciones a un usuario se calcula como la proporción de productos sugeridos por la herramienta y aceptados[2] por éste. Por tanto, la precisión de las recomendaciones en relación a una clase C_p puede definirse como el número de productos de la clase C_p aceptados por el usuario, respecto al número total de productos de esa clase sugeridos por la herramienta, tal como muestra la siguiente expresión:

$$P_{C_p} = \frac{\#Productos\ aceptados_{C_p}}{\#Productos\ recomendados_{C_p}} \tag{6.11}$$

Sean $t+1$ el instante temporal en el que se presenta una sugerencia al usuario activo U de un producto clasificado bajo la clase C_p en la ontología, y t el correspondiente a la recomendación anterior de otro producto de esa clase. Sean $P_{C_p}(t)$ y $P_{C_p}(t+1)$ los valores de precisión correspondientes a ambas recomendaciones, y $\beta^{C_p}_{Pred_U}(t)$ y $\beta^{C_p}_{Pred_U}(t+1)$ los valores de los umbrales utilizados en éstas. En este escenario, el valor del umbral de recomendación se actualiza en $t+1$ de acuerdo a la expresión mostrada en la Ec. 6.12.

$$\beta^{C_p}_{Pred_U}(t+1) = \begin{cases} \omega_3 & \text{si } \beta^{C_p}_{Pred_U}(t) - \Delta^{C_p}_P(t+1,t) < \omega_3 \\ \omega_2 & \text{si } \beta^{C_p}_{Pred_U}(t) - \Delta^{C_p}_P(t+1,t) > \omega_2 \\ \beta^{C_p}_{Pred_U}(t) - \Delta^{C_p}_P(t+1,t) & \text{en otro caso} \end{cases} \tag{6.12}$$

donde $\Delta^{C_p}_P(t+1,t) = P_{C_p}(t+1) - P_{C_p}(t)$.

6.4 Ejemplo de aplicación

Con el objetivo de ilustrar la estrategia propuesta con mayor detalle, en esta sección describiremos cuantitativamente el ejemplo desarrollado a lo largo del capítulo. De manera análoga, calcularemos en primer lugar la *variabilidad cuantitativa* de los usuarios N_1 y N_2 respecto a la clase *Queso* (ver Tabla 6.1) y, a continuación, nos centraremos en la obtención de sus medidas de *variabilidad cualitativa* en relación a dicha clase. Para concluir, compararemos las medidas de experiencia práctica que se obtienen para estos usuarios y destacaremos la importancia de ambas métricas de variabilidad en nuestro enfoque, planteando escenarios de aplicación alternativos.

[2]Nuestro enfoque considera que una recomendación es aceptada por el usuario si el índice DOI calculado para el producto sugerido —a partir de la realimentación de relevancia explícita o implícita adquirida por el sistema— es positivo (ver Sección 3.3.4).

6.4.1 Variabilidad cuantitativa de N_1 y N_2

Tal como describimos en la Sección 6.2.1, la variabilidad cuantitativa de un usuario depende de la *variabilidad* de instancias de productos registrados en el perfil de dicho usuario, pertenecientes a la clase hoja en cuestión, y del *factor de instanciación* de la ontología del sistema (ver Ec. 6.2).

En este ejemplo, por simplicidad, asumiremos que la clase *Queso* está suficientemente instanciada en la ontología ($f_p(Queso) = 1$), siendo el número de instancias definidas en la misma 50 —i.e., se han clasificado 50 marcas diferentes de queso—. Por tanto, atendiendo a la columna izquierda de la Tabla 6.1, la variabilidad cuantitativa de N_1 y N_2 se obtiene como sigue:

$$QnVar_{N_1} = \frac{\#I_{N_1}(Queso)}{\#I(Queso)} \cdot f_p(Queso) = \frac{6}{50} \cdot 1 = 0{,}12$$

$$QnVar_{N_2} = \frac{\#I_{N_2}(Queso)}{\#I(Queso)} \cdot f_p(Queso) = \frac{2}{50} \cdot 1 = 0{,}02$$

En efecto, dado que N_1 ha probado más marcas de queso que N_2, la variabilidad cuantitativa de N_1 respecto a la clase *Queso* es mayor que la de N_2. Nótese, sin embargo, que N_1 no obtiene un valor elevado, puesto que se trata de un producto con una oferta en el mercado considerable y dicho usuario tiene registrado en su perfil solamente una pequeña proporción.

6.4.2 Variabilidad cualitativa de N_1 y N_2

A fin de obtener la variabilidad cualitativa de N_1 y N_2 hemos de desarrollar la Ec. 6.8 atendiendo a la parte superior e inferior de la Tabla 6.1, respectivamente.

Variabilidad cualitativa de N_1

Tal como detallamos en la Sección 6.2.2, con miras a inferir la variabilidad cualitativa de un usuario en una clase hoja, la estrategia propuesta explora la ontología para calcular la disimilitud semántica de todas las combinaciones distintas de pares de instancias registradas en el perfil del usuario, pertenecientes a dicha clase.

En particular, para cuantificar la variabilidad cualitativa de N_1 en la clase *Queso*, debemos obtener la disimilitud semántica entre cada par de instancias representadas en la Tabla 6.2. Recordemos que, como $SemDis(i_x, i_y) = SemDis(i_y, i_x)$, el número de combinaciones distintas se corresponde con el coeficiente binomial de $\#I_U(C_L)$ —número de instancias de la clase considerada registradas en el perfil del usuario— en dos. Por tanto, en este caso:

$$\binom{\#I_{N_1}(Queso)}{2} = \binom{6}{2} = \frac{6!}{2! \cdot (6-2)!} = 15$$

i_x	i_y	$SemDis(i_x, i_y)$
$Mozzarella_A$	$Manchego_C$	0.92
$Mozzarella_A$	$Manchego_B$	0.92
$Mozzarella_A$	$Roquefort_B$	0.67
$Mozzarella_A$	$Gruyere_A$	0.54
$Mozzarella_A$	$Camembert_A$	0.67
$Manchego_C$	$Manchego_B$	0
$Manchego_C$	$Roquefort_B$	0.54
$Manchego_C$	$Gruyere_A$	0.42
$Manchego_C$	$Camembert_A$	0.79
$Manchego_B$	$Roquefort_B$	0,54
$Manchego_B$	$Gruyere_A$	0.42
$Manchego_B$	$Camembert_A$	0.79
$Roquefort_B$	$Gruyere_A$	0.79
$Roquefort_B$	$Camembert_A$	0.57
$Gruyere_A$	$Camembert_A$	0.54

Tabla 6.2: Disimilitud semántica entre instancias del perfil de N_1

La disimilitud semántica entre dos instancias se cuantifica mediante la Ec. 6.6, a partir de las medidas disimilitud semántica entre cada par de atributos comparables de dichas instancias —i.e., aquellos que pertenecen a la misma jerarquía de atributos—.

Por ejemplo, los atributos comparables entre las instancias $Mozzarella_A$ y $Roquefort_B$ (representados en la Fig. 6.1) son: *Roma* y *R-s-S* (pertenecientes a la jerarquía *Región*), *Lac-O* y *M-B* (de la jerarquía *Tipo de Leche*), F_1 y $S\text{-}c_6$ (clasificados bajo la jerarquía *Curación*), y $S\text{-}s_1$ y $S\text{-}s_5$ (de la jerarquía *Textura*).

Tal como expusimos en la Sección 6.2.2, nuestro enfoque cuantifica la disimilitud semántica entre cada par de atributos mediante la Ec. 6.5, atendiendo a la profundidad de los mismos y de su correspondiente *LCA*. En el escenario representado en la Fig. 6.1 pueden verse, rodeados, los *LCAs* identificados entre los pares de atributos en cuestión.

Por tanto, en este caso, el desarrollo de la Ec. 6.6 resulta como sigue:

$$\begin{aligned} SemDis(Mozzarella_A, Roquefort_B) = & \frac{1}{4}(SemDis(Roma, R\text{-}s\text{-}S) + SemDis(Lac\text{-}O, M\text{-}B) \\ & + SemDis(F_1, S\text{-}c_6) + SemDis(S\text{-}s_1, S\text{-}s_5)) \\ & = 0{,}67 \end{aligned}$$

siendo:

$$SemDis(Roma, R\text{-}s\text{-}S) = 1 - \frac{depth(LCA_{Roma,R\text{-}s\text{-}S})}{max(depth(Roma), depth(R\text{-}s\text{-}S))} = 1 - \frac{1}{3} = 0{,}67$$

$$SemDis(Lac\text{-}O, M\text{-}B) = 1 - \frac{depth(LCA_{Lac\text{-}O,M\text{-}B})}{max(depth(Lac\text{-}O), depth(M\text{-}B))} = 1 - 0 = 1$$

$$SemDis(F_1, S\text{-}c_6) = 1 - \frac{depth(LCA_{F_1,S\text{-}c_6})}{max(depth(F_1), depth(S\text{-}c_6))} = 1 - 0 = 1$$

$$SemDis(S\text{-}s_1, S\text{-}s_5) = 1 - \frac{depth(LCA_{S\text{-}s_1,S\text{-}s_5})}{max(depth(S\text{-}s_1), depth(textitS - s_5))} = 1 - \frac{2}{2} = 0$$

De manera análoga se obtienen el resto de valores de disimilitud semántica entre instancias recogidos en la Tabla 6.2. Reuniendo las contribuciones de tales valores de disimilitud en la Ec. 6.8 resulta la medida de variabilidad cualitativa de N_1 respecto a la clase *Queso*, tal como se muestra a continuación:

$$\begin{aligned}
QlVar_{N_1}(Queso) = & \frac{1}{15}(SemDis(Mozzarella_A, Manchego_C) + SemDis(Mozzarella_A, \\
& Manchego_B) + SemDis(Mozzarella_A, Roquefort_B) + \\
& SemDis(Mozzarella_A, Gruyere_A) + SemDis(Mozzarella_A, \\
& Camembert_A) + SemDis(Manchego_C, Manchego_B) + \\
& SemDis(Manchego_C, Roquefort_B) + SemDis(Manchego_C, \\
& Gruyere_A) + SemDis(Manchego_C, Camembert_A) + \\
& SemDis(Manchego_B, Roquefort_B) + SemDis(Manchego_B, \\
& Gruyere_A) + SemDis(Manchego_B, Camembert_A) + \\
& SemDis(Roquefort_B, Gruyere_A) + SemDis(Roquefort_B, \\
& Camembert_A) + SemDis(Gruyere_A, Camembert_A)) \\
& = \frac{1}{15}(0{,}92 + 0{,}92 + 0{,}67 + 0{,}54 + 0{,}67 + 0 + 0{,}54 + 0{,}42 + \\
& 0{,}79 + 0{,}54 + 0{,}42 + 0{,}79 + 0{,}79 + 0{,}57 + 0{,}54) = 0{,}61
\end{aligned}$$

Variabilidad cualitativa de N_2

En el caso de N_2, dado que únicamente tiene registradas en su perfil dos instancias de la clase *Queso*, correspondientes a los productos $Gruyere_A$ y $Parmesano_A$ (ver Tabla 6.1), la variabilidad cualitativa de este usuario equivale a la medida de disimilitud semántica inferida entre ellas.

En la Fig. 6.2 se muestran los atributos de dichas instancias —clasificados bajo sus pertinentes jerarquías— y, rodeados, los *LCAs* de cada par de atributos comparables.

Por consiguiente, el desarrollo de la Ec. 6.8 para N_2 respecto a la clase *Queso* resulta como sigue:

$$QlVar_{N_2}(Queso) = SemDis(Gruyere_A, Parmesano_A) = \frac{1}{4}(SemDis(Berna, Parma) + SemDis(G\text{-}V, P\text{-}V) + SemDis(C_2, C_5) + SemDis(C\text{-}c_1, G_3)) = 0{,}29$$

siendo:

$$SemDis(Berna, Parma) = 1 - \frac{depth(LCA_{Berna,Parma})}{max(depth(Berna), depth(Parma))} = 1 - \frac{1}{3} = 0{,}67$$

$$SemDis(G\text{-}V, P\text{-}V) = 1 - \frac{depth(LCA_{G\text{-}V,P\text{-}V})}{max(depth(G\text{-}V), depth(P\text{-}V))} = 1 - \frac{2}{2} = 0$$

$$SemDis(C_2, C_5) = 1 - \frac{depth(LCA_{C_2,C_5})}{max(depth(C_2), depth(C_5))} = 1 - \frac{2}{2} = 0$$

$$SemDis(C\text{-}c_1, G_3) = 1 - \frac{depth(LCA_{C\text{-}c_1,G_3})}{max(depth(C\text{-}c_1), depth(G_3))} = 1 - \frac{1}{2} = 0{,}5$$

En efecto, tal como comentamos en la Sección 6.2.2, la variabilidad cualitativa de N_2 respecto a la clase *Queso* es mucho menor que la de N_1, puesto que las características semánticas de los quesos registrados en su perfil son muy similares —ambos son quesos *Curados* hechos con leche de *Vaca* que tienen una textura *Dura* (ver Fig. 6.2)—.

6.4.3 Experiencia práctica

Es inmediato deducir en el presente ejemplo que la experiencia práctica inferida para N_1 respecto a la clase *Queso* será mayor que la de N_2, independientemente del peso dado a cada componente en la Ec. 6.1, ya que N_1 obtiene un mayor valor tanto en variabilidad cuantitativa como en variabilidad cualitativa.

Sin embargo, conviene destacar que la variaribilidad cuantitativa y la variabilidad cualitativa no son medidas necesariamente proporcionales —de ahí la necesidad de definir ambas métricas—.

- Supongamos, por ejemplo, que N_2 tuviera registrados en su perfil los quesos $Mozzarella_A$ y $Roquefort_B$ en lugar de $Gruyere_A$ y $Parmesano_A$. Tal como describimos previamente, los primeros obtienen un valor de disimilitud semántica mucho mayor (en concreto $SemDis(Mozzarella_A, Roquefort_B) = 0{,}67$) y, en consecuencia, la variabilidad cualitativa de N_2 resultaría mayor que la de N_1. En ese escenario, no es razonable asumir que N_2 tenga una mayor experiencia práctica

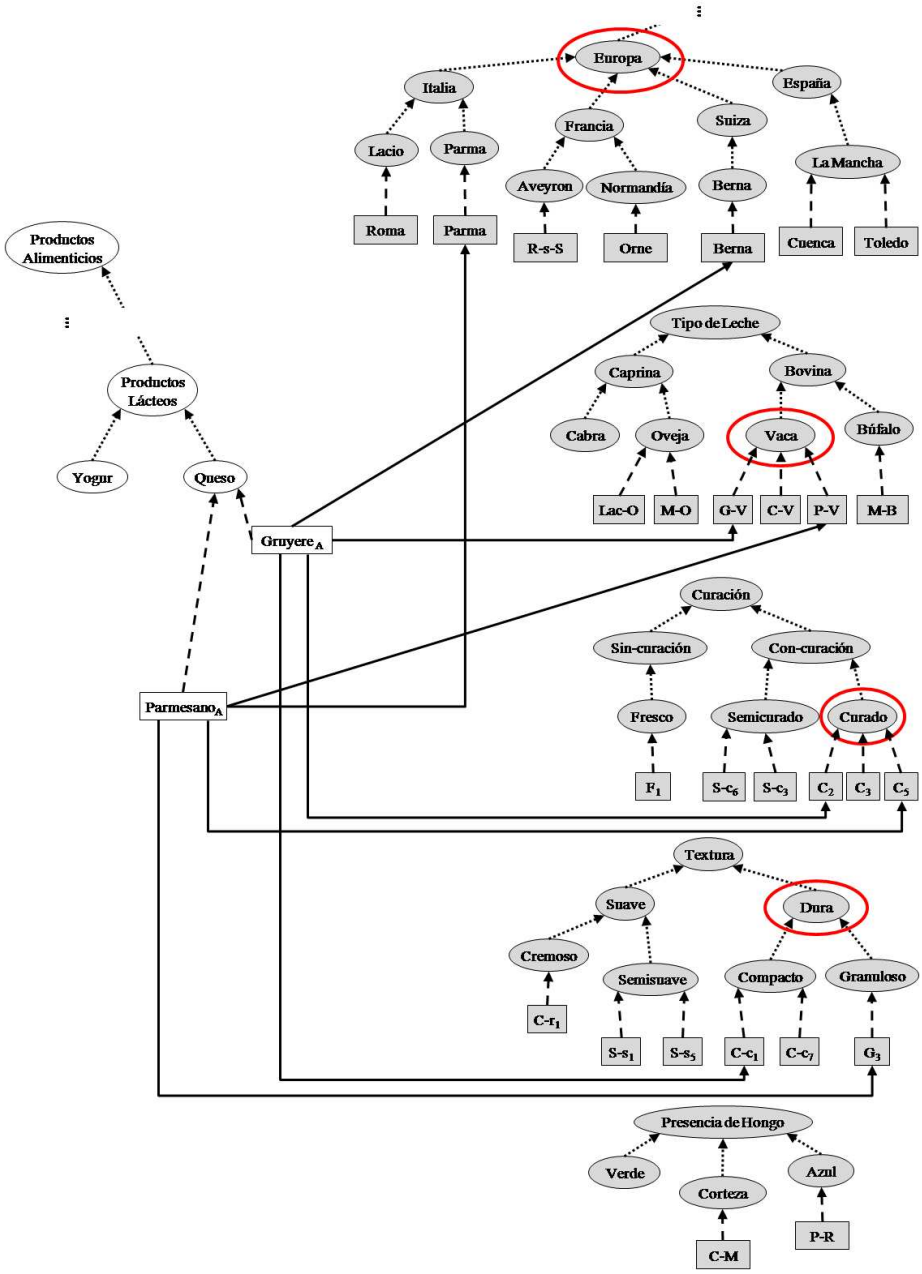


Figura 6.2: Disimilitud semántica entre $Gruyere_A$ y $Parmesano_A$

que N_1, puesto que este último no sólo obtiene un valor de variabilidad cualitativa significativo, además el número de instancias de la clase *Queso* registradas en su perfil es mucho mayor. Precisamente la variabilidad cuantitativa permite identificar esta situación.

- Igualmente, la variabilidad cuantitativa por sí misma no es suficiente para inferir la experiencia práctica de un usuario. Consideremos, por ejemplo, dos usuarios N_3 y N_4 que tienen registrados en sus perfiles un número similar de instancias diferentes de la clase *Queso*. No obstante, N_3 únicamente ha probado quesos hechos con leche de *Vaca*, de curación *Semicurado* y con textura *Semisuave*, mientras que el perfil de N_4 contiene quesos de características muy diferentes. En este escenario, no es razonable concluir que ambos tienen una experiencia práctica similar en quesos, puesto que la adquirida por N_3 será presumiblemente más limitada —ha probado varias marcas diferentes, pero de un mismo tipo de quesos—. La variabilidad cualitativa posibilita distinguir una mayor experiencia de N_4 en ese caso.

6.5 Beneficios de la estrategia

El beneficio más destacable de la estrategia presentada en este capítulo es su carácter implícito. Gracias al enfoque semántico utilizado, se infieren medidas de experiencia práctica de los usuarios partiendo de información disponible en cualquier sistema recomendador —el historial de consumo registrado en sus perfiles—. En efecto, se estima el conocimiento práctico de los usuarios sobre productos comerciales sin requerir que éstos proporcionen ningún tipo de información adicional.

Aplicando técnicas de razonamiento semántico sobre la información de consumo de un usuario, nuestra estrategia cuantifica su experiencia en base a los productos que éste ha probado hasta el momento. Más concretamente, dada una clase de productos, la métrica propuesta tiene en cuenta la variedad de productos de esa clase que el usuario conoce. Esta premisa es razonable, puesto que cuantos más productos diferentes de un mismo tipo un individuo ha probado, más conocimiento práctico puede haber adquirido sobre ese tipo de productos.

Precisamente la exploración de la ontología de productos permite incorporar una gran cantidad de conocimiento útil al proceso. En particular, posibilita identificar dos tipos de variabilidad de productos: la variabilidad cuantitativa y la variabilidad cualitativa. En ambos casos, la estrategia propuesta examina las instancias de productos específicos registrados en el perfil del usuario que pertenecen a una cierta clase hoja de la ontología. De este modo, se obtienen medidas de experiencia precisas, respecto a las clases más específicas que contienen los productos que el usuario ha probado.

Por un lado, mediante la variabilidad cualitativa nuestra estrategia cuantifica la cantidad de productos concretos diferentes de una cierta clase hoja almacenados en el perfil del usuario. Esto es, si las instancias representan marcas comerciales, esta medida depende del número de marcas distintas del producto en cuestión.

Por otro lado, la variabilidad cualitativa permite cuantificar la diversidad de atributos asociados a los productos específicos examinados. A tal fin, nuestro enfoque explora las jerarquías de atributos formalizadas en la ontología para obtener medidas de disimilitud entre las características semánticas de dichos productos.

Capítulo 7

Hibridación del proceso de selección de vecindarios con una componente basada en redes sociales

El propósito de este capítulo es doble: por un lado, en él se describe una estrategia para la cuantificación de relaciones de confianza personal a partir de la información disponible en redes sociales, cuyo objetivo es combatir el problema de cold start users presente en los sistemas recomendadores, aún abierto en el estado del arte; por otra parte, se detalla en él la expresión que combina las relaciones obtenidas mediante esta estrategia con las obtenidas por medio de los métodos descritos en capítulos previos, de forma que el proceso de selección de vecindarios se adapte a la evolución de cada usuario activo en el sistema.

7.1 Introducción

Las estrategias para la mejora del proceso de selección de vecindarios presentadas hasta ahora (detalladas en los Capítulos 4 y 5) tienen un fundamento común: su enfoque semántico. En ambos casos, la explotación de técnicas de razonamiento sobre la ontología de productos comerciales posibilita el descubrimiento de nuevo conocimiento, a partir del ya adquirido por el sistema. En la primera estrategia, se hace uso de la semántica a fin de incorporar en el proceso información adicional sobre las preferencias de los usuarios y así mitigar ciertas limitaciones identificadas en él —el problema de falsos vecindarios y el *sparsity problem*—. En la segunda, se descubren relaciones de confianza significativas entre los usuarios explotando la semántica asociada a los resultados de recomendaciones anteriores, registrados en sus perfiles. Esto es, son estrategias que enriquecen el proceso de selección de vecindarios de forma transparente a los usuarios, sin requerir que éstos proporcionen ningún tipo de información adicional.

Sin embargo, el carácter implícito de estas estrategias hace que se vean afectadas por el conocido problema de *cold start users* (descrito en la Sección 2.4) que ha afectado a los sistemas recomendadores desde sus orígenes, y que continúa abierto en el

estado del arte. Ciertamente, cuando el perfil de un cierto usuario activo no está lo suficientemente detallado, no es posible averiguar sus preferencias y, por tanto, la posibilidad de descubrir nuevo conocimiento sobre ellas se ve reducida. En consecuencia, los perfiles de usuario incompletos limitan en gran medida los beneficios obtenidos por nuestras estrategias.

Para combatir este problema, es necesario incluir en el proceso de selección de vecindarios información externa al sistema recomendador. Tal como mencionamos en la Sección 2.4, encontramos en la literatura numerosos trabajos que emplean estrategias *trust-aware* a tal fin, seleccionando como vecinos a aquellos individuos en los que el usuario activo confía de forma directa o indirecta. Sin embargo, los enfoques existentes dependen de redes de confianza *explícita* y, por tanto, requieren que los usuarios especifiquen expresamente cuánto confían en otros individuos. En consecuencia, dichos enfoques están limitados a aquellos casos en los que los usuarios están dispuestos a proporcionar tal información.

En este capítulo proponemos una estrategia que aprovecha la masiva adopción de tecnologías Web 2.0 en la actualidad (ver Sección 2.5.2) para obtener relaciones de confianza *personal* entre usuarios, a partir de la información disponible en las redes sociales —sin requerir a los usuarios que proporcionen ningún tipo de información de forma expresa—. Tomando como base las relaciones explícitas existentes en una red social (recordemos que cada conexión representa el acuerdo mutuo de los usuarios vinculados a convertirse en *contactos*), nuestra estrategia cuantifica sus correspondientes valores de confianza mediante el análisis de la información contenida en los perfiles sociales a los que tiene acceso.

Teniendo en cuenta que muchos de los usuarios que forman parte del sistema recomendador pueden a su vez estar registrados en una o más redes sociales, nuestro enfoque incorpora sus pertinentes relaciones de confianza personal en el proceso de selección de vecindarios. El objetivo perseguido es combinar estas relaciones de confianza personal con las de similitud semántica y confianza implícita (inferidas según los métodos descritos en la Sección 4.3 y la Sección 5.2, respectivamente), de modo que se tomen en mayor consideración en el proceso de selección de vecindarios, bien las relaciones extraídas de las redes sociales, o bien las obtenidas dentro del propio sistema recomendador, dependiendo de la evolución en el sistema del usuario activo considerado.

Precisamente, cuando el usuario activo en cuestión acabe de darse de alta en el sistema recomendador y, presumiblemente, su perfil no esté lo suficientemente detallado, nuestro enfoque aprovechará la información externa al sistema que tenga a su disposición —la que pueda extraerse de las redes sociales en las que esté registrado dicho usuario—. A medida que el usuario interaccione con el sistema, su perfil se completará cada vez más y nuestro enfoque podrá aprovechar los beneficios de las estrategias semánticas desarrolladas para incorporar nuevo conocimiento y enriquecer el proceso de selección de vecindarios.

Con el propósito de determinar la evolución de un usuario en el sistema, se definen dos factores (que serán descritos en la Sección 7.3): la *completitud del perfil* de un usuario y la *madurez* de las relaciones de confianza implícita. El primero hace referencia al nivel de detalle del perfil de un usuario en un momento dado, atendiendo a la cantidad y *heterogeneidad* de la información registrada en el mismo. En cuanto al factor de madurez, su propósito es reflejar la fiabilidad de las relaciones de confianza implícita inferidas hasta el momento.

7.2 Cuantificación de relaciones de confianza personal entre usuarios

Tal como adelantamos en la sección anterior, a fin de mitigar el problema de *cold start users* de forma transparente a los usuarios, proponemos una estrategia para inferir relaciones de confianza personal entre los usuarios de nuestro sistema que también están registrados en una o más redes sociales y sus respectivos contactos en dichas redes. Para ello, nuestra estrategia examina la información a la que tiene acceso en la red social para cuantificar la intensidad de las relaciones que esos usuarios tienen con otros individuos a los que, por norma general, conocen personalmente.

Más concretamente, la métrica propuesta cuantifica la relación de confianza personal (*PT —Personal Trust—*) de un usuario A a un usuario B atendiendo al historial de interacciones entre ellos y a las configuraciones de privacidad que han especificado en sus respectivos perfiles sociales. Con esta idea en mente, se incluyen dos componentes en dicha métrica: el *factor de interacción*, $f_{INTERACTION}(A,B)$, y el *factor de privacidad*, $f_{PRIVACY}(A,B)$, ponderados mediante el parámetro de combinación σ, tal como se muestra en la Ec. 7.1.

$$PT_{A,B} = \sigma \cdot f_{INTERACTION}(A,B) + (1-\sigma) \cdot f_{PRIVACY}(A,B) \tag{7.1}$$

Cabe destacar que, si bien nuestra estrategia parte de relaciones *no-direccionales* (recordemos que cada enlace en la red social representa el acuerdo mutuo de los usuarios conectados a convertirse en contactos), obtiene relaciones *direccionales*. Por tanto, el método propuesto está en concordancia con la propiedad de *asimetría* inherente a la confianza —la confianza de A a B puede ser diferente de la de B a A (ver Sección 2.4.2)—. A modo de ilustración, la Fig. 7.1 escenifica cómo un vínculo entre dos usuarios en la red social (i.e., no direccional) se transforma en dos arcos direccionales, uno de A a B, y otro de B a A. Nuestro enfoque conduce a relaciones direccionales gracias a las desigualdades en las interacciones entre los usuarios (por ejemplo, uno está más pendiente del otro) y en sus configuraciones de privacidad (por ejemplo, uno permite al otro el acceso total a su perfil, mientras que éste le restringe el acceso a algunas de sus publicaciones al primero).

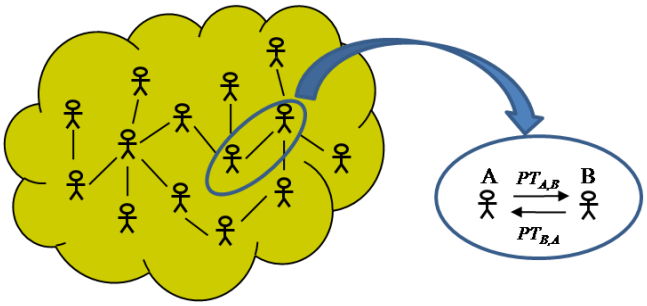


Figura 7.1: Relaciones de confianza personal

7.2.1 Factor de interacción

El factor de interacción de A a B dependerá del número de acciones de A en la red social que tengan que ver con B. Aunque los parámetros específicos incluidos en esta medida varían ligeramente según la red social que esté siendo analizada, generalmente abarcan los considerados en la siguiente expresión, donde w_i ($i \in [1,7]$) son valores configurables que modulan el peso de cada componente, de forma que $\sum_i w_i = 1$.

$$\begin{aligned} f_{INTERACTION}(A,B) = {} & w_1 \cdot \frac{X_{Posts}(A,B)}{X_{Posts}(A)} + w_2 \cdot \frac{X_{C_{Photos}}(A,B)}{X_{C_{Photos}}(A)} \\ & + w_3 \cdot \frac{X_{T_{Photos}}(A,B)}{X_{T_{Photos}}(A)} + w_4 \cdot \frac{X_{Wall}(A,B)}{X_{Wall}(A)} + w_5 \cdot \frac{X_{Groups}(A,B)}{X_{Groups}(A)} \\ & + w_6 \cdot \frac{X_{Mentions}(A,B)}{X_{Mentions}(A)} + w_7 \cdot \frac{X_{Likes}(A,B)}{X_{Likes}(A)} \end{aligned} \tag{7.2}$$

En definitiva, se cuantifica:

- La proporción de mensajes que A ha enviado a B, $X_{Posts}(A,B)$, respecto al número total de mensajes enviados por A a cualquiera de sus contactos, $X_{Posts}(A)$.
- La proporción de fotografías publicadas por B que han sido comentadas por A, $X_{C_{Photos}}(A,B)$, respecto al número total de fotografías comentadas por A, $X_{C_{Photos}}(A)$.
- La proporción de fotografías publicadas por A en las que B está etiquetado, $X_{T_{Photos}}(A,B)$, respecto al número total de fotografías publicadas por A, $X_{T_{Photos}}(A)$.
- La proporción de publicaciones de A en el perfil de B, $X_{Wall}(A,B)$, respecto al total de publicaciones de A en los perfiles de cualquiera de sus contactos, $X_{Wall}(A)$.
- La proporción de grupos a los que pertenecen A y B, $X_{Groups}(A,B)$, respecto al número total de grupos a los que pertenece A, $X_{Groups}(A)$.
- La proporción de veces que A ha etiquetado a B en sus publicaciones, $X_{Mentions}(A,B)$, respecto al número total de veces que A ha etiquetado a alguien en sus publicaciones, $X_{Mentions}(A)$.
- La proporción de veces que A ha expresado "me gusta" en publicaciones hechas por B, $X_{Likes}(A,B)$, respecto al número total de veces que A ha expresado "me gusta", $X_{Likes}(A)$.

Por tanto, cuanto mayor sea el número de acciones de A en la red social que involucren a B, mayor será el factor de interacción de A a B. Nótese que, al considerar valores proporcionales, la métrica tiene en cuenta el comportamiento habitual de los usuarios en la red social —las acciones de un usuario con un gran nivel de actividad en la red ofrecen menos información que aquellas de un usuario que es muy selectivo a la hora de interactuar—.

7.2.2 Factor de privacidad

Para definir el factor de privacidad de A a B, nuestra estrategia analiza los grupos de contactos que A ha especificado en su configuración de privacidad (a fin de restringir el acceso a cierta información de su perfil) y las implicaciones que conlleva para B la mencionada configuración. El factor tomará un valor positivo si B está incluido en un grupo con acceso a información que está restringida para otros. En ese caso, cuanto menor sea el número de contactos de A incluidos en dicho grupo, mayor será el valor resultante. Por el contrario, el factor tomará un valor negativo cuando el grupo en cuestión es utilizado por A para evitar que los contactos que lo forman accedan a cierta información, abierta para otros.

Dicho de otro modo, el factor de privacidad de A a B depende del *tamaño* del grupo en el que B está incluido (G_{Size}) y del *acceso* que tiene ese grupo a los contenidos[1] publicados por A en su perfil (G_{Access}), de forma que:

- Se anula cuando, independientemente de su tamaño, el grupo no está siendo usado por A para restringir el acceso a contenidos —i.e., la información mostrada a ese grupo es la misma que la abierta al resto de contactos de A—.
- Toma un valor positivo máximo cuando el grupo solamente está formado por un usuario, que es B, y todas las publicaciones de A son mostradas en exclusiva a dicho grupo, esto es, a B.
- Toma un valor negativo máximo cuando el grupo está formado únicamente por B, y toda la información contenida en el perfil de A es accesible a todos los contactos de A, salvo B.

En particular, se define el parámetro G_{Size} mediante la Ec. 7.3, donde $\#SC_{A_G}$ es el número de contactos de A incluidos en el grupo en cuestión, y $\#SC_A$ el número total de contactos de A en la red social.

$$G_{Size} = \frac{\#SC_{A_G} - 1}{\#SC_A} \tag{7.3}$$

Con respecto a G_{Access}, nuestra estrategia atiende al número de contenidos a los que el grupo considerado tiene acceso, $\#P_{A_G}$, respecto al número total de contenidos en el perfil de A, $\#P_A$, teniendo además en cuenta el número de contenidos que A muestra a todos sus contactos, $\#P_{A_{Open}}$, tal como se describe en la Ec. 7.4.

$$G_{Access} = \frac{\#P_{A_G} - \#P_{A_{Open}}}{\#P_A} \tag{7.4}$$

En consecuencia, y teniendo en cuenta que B puede estar incluido en más de un grupo de los definidos por A en su configuración de privacidad, se obtiene su correspondiente factor de privacidad mediante la Ec. 7.5.

[1] Entendemos por *contenido* cualquier publicación del usuario en su perfil social (por ejemplo, fotos, actualizaciones de estado, información personal, comentarios, etc.).

$$f_{PRIVACY}(A,B) = \frac{1}{N_B}\sum_{i=1}^{N_B}(1 - G_{Size_i}) \cdot G_{Access_i} \tag{7.5}$$

En esta expresión, N_B es el número de grupos definidos por A que incluyen a B, G_{Size_i} el tamaño asociado al grupo i-ésimo al que pertenece B, y G_{Access_i} el acceso inferido para dicho grupo.

7.3 Selección de vecindarios híbrida

Tal como apuntábamos al comienzo del capítulo, las relaciones de confianza personal extraídas de las redes sociales (por medio de la estrategia descrita en la sección previa) son especialmente beneficiosas para mejorar las recomendaciones iniciales a usuarios que son nuevos en el sistema, esto es, para aliviar el problema de *cold start users* inherente a todo sistema recomendador. En efecto, cuando las preferencias e historial de consumo registrados en el perfil de un cierto usuario activo son escasos, formar su correspondiente vecindario con individuos de su círculo social puede mejorar en gran medida los resultados de las recomendaciones. Por un lado, se seleccionan como vecinos individuos en los que el usuario confía de forma subjetiva. Por otra parte, es habitual que esos individuos compartan intereses con el usuario activo en cuestión y, por tanto, sus perfiles sociales pueden contener información útil para el proceso.

Con el paso del tiempo, a medida que los nuevos usuarios interactúen con el sistema, se dispondrá de cada vez más información acerca de sus preferencias, y los otros dos elementos clave de nuestro enfoque híbrido de selección de vecindarios —la similitud semántica entre usuarios y las relaciones de confianza implícita— ganarán más y más importancia en el proceso. Nótese que, a diferencia de las relaciones de confianza personal (que reflejan confianza subjetiva de un individuo en otro, consecuencia de la relación social existente entre ellos), las relaciones de confianza implícita son inferidas por el sistema (entre dos individuos que, por norma general, no se conocen) en base a la fiabilidad de uno cuando ha desempeñado el rol de vecino en recomendaciones hechas al otro, y viceversa.

A fin de combinar los citados elementos de la forma más conveniente, en función del progreso en el sistema del usuario activo considerado en un momento dado, nuestro enfoque define los dos factores de ponderación descritos a continuación: la *completitud del perfil* de un usuario y la *madurez* de las relaciones de confianza implícita.

7.3.1 Completitud del perfil

Tal como su nombre indica, la completitud del perfil de un usuario hace referencia al nivel de detalle del mismo. En otras palabras, el propósito de dicho factor es reflejar la cantidad de información registrada en los perfiles de los usuarios acerca de sus preferencias, hasta un momento dado.

La completitud de los perfiles determinará el peso de los valores de confianza personal frente a la similitud semántica entre usuarios en la selección de vecindarios, de forma que:

- Cuando el perfil de un usuario activo es todavía muy incompleto (e.g., debido a que el usuario acaba de registrarse en el sistema) y, por tanto, no es posible

inferir medidas de similitud fiables, las relaciones de confianza personal tendrán un mayor peso en el proceso.

- En cuanto el perfil alcance un nivel de completitud suficiente, las medidas de similitud semántica tendrán un mayor impacto. Tal premisa es razonable, ya que nuestra métrica de similitud semántica posibilita identificar a individuos con preferencias similares a las del usuario activo en la clase de productos considerada en la recomendación, mientras que las relaciones de confianza personal no implican necesariamente similitud —por ejemplo, dos buenos amigos pueden tener intereses diferentes—.

A fin de medir la completitud del perfil de un usuario, nuestra estrategia tiene en cuenta, además de la cantidad de instancias de productos registradas en dicho perfil, su *heterogeneidad*. En otras palabras, la métrica formulada no sólo atiende al número de productos almacenados en el perfil, sino que también considera cómo están clasificados esos productos en la ontología.

Heterogeneidad

Para definir la heterogeneidad de un perfil, la estrategia propuesta se vale de la ontología de productos comerciales, más concretamente, del subárbol propio de la clase C_p considerada en la recomendación —i.e., la clase bajo la que está clasificada el producto objetivo de la recomendación, en el nivel de abstracción seleccionado (ver Sección 3.2)—. A modo de ilustración, en la Fig. 7.2 están representados los subárboles de la clase C_p, delineados en negro, correspondientes a los perfiles de dos usuarios. Nótese que ambos perfiles contienen cuatro instancias de productos en el subárbol de C_p: i_1, i_2, i_3, i_4, uno, y j_1, j_2, j_3, j_4, el otro.

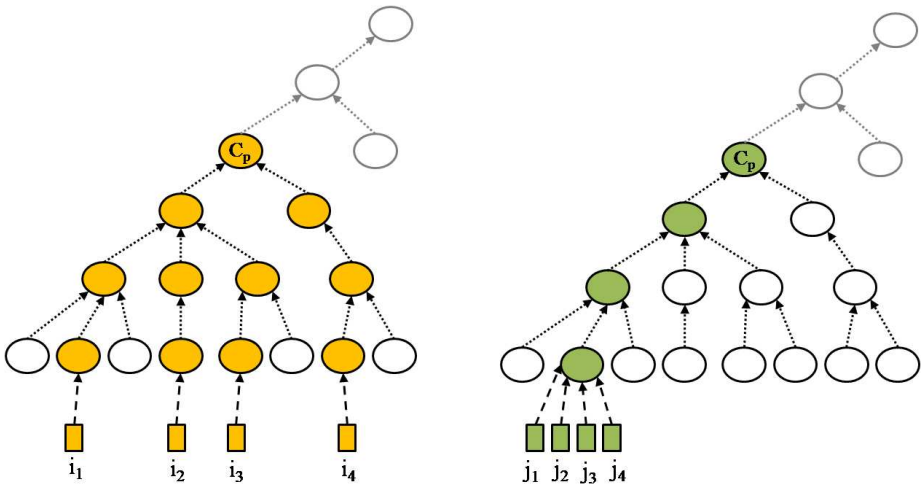


Figura 7.2: Heterogeneidad

En particular, nuestra estrategia atiende al número de clases diferentes *activadas* por los productos de un cierto perfil, en el subárbol considerado. Dada una instancia registrada en el perfil de un usuario, se consideran activadas: la clase hoja bajo la que está clasificada dicha instancia en la ontología, y todas sus superclases hasta C_p. Es decir, se reconocen las clases (ascendiendo en la jerarquía) desde la instancia que origina la activación hasta el raíz del subárbol considerado, C_p. En el escenario mostrado en la Fig. 7.2, las clases activadas en cada perfil están coloreadas.

Cuanto mayor sea el número de clases diferentes activadas por los productos de un perfil, mayor será su heterogeneidad. Por tanto, una instancia nueva aumentará la heterogeneidad del perfil siempre y cuando provoque nuevas activaciones. A la vista de la Fig. 7.2, es inmediato deducir que el perfil representado en la parte izquierda es más heterogéneo que el representado en la parte derecha, dado que el número de clases activadas es mucho mayor en el primero. En efecto, si consideramos que el perfil de la parte derecha de la figura inicialmente sólo contiene la instancia j_1 y, a continuación, son incorporadas j_2, j_3 y j_4, podemos observar como ninguna de estas instancias aumenta la heterogeneidad del perfil, puesto que ninguna de ellas provoca nuevas activaciones. Por el contrario, si en el perfil de la parte izquierda de la figura consideramos en un principio únicamente la instancia i_1, la incorporación de cualquiera de las otras instancias (i_2, i_3 e i_4) provoca nuevas activaciones y, por tanto, aumenta la heterogeneidad del perfil. En concreto, la incorporación de la instancia i_4 aumenta en mayor medida dicha heterogeneidad, ya que activará tres clases nuevas frente a las dos que activan i_2 e i_3.

Esta idea está formalizada en la Ec. 7.6, mediante la que se cuantifica la heterogeneidad del perfil de un usuario A recorriendo el subárbol pertinente en la jerarquía de clases nivel a nivel —desde el nivel correspondiente al raíz del subárbol, l_{C_p}, hasta el de clases hoja, l_{C_L}—. En cada nivel, se contabiliza el número de clases activadas y se asigna un mayor peso cuanto mayor sea la profundidad del nivel considerado. De este modo, la Ec. 7.6 conduce a mayores valores de heterogeneidad cuando la proporción de clases activadas sea alta en los niveles que representan conceptos más concretos. Esta premisa es razonable, ya que el perfil contendrá una mayor variedad de productos cuando las instancias correspondientes estén más distribuidas en las diferentes clases hoja de la ontología. Por tanto, en dicha expresión, $\#C_i^{T_A}(l)$ es el número de clases del nivel l del subárbol activadas en el perfil de A, $\#C_i(l)$ el número total de clases de ese nivel pertenecientes a dicho subárbol, y f_N un factor de normalización que fija el valor máximo de heterogeneidad a 1.

$$H_A(C_p) = \frac{1}{f_N} \sum_{l=l_{C_p}}^{l_{C_L}} \frac{\#C_i^{T_A}(l)}{\#C_i(l)} \cdot \frac{l}{l_{C_L}} \qquad \forall C_i \in subtree(C) \tag{7.6}$$

donde:

$$f_N = \frac{1}{l_{C_L}} \sum_{l=l_{C_p}}^{l_{C_L}} l$$

Dados dos perfiles que contienen el mismo número de productos clasificados bajo el subárbol en cuestión, nuestro enfoque considera más completo aquel cuya heterogeneidad sea mayor, puesto que su diversidad facilita la identificación de otros usuarios con preferencias similares en ese tipo de productos.

Multiplicidad de instancias

En cuanto a la consideración del número de instancias incluidas en el perfil, cuanto mayor sea el número de instancias clasificadas bajo cada clase hoja del subárbol de C_p, mayor será la completitud de ese perfil respecto a C_p. En concreto, se define la *multiplicidad de instancias* del perfil de un usuario A como el promedio de los factores de multiplicidad asociados a las clases hoja del subárbol en cuestión, $f_M(C_{Li})$, tal como muestra la Ec. 7.7, siendo N es el número de clases hoja en dicho subárbol y C_{Li} la i-ésima de esas clases hoja. El factor de multiplicidad de una cierta clase hoja depende del número de productos de esa clase registrados en el perfil de A, $\#I_A(C_{Li})$, respecto a $\#I_m(C_{Li})$, un parámetro configurable que delimita el número de instancias de esa clase a partir del cual ésta se considera lo suficientemente incluida en el perfil.

$$IM_A(C_p) = \frac{1}{N}\sum_{i=1}^{N} f_M(C_{Li}) \tag{7.7}$$

donde:

$$f_M(C_{Li}) = \begin{cases} 1 & \text{si } \#I_A(C_{Li}) \geq \#I_m(C_{Li}) \\ \frac{\#I_A(C_{Li})}{\#I_m(C_{Li})} & \text{si } \#I_A(C_{Li}) < \#I_m(C_{Li}) \end{cases}$$

Expresión matemática de la completitud del perfil

Reuniendo las contribuciones de los dos factores presentados, $H_A(C_p)$ e $IM_A(C_p)$, resulta la expresión para el cálculo de la completitud del perfil del usuario A respecto a C_p, mostrada en la Ec. 7.8.

$$PC_A(C_p) = H_A(C_p) \cdot IM_A(C_p) \tag{7.8}$$

Por tanto, la completitud de un perfil respecto a una cierta clase C_p será mayor cuanto mayor sea su heterogeneidad y mayor sea su multiplicidad de instancias, en relación a esa clase —i.e., atendiendo a los productos registrados en el perfil que están clasificados bajo el subárbol propio de dicha clase en la ontología—. A modo de ejemplo, comparemos el subárbol de la clase C_p representado en la Fig. 7.3, correspondiente al perfil de un tercer usuario, con los descritos en la Fig. 7.2. Recordemos que la Fig. 7.2 fue utilizada para comparar la heterogeneidad de los perfiles de dos usuarios respecto a la clase C_p, evidenciando la mayor heterogeneidad del perfil representado en la parte izquierda de la figura. A la vista de la Fig. 7.3, es sencillo deducir que el perfil mostrado en ella obtendrá un mayor valor de completitud respecto a C_p que los otros dos, ya que teniendo la misma heterogeneidad que el perfil de la parte izquierda de la Fig. 7.2, tiene una mayor multiplicidad de instancias. En efecto, distribuidas en las mismas clases hoja, frente a las cuatro instancias de aquel (i_1, i_2, i_3, i_4), el perfil representado en la Fig. 7.3 tiene nueve (k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8 y k_9).

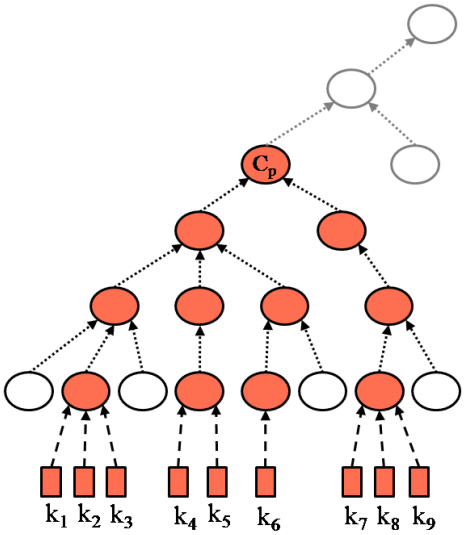


Figura 7.3: Completitud del perfil

7.3.2 Madurez

La medida de madurez definida en nuestro enfoque refleja la fiabilidad de cada relación de confianza implícita inferida hasta el momento (mediante la estrategia descrita en la Sección 5.2). En vista de que dichas relaciones son estimadas a partir de los resultados obtenidos en recomendaciones previas, cuanto mayor sea el número de resultados en el que está basada una cierta relación de confianza implícita, más fiable será su valor. Es decir, cuanto mayor sea el historial de éxitos y fracasos (concernientes a los usuarios considerados) disponible en el proceso de inferencia de una determinada relación de confianza implícita, mayor será la madurez de ésta.

En particular, se define la madurez de la relación de confianza implícita inferida entre los usuarios A y B, respecto a la clase C_p, mediante la Ec. 7.9, donde N es el número de clases hoja en el subárbol de C_p, C_{Li} la i-ésima de esas clases hoja y $f_R(C_{Li})$ el factor de madurez asociado a dicha clase. Este factor depende del número de resultados de recomendaciones anteriores en que se basa el cálculo de la relación de confianza implícita relativa a la clase hoja C_{Li} —$M(C_{Li})$— respecto a r, un parámetro configurable que delimita el número de resultados a partir del cual se considera el valor de confianza inferido lo suficientemente maduro.

$$M_{IT_{A,B}}(C_p) = \frac{1}{N}\sum_{i=1}^{N} f_R(C_{Li}) \tag{7.9}$$

donde:

$$f_R(C_{Li}) = \begin{cases} 1 & \text{si } M(C_{Li}) \geq r \\ \frac{M(C_{Li})}{r} & \text{si } M(C_{Li}) < r \end{cases}$$

Estas medidas de madurez permiten tener en cuenta, en el proceso de selección de vecindarios, no sólo la existencia de relaciones de confianza implícita entre el usuario activo y los candidatos a vecinos, también la fiabilidad de las mismas. En efecto, cuanto mayor es la madurez de una relación de confianza implícita, dicha relación refleja un mayor número de experiencias previas entre los usuarios considerados y, por tanto, más significativo es el valor asociado a la misma. De ahí que, en nuestro enfoque, el peso de cada medida de confianza implícita en el cálculo de la idoneidad de los candidatos a vecinos estará condicionado a su madurez.

Nótese que las relaciones de confianza implícita inferidas en nuestra metodología proporcionan una medida de precisión, puesto que reflejan el historial de éxitos y fracasos de los usuarios en sus contribuciones como vecinos en recomendaciones de un determinado tipo de productos. Por este motivo, tal como describiremos a continuación, en cuanto las relaciones de confianza implícita obtenidas respecto al usuario activo en cuestión superan un cierto nivel de madurez, tienen un gran impacto en el proceso de selección de vecindarios.

7.4 Reuniendo las piezas del puzle

Llegados a este punto, nuestro enfoque dispone de un conjunto de estrategias independientes desarrolladas a fin de incrementar la fiabilidad del proceso de selección de vecindarios y, en consecuencia, la calidad de las recomendaciones ofrecidas. Dichas estrategias permiten obtener, de forma transparente a los usuarios, gran cantidad de información útil acerca de los mismos y de las relaciones establecidas entre ellos:

- **Medidas de similitud semántica**. El enfoque semántico adoptado en el método de cuantificación de medidas de similitud entre usuarios, descrito en la Sección 4.3, permite inferir relaciones de afinidad significativas entre los usuarios del sistema. Precisamente, la explotación de la ontología de dominio posibilita la incorporación de nuevo conocimiento acerca de las preferencias de los usuarios en dicha métrica, permitiendo mitigar dos importantes limitaciones propias de los sistemas colaborativos: el problema de falsos vecindarios y el *sparsity problem*. Se trata de medidas de similitud condicionadas a clases de productos, por lo que se utilizarán las más apropiadas en cada proceso de selección de vecindarios en función del producto objetivo de la recomendación.

- **Relaciones de confianza implícita**. La aplicación de técnicas de razonamiento semántico sobre los resultados obtenidos en recomendaciones previas, tal como se describe en la Sección 5.2, permite inferir relaciones de confianza implícita de calidad entre los usuarios del sistema. Dichas relaciones denotan medidas de fiabilidad de los usuarios en su rol de vecinos, reflejando su historial de éxitos y fracasos en contribuciones a recomendaciones de productos de un cierto tipo.

- **Relaciones de confianza personal**. Tal como detallamos en la Sección 7.2, aprovechando la información contenida en redes sociales accesible al sistema, la estrategia propuesta cuantifica la intensidad de las relaciones existentes entre los usuarios del recomendador, que también estén registrados en alguna de estas redes, y sus respectivos contactos sociales. En concreto, las denominadas relaciones de confianza personal se estiman atendiendo al historial de interacciones entre los individuos y a las configuraciones de privacidad especificadas en sus perfiles sociales. Dado que estas relaciones son obtenidas a partir de información externa al sistema recomendador, permiten mitigar el problema de *cold start users* típico de las técnicas de filtrado colaborativas.

- **Medidas de completitud de los perfiles**. El factor de completitud del perfil de un usuario, obtenido mediante el método descrito en la Sección 7.3.1, ofrece una medida de la evolución de dicho usuario en el sistema recomendador. Cuando un usuario acaba de darse de alta en el sistema, su perfil contendrá únicamente la información que haya proporcionado en el momento del registro (ver Sección 3.3.2). Con el paso del tiempo, el perfil registrará cada vez más información acerca del usuario, adquirida a partir de la realimentación explícita y/o implícita del usuario (ver Sección 3.3.3). El enfoque semántico adoptado al cuantificar la completitud de un perfil posibilita: por un lado, obtener valores condicionados a clases de productos (en consonancia con las medidas de similitud semántica y las relaciones de confianza implícita inferidas en nuestra metodología); y, por otra parte, considerar no sólo el número de productos almacenados en dicho perfil en el cálculo, sino también su heterogeneidad.

- **Medidas de madurez de las relaciones de confianza implícita**. Tal como expusimos en la sección previa, el factor de madurez de una relación de confianza implícita proporciona una medida de su fiabilidad. Al igual que las medidas de completitud de los perfiles, éstas también son un reflejo de la evolución del usuario en el sistema. En efecto, a medida que un usuario recibe recomendaciones por parte del sistema, se establecen nuevas relaciones de confianza implícita con los usuarios que han formado parte de su vecindario y se afianzan las ya existentes con aquellos que han sido seleccionados como vecinos suyos repetidas veces, aumentando su madurez.

El objetivo final perseguido es, tal como hemos comentado anteriormente, combinar las distintas piezas del puzle de forma que se optimice el proceso de selección de vecindarios en función del usuario activo considerado. Para ello, nuestro enfoque considera dos premisas:

- Cuando la completitud de los perfiles es baja y, por tanto, los perfiles no están lo suficientemente detallados para obtener medidas de similitud entre usuarios fiables, se debe asignar un mayor peso a aquellos candidatos que gozan de una mayor confianza subjetiva por parte del usuario activo considerado (i.e., aquellos individuos de su círculo social cuya relación de confianza personal es más positiva), frente a los candidatos que obtienen mayores valores de similitud, y viceversa.

- El peso de las relaciones de confianza implícita ha de estar condicionado a su madurez: cuanto mayor sea la madurez de las relaciones, mayor será el peso que se les otorgue en la selección del vecindario. Teniendo en cuenta que dichas

relaciones proporcionan una medida de precisión en los resultados de las recomendaciones previas, cuando alcancen un cierto nivel de madurez se les asignará el mayor peso en el proceso.

Conforme a estas premisas, reuniendo todos los conceptos y parámetros presentados hasta ahora, resulta la expresión matemática que constituye el quid del enfoque híbrido propuesto para la selección de vecindarios. Dicha expresión, mostrada en la Ec. 7.10, permite cuantificar la idoneidad de un candidato CN_i a formar parte del vecindario de un usuario activo U para la recomendación de productos de clase C_p.

$$S_U^{C_p}(CN_i) = M_{IT_{U,CN_i}}(C_p) \cdot IT_{U,CN_i}(C_p) + (1 - M_{IT_{U,CN_i}}(C_p)) \cdot \\ \cdot (PC_{U,CN_i}(C_p) \cdot Sim_{U,CN_i}(C_p) + (1 - PC_{U,CN_i}(C_p)) \cdot PT_{U,CN_i}) \tag{7.10}$$

donde:

$$PC_{U,CN_i}(C_p) = PC_U(C_p) \cdot PC_{CN_i}(C_p)$$

En definitiva, en esta expresión:

- PT_{U,CN_i} es la relación de confianza personal de U a CN_i, obtenida según el método descrito en la Sección 7.2.
- $Sim_{U,CN_i}(C_p)$ se refiere a la similitud semántica entre las preferencias de U y CN_i en productos clasificados bajo la clase C_p en la ontología de productos, inferida mediante la estrategia detallada en la Sección 4.3.
- $IT_{U,CN_i}(C_p)$ es el valor asociado a la relación de confianza implícita, condicionada a C_p, inferida de U a CN_i según el método presentado en la Sección 5.2.
- $M_{IT_{U,CN_i}}(C_p)$ es la madurez de dicha relación de confianza implícita, obtenida tal como expusimos en la Sección 7.3.2.
- $PC_{U,CN_i}(C_p)$ es la completitud de los perfiles de U y CN_i respecto a C_p, calculada como el producto de la completitud del perfil de U respecto a C_p, $PC_U(C_p)$, por la completitud del perfil de CN_i respecto a C_p, $PC_{CN_i}(C_p)$ —obtenidas según el método descrito en la Sección 7.3.1—.

Como resultado, el vecindario del usuario U para recomendaciones de productos de clase C_p estará formado por los K candidatos a vecino que obtengan los valores más positivos de $S_U^{C_p}(CN_i)$[2].

A modo de ilustración, la Fig. 7.4 muestra una visión general del proceso de selección de vecindarios híbrido propuesto, en la que se incluyen las interacciones con los elementos clave del sistema:

[2]Recordemos que, en la fase de predicción, el peso de cada uno de los vecinos seleccionados se modulará según la experiencia práctica de los mismos, de acuerdo a lo expuesto en el Capítulo 6.

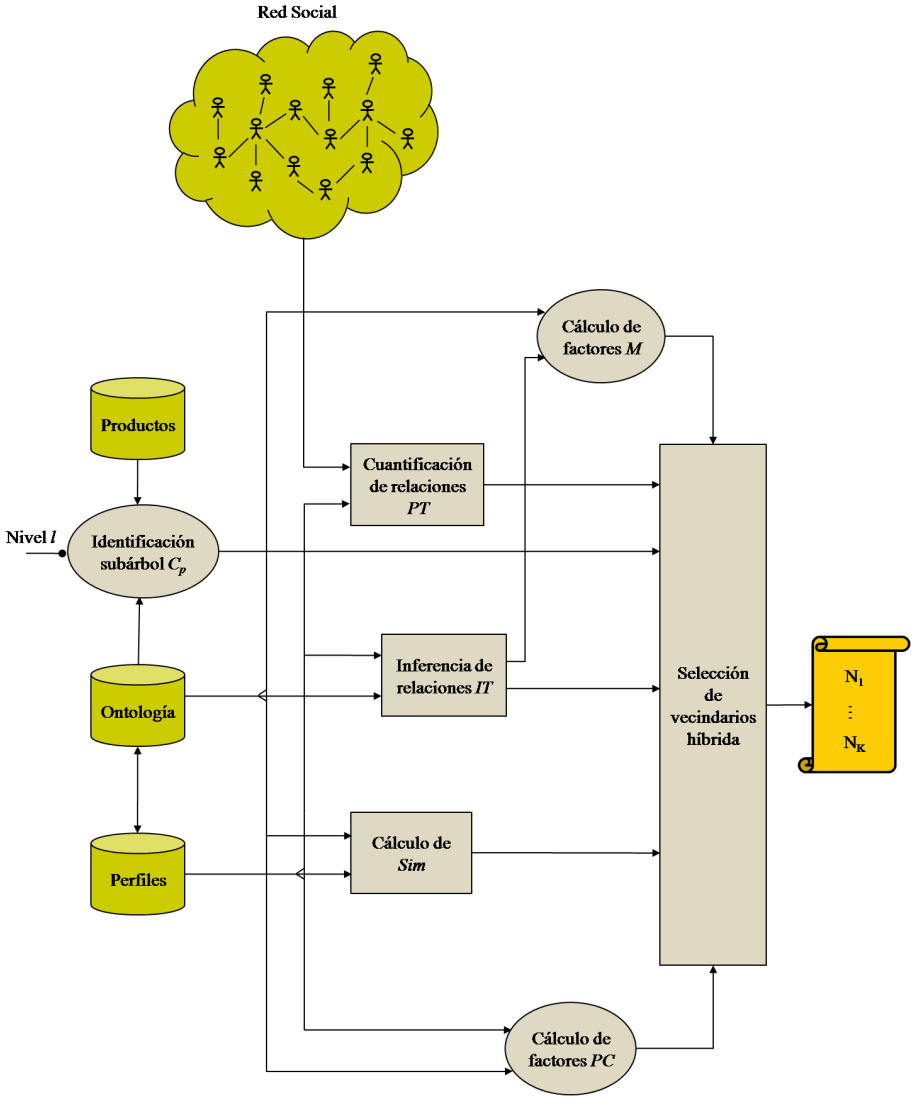


Figura 7.4: Proceso de selección de vecindarios híbrida

- En primer lugar, dado un cierto producto objetivo, nuestra metodología identifica, en función del nivel de abstracción considerado, la clase C_p correspondiente (en la figura, *Identificación subárbol* C_p). Dicha clase determina las medidas de similitud, las relaciones de confianza implícita y los factores de madurez y completitud de perfiles que serán utilizados en el proceso.

- En consonancia con el enfoque semántico adoptado en sus respectivas estrategias, tanto el proceso de inferencia de relaciones de confianza implícita (*Inferencia de relaciones IT*), como el de cuantificación de medidas de similitud entre usuarios (*Cálculo de Sim*), como los procesos de cálculo de factores de madurez y completitud de perfiles (en la Fig. 7.4, *Cálculo de factores M* y *Cálculo de factores PC*, respectivamente), se valen de la ontología de dominio del sistema para derivar nuevo conocimiento acerca de los usuarios.

- Por otro lado, de acuerdo con lo expuesto a lo largo de este capítulo, el proceso de cuantificación de relaciones de confianza personal (*Cuantificación de relaciones PT*) se apoya en la información contenida en las redes sociales a las que pertenezca el usuario activo considerado.

7.5 Beneficios de la estrategia

El incorporar medidas de confianza personal obtenidas a partir de la información disponible en redes sociales (i.e., información externa al sistema recomendador) en el proceso de selección de vecindarios, permite mitigar el problema de *cold start users* inherente a los sistemas recomendadores y, por tanto, obtener recomendaciones de mayor calidad para aquellos usuarios que acaban de registrarse en el sistema.

La gran ventaja asociada a la estrategia propuesta para la cuantificación de dichas relaciones de confianza es que, aun basándose en conexiones explícitas entre usuarios (las establecidas entre un individuo y sus contactos en una red social), son obtenidas sin necesidad de requerir a los usuarios que proporcionen ningún tipo de información de forma expresa. Dicho de otro modo, se consigue mejorar la calidad de las recomendaciones de forma transparente al usuario.

En cuanto a la combinación de las relaciones de confianza personal resultantes con los otros dos elementos claves de nuestra estrategia final de selección de vecindarios (las medidas de similitud semántica entre usuarios y las relaciones de confianza implícita), el enfoque semántico empleado en la definición de los factores de ponderación presentados en este capítulo permite reunir dichos elementos de forma significativa, teniendo en cuenta el progreso en el sistema del usuario activo considerado en un momento dado.

Parte IV

Conclusiones

Capítulo 8

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 a las estrategias propuestas en nuestra metodología. Seguidamente, se presentan algunas líneas de investigación futura que extenderían, de manera natural, el trabajo desarrollado.

8.1 Conclusiones

El trabajo desarrollado en esta tesis arrancó con el objetivo de enriquecer las capacidades de personalización ofrecidas por los sistemas recomendadores actuales, superando las limitaciones identificadas en ellos de forma transparente a los usuarios, esto es, sin requerir que éstos proporcionen ningún tipo de información adicional de forma explícita. A tal fin, nuestro enfoque recurre a técnicas de razonamiento, tradicionalmente vinculadas a la Web Semántica, para inferir nuevo conocimiento acerca de los usuarios y de las relaciones existentes entre ellos.

Si bien la metodología propuesta es suficientemente flexible y genérica para ser reutilizada en diferentes dominios de aplicación, hemos adoptado un contexto específico para poder ilustrar el enfoque con claridad. En concreto, puesto que en ella se combate el problema de falsos vecindarios (ver Capítulo 4), derivado de la aplicación de sistemas recomendadores colaborativos en dominios donde existe una gran diversidad en la naturaleza de los productos disponibles, se ha seleccionado a tal fin el dominio del comercio electrónico —caracterizado por la heterogeneidad de sus productos—. De ahí que la base de conocimiento sobre la que se asienta el sistema sea una ontología de productos comerciales.

Con el propósito de explotar el potencial de la semántica representada en la ontología, nuestro enfoque modela las preferencias de cada usuario a partir de la misma. En concreto, junto a los niveles de interés del usuario (que pueden ser declarados explícitamente por éste, o bien inferidos por el sistema a partir de su comportamiento de consumo), cada producto almacenado en su perfil tiene asociada una referencia única a la instancia correspondiente a dicho producto en la conceptualización OWL. Mediante tales referencias es posible consultar en la ontología todas las descripciones semánticas asociadas al producto considerado directamente.

Revisamos a continuación cada una de las estrategias desarrolladas en nuestra metodología, así como la estrategia final de selección de vecindarios híbrida que combina las contribuciones pertinentes. Teniendo en cuenta que la selección de vecindarios es la etapa clave de un proceso de recomendación colaborativo, ya que de ella depende el éxito en la recomendación al usuario activo, no es de extrañar que la mayoría de las técnicas presentadas en esta tesis estén enfocadas a enriquecer dicha etapa.

8.1.1 Estrategia para combatir el problema de falsos vecindarios

En esta primera estrategia, el objetivo es evitar una selección de vecindarios contraproducente cuando en el dominio de aplicación del sistema recomendador colaborativo existe una gran diversidad de productos. En un dominio como el del comercio electrónico, las distintas categorías de productos delimitan artículos de muy diferente naturaleza (como libros, música o productos alimenticios), por lo que tener en cuenta todas las preferencias registradas en los perfiles dos usuarios a la hora de estimar su similitud de intereses provoca la selección de falsos vecinos, esto es, usuarios afines al usuario activo en gran cantidad de ítems pero con intereses opuestos respecto al tipo de productos considerado en la recomendación en curso.

Para evitar la elección de falsos vecinos y, al mismo tiempo, detectar candidatos afines respecto al producto objetivo que pasarían desapercibidos por no compartir otros intereses con el usuario activo, la estrategia propuesta selecciona el vecindario dinámicamente, en función del tipo de producto contemplado en la recomendación. Construyendo distintos vecindarios para cada usuario activo, según los diferentes tipos de productos disponibles, logramos identificar qué candidatos a vecinos son los más apropiados en cada recomendación concreta.

Nuestro enfoque aprovecha la estructura jerárquica definida en la ontología para clasificar de forma significativa los productos a considerar en cada selección de vecindario. Por su naturaleza, dicha estructura permite dotar de gran flexibilidad al proceso, en función del nivel de abstracción considerado. De este modo, la estrategia desarrollada posibilita realizar una formación de vecindarios por categorías de productos o por subcategorías más concretas.

Además, explotando técnicas de razonamiento sobre la ontología, es posible inferir gran cantidad de conocimiento útil sobre las preferencias de los usuarios. El descubrimiento de asociaciones semánticas complejas entre los productos almacenados en el perfil del usuario activo y los perfiles de los candidatos a vecinos nos permite identificar similitudes entre ellos ocultas en la base de conocimiento, mejorando notablemente el proceso de selección de vecindarios. Estas medidas de similitud semántica entre productos nos permiten enriquecer las métricas de similitud entre usuarios tradicionales (la *similitud del coseno* y la *correlación Pearson-r*), mitigando los efectos adversos del *sparsity problem*. En efecto, en lugar de seleccionar los vecinos de un usuario basándose únicamente en los productos específicos incluidos en sus respectivos perfiles, nuestro enfoque permite detectar similitudes entre sus preferencias aunque dichos perfiles no tengan registrados exactamente los mismos productos.

8.1.2 Estrategia para la inferencia de redes de confianza implícitas

El objetivo perseguido con esta estrategia es la incorporación de un modelo de confianza en el proceso de selección de vecindarios de forma transparente a los usuarios.

A tal fin, nuestro enfoque toma como punto de partida los resultados obtenidos en recomendaciones previas, dado que en ellas se establecen conexiones virtuales entre cada usuario activo y sus respectivos vecinos. En particular, una relación de confianza implícita entre un usuario activo y un cierto vecino toma un valor positivo o negativo (i.e., se traduce en una relación de confianza o de desconfianza) dependiendo de si las contribuciones del último en recomendaciones al primero han sido o no exitosas.

Para superar las limitaciones asociadas al uso de una aproximación sintáctica, que únicamente permite definir relaciones condicionadas a productos concretos (difícilmente reutilizables) o generalizar a partir de éstas (incurriendo en una *sobregeneralización*), nuestra estrategia emplea técnicas de razonamiento semántico sobre los resultados de partida. La exploración de la ontología de productos comerciales permite obtener relaciones de confianza implícita de mayor calidad, condicionadas a clases de productos de cualquier nivel de abstracción.

Estructurando todas las relaciones de confianza inferidas entre pares de usuarios respecto a cada clase de productos, se construyen redes de confianza implícita entre los usuarios del sistema recomendador. Precisamente la posibilidad de considerar diferentes niveles de abstracción a la hora de inferir las relaciones que conforman las redes, provee de una gran flexibilidad al proceso. En efecto, se establece una relación de compromiso entre precisión y *sparsity*. Por un lado, cuanto mayor sea el nivel considerado, mayor será el número de redes construidas y habrá una mayor dispersión de datos. Por otra parte, a medida que descendemos en la jerarquía las clases representan conceptos más específicos, lo que conducirá a descubrimientos más precisos.

Todavía cabe señalar que, a través de la exploración de las redes resultantes, nuestro enfoque cuantifica la medida de reputación de cada usuario respecto a la clase de productos correspondiente. Para ello, la métrica propuesta reúne las valoraciones de confianza que un usuario recibe de otros, teniendo en cuenta a su vez la reputación de estos últimos. Esto permite disponer de información adicional sobre la fiabilidad de los candidatos a vecinos en el proceso de selección de vecindarios: lo fiables que son considerados por la comunidad en conjunto.

8.1.3 Estrategia para la cuantificación de la experiencia práctica de los usuarios

El beneficio más destacable de esta estrategia es su carácter implícito, esto es, su capacidad de evaluar la experiencia de los usuarios del sistema recomendador sin necesidad de que éstos proporcionen ningún tipo de información adicional. En efecto, gracias al enfoque semántico utilizado, no se requiere más información que el historial de consumo del usuario en cuestión —disponible en cualquier sistema recomendador—.

Aplicando técnicas de razonamiento semántico sobre dicha información de consumo, la estrategia propuesta cuantifica la experiencia práctica de los usuarios en base a los productos que éstos han probado hasta el momento.

En particular, nuestra métrica infiere la experiencia práctica de un usuario sobre una determinada clase hoja de la ontología de productos, en función de dos factores: la *variabilidad cuantitativa*, que depende de la variedad de productos clasificados bajo esa clase registrados en el perfil del usuario; y la *variabilidad cualitativa*, que explora las características semánticas definidas en la ontología relacionadas con esos productos a fin de estimar lo diferentes que son —i.e., depende de la diversidad de productos, de la clase en cuestión, que el usuario ha probado—.

El objetivo último de la estrategia es incorporar las susodichas medidas de experiencia en la fase de predicción del proceso de recomendación colaborativo, valorando la experiencia de los usuarios que forman parte del vecindario oportuno, a la hora de estimar el nivel de interés potencial del usuario activo en el producto objetivo de esa recomendación. Más concretamente, se trata de ponderar la contribución de cada vecino (i.e., su opinión acerca de si ese producto debe ser sugerido al usuario activo o no) dependiendo de su experiencia práctica en la clase hoja de dicho producto.

8.1.4 Estrategia para la cuantificación de relaciones de confianza personal entre usuarios

La finalidad de esta estrategia es mitigar el problema de *cold start users* que ha afectado a los sistemas recomendadores desde sus orígenes, derivado de la falta de información acerca de las preferencias de los usuarios cuando éstos acaban de registrarse en el sistema. Si bien el empleo de técnicas de razonamiento semántico en nuestra metodología permite incorporar gran cantidad de conocimiento útil en el proceso de recomendación y, por tanto, obtener recomendaciones de mayor calidad incluso en presencia de *sparsity*; es necesario disponer de un mínimo de información acerca de los intereses del usuario[1].

Para combatir este problema, es necesario incluir en el proceso de selección de vecindarios información externa al sistema recomendador. La estrategia propuesta aprovecha la masiva adopción de tecnologías Web 2.0 en la actualidad para obtener relaciones de confianza personal entre usuarios, a partir de la información disponible en las redes sociales. Tomando como base las relaciones explícitas existentes en una red social (cada conexión representa el acuerdo mutuo de dos usuarios a convertirse en contactos), nuestra estrategia examina la información contenida en los perfiles sociales a los que tiene acceso para cuantificar la intensidad de las relaciones de los usuarios del sistema con otros individuos a los que, por norma general, conocen personalmente.

En particular, la métrica especificada cuantifica las relaciones de confianza personal entre dos usuarios atendiendo al historial de interacciones entre ellos (i.e., las acciones de uno en la red social que involucren al otro) y a las configuraciones de privacidad que han especificado en sus respectivos perfiles sociales (más concretamente, los grupos de contactos que uno ha definido a fin de restringir el acceso a cierta información de su perfil y las implicaciones que conlleva para el otro dicha configuración). Cabe destacar que nuestro enfoque conduce a relaciones direccionales (i.e., en concordancia con la propiedad de asimetría inherente a la confianza) gracias a las desigualdades en las interacciones entre los usuarios y en sus configuraciones de privacidad.

La gran ventaja asociada a esta estrategia de cuantificación de relaciones de confianza es que son obtenidas, de forma externa al sistema recomendador, sin requerir a los usuarios que proporcionen ningún tipo de información de forma expresa.

8.1.5 Selección de vecindarios híbrida

Tal como acabamos de comentar, las relaciones de confianza personal extraídas de las redes sociales son especialmente beneficiosas para mejorar las recomendacio-

[1] Recordemos que, en el sistema propuesto, dicha información no tiene por qué ser proporcionada por el usuario de forma explícita, gracias a nuestro mecanismo de cuantificación de niveles de interés (llamados índices *DOI*) a partir de la realimentación de relevancia implícita obtenida por el sistema (ver Sección 3.3.4).

nes iniciales a usuarios que son nuevos en el sistema, esto es, para aliviar el problema de *cold start users* inherente a todo sistema recomendador. En efecto, cuando las preferencias e historial de consumo registrados en el perfil de un cierto usuario activo son escasos, formar su correspondiente vecindario con individuos de su círculo social puede mejorar en gran medida los resultados de las recomendaciones. Por un lado, se seleccionan como vecinos individuos en los que el usuario confía de forma subjetiva. Por otra parte, es habitual que esos individuos compartan intereses con el usuario activo en cuestión y, por tanto, sus perfiles sociales pueden contener información útil para el proceso.

Con el paso del tiempo, a medida que los nuevos usuarios interactúan con el sistema recomendador, éste recopila cada vez más información acerca de sus preferencias. Entonces, el sistema puede valerse de las estrategias semánticas propuestas en nuestro trabajo para obtener relaciones entre los usuarios de mayor calidad y enriquecer el proceso de selección de vecindarios.

El objetivo final perseguido ha sido, por tanto, idear una estrategia de selección de vecindarios híbrida, en la que se combinen las relaciones de confianza personal con las de similitud semántica y confianza implícita (inferidas según los métodos descritos en las Secciones 4.3 y 5.2, respectivamente), de modo que se tomen en mayor consideración, bien las relaciones extraídas de las redes sociales, o bien las obtenidas dentro del propio sistema recomendador, dependiendo de la evolución en el sistema del usuario activo considerado.

Para ello, se han definido dos factores de ponderación: la *completitud del perfil* de un usuario y la *madurez* de las relaciones de confianza implícita. El primero determina el peso de las relaciones de confianza personal frente a las de similitud semántica, de forma que: si el perfil de un usuario activo es todavía muy incompleto, se recurre a la información extraída de las redes sociales; y si ya es posible inferir medidas de similitud fiables, éstas tienen un mayor impacto en el proceso, puesto que identifican usuarios afines respecto al tipo de productos considerado en la recomendación (mientras que las relaciones de confianza personal no implican necesariamente una similitud de intereses). En cuanto al factor de madurez, refleja la fiabilidad de las relaciones de confianza implícita inferidas hasta un momento dado. Teniendo en cuenta que las relaciones de confianza implícita proporcionan una medida de precisión (porque reflejan el historial de éxitos y fracasos de los usuarios en sus contribuciones como vecinos en recomendaciones de un determinado tipo de productos), en cuanto superan un cierto nivel de madurez, tienen un gran peso en el proceso de selección de vecindarios.

8.2 Líneas de trabajo futuro

Son varios los frentes que podemos acometer para extender el trabajo presentado:

- En primer lugar, apuntamos una línea de trabajo que daría una nueva dimensión a la metodología desarrollada en esta tesis. Nos referimos a la combinación de la estrategia de personalización colaborativa incluida en nuestro enfoque con una estrategia basada en contenido, a fin de aprovechar en el proceso de recomendación las ventajas asociadas a estas últimas, mitigando sus inconvenientes (ver Sección 2.3.2).

  Precisamente, encontramos en la literatura una propuesta que explota la semántica formalizada en la ontología de dominio con el objetivo de aliviar el problema

de *sobreespecialización* inherente a las estrategias basadas en contenido [17]. Para ello, dicho enfoque emplea técnicas SA (*Spreading Activation*) sobre la base de conocimiento del sistema para descubrir contenidos relacionados de forma significativa con las preferencias del usuario activo, pero no necesariamente iguales a aquellos que le interesaron en el pasado.

Si bien la citada propuesta semántica se enmarca en el dominio de la Televisión Digital, podría ser adoptada como punto de partida y estudiar las implicaciones que supone su aplicación en dominios de aplicación heterogéneos —como el comercio electrónico—.

- Otra de las líneas de investigación que convendría ampliar es la búsqueda y procesamiento de datos en la Web 2.0, dado su enorme potencial como fuente de información externa al sistema recomendador acerca de los usuarios.

  En efecto, más allá de extraer información sobre sus preferencias interpretando qué tipos de productos o marcas comerciales se esconden detrás de las páginas hacia las que un cierto usuario ha mostrado interés en una red social (e.g., expresando "me gusta"), la utilización de técnicas de procesamiento del *lenguaje natural* permitiría descubrir gran cantidad de conocimiento adicional. Los comentarios que un usuario publica en su perfil social o en los de sus contactos en dichas redes, sus participaciones en foros de debate o blogs y, especialmente, las reseñas publicadas en sitios web de comercio electrónico, pueden contener abundante información acerca de qué productos interesan —y en qué medida— a dicho usuario. Esta información podría ser utilizada por nuestro sistema para enriquecer los perfiles de los usuarios, como una nueva forma de realimentación de relevancia implícita.

- La última línea de trabajo que apuntamos también está relacionada con las tecnologías de la Web 2.0, más concretamente, con su explotación como soporte publicitario. Cabe destacar, en el contexto de las redes sociales, el potencial del *marketing viral* para llegar a nuevos consumidores. Se trata de técnicas que intentan provocar la propagación del producto o la marca anunciada a manos de los usuarios, mediante un efecto similar al *boca-a-boca*, a través de la red.

  La metodología propuesta en esta tesis puede ser adoptada para dotar de una mayor focalización[2] a dicha publicidad social, puesto que la personalización de las soluciones actuales está limitada a identificar la audiencia objetivo de una campaña atendiendo únicamente a sus características demográficas.

  Por otro lado, nuestra estrategia para la cuantificación de relaciones de confianza personal entre usuarios puede ser utilizada como punto de partida para valorar la influencia de los usuarios en la red, a fin de optimizar los beneficios de la promoción del producto o marca en cuestión. Por ejemplo, en una campaña de distribución de cupones de descuento, los usuarios considerados más influyentes serán aquellos que tengan el potencial de desencadenar un mayor número de canjes de cupones del producto promocionado.

[2]Entendiendo *focalización* como la capacidad de ofrecer a cada individuo productos que encajan con sus intereses o necesidades. El objetivo perseguido no es otro que promocionar un cierto producto a sus consumidores potenciales.

Bibliografía

[1] Adomavicius G. and 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. and Srikant R. Fast Algorithms for Mining Association Rules. In *20th International Conference on Very Large Data Bases (VLDB-94)*, pages 487–499, 1994.

[3] Alpert J. and Hajaj N. *Google Official Blog. Disponible en http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html*. 2008.

[4] Amazon. *Disponible en http://www.amazon.com*. 2001.

[5] Anyanwu K. and Sheth A. ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. In *12th International World Wide Web Conference (WWW-03)*, pages 115–125, 2003.

[6] Arbib M. A. *The handbook of brain theory and neural networks*. Cambridge, 1995.

[7] Ardissono L., Gena C., Torasso P., Bellifemine F., Difino A. and Negro B. User modeling and recommendation techniques for personalized electronic program guides. In *Personalized Digital Television: Targeting Programs to Individual Viewers*, pages 3 – 26. Kluwer Academics Publishers, 2004.

[8] Armstrong R., Freitag D., Joachims T. and Mitchell T. WebWatcher: A Learning Apprentice for the World Wide Web. In *AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments*, pages 6–12, 1995.

[9] Asnicar F. and Tasso C. IfWeb: A Prototype of User-Models-Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web. In *Proceedings of 6th International Conference on User Modeling (UM-97)*, pages 14–24, 1997.

[10] Balabanovic M. and Shoham Y. Combining Content-Based and Collaborative Recommendation. *Communications of the ACM*, 40(3):1–9, 1997.

[11] Basu C., Hirsh H. and Cohen W. Recommendation as Classification: Using Social and Content-based Information in Recommendation. In *15th National Conference on Artificial Intelligence (AAAI-98)*, pages 714–720, 1998.

[12] Baudisch P. and Bruekner L. TV Scout: guiding users from printed TV program guides to personalized TV recommendation. In *1st Workshop on Personalization in Future TV*, pages 151 – 160, 2001.

[13] Billsus D. and Pazzani M. Learning Collaborative Information Filters. In *15th International Conference on Machine Learning (ICML-98)*, pages 46–54, 1998.

[14] Billsus D. and Pazzani M. A personal news agent that talks, learns and explains. In *3rd International Conference on Autonomous Agents (Agents-99)*, pages 268–275, 1999.

[15] Björkman M., Aroyo L, Bellekens M., Dekker T., Loef E. and Pulles R. Personalised Home Media Centre Using Semantically Enriched TV-Anytime Content. In *4th European Conference on Interactive Television (EuroITV-06)*, pages 195–173, 2006.

[16] Blanco-Fernández Y. *Propuesta metodológica para el razonamiento semántico en sistemas de recomendación personalizada y automática. Aplicación al caso de contenidos audiovisuales*. PhD thesis, Departamento de Enxeñería Telemática. Universidad de Vigo, 2007.

[17] Blanco-Fernández Y., López-Nores M., Gil-Solla A., Ramos-Cabrer M. and Pazos-Arias J. J. Exploring synergies between content-based filtering and spreading activation techniques in knowledge-based recommender systems. *Information Sciences*, 181(21):4823–4846, 2011.

[18] Boone G. Concept Features in RE:Agent, an Intelligent Email Agent. In *2nd International Conference on Autonomous Agents (Agents-98))*, pages 141–148, 1998.

[19] Bradley K. and Smyth B. Improving recommendation diversity. In *12th National Conference in Articial Intelligence and Cognitive Science (AICS-01)*, pages 75–84, 2001.

[20] Breese J., Heckerman D. and Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, 1998.

[21] Burke R. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[22] Carchiolo V., Longheu A., Malgeri M. and Mangioni G. Context-based global expertise in recommendation systems. *Informatica*, 34(4):409–417, 2010.

[23] CDNow. *Disponible en http://www.cdnow.com*. 2001.

[24] Chen L. and Sycara K. WebMate: A Personal Agent for Browsing and Searching. In *2nd International Conference on Autonomous Agents (AGENTS-98)*, pages 132–139. ACM Press, 1998.

[25] Cho Y. H., Kim J. K. Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications*, 26(2):233–246, 2004.

[26] Cooley R., Tan P. N. and Srivastava J. WebSift: The Web Site Information Filter System. In *KDD Workshop on Web Mining*, pages 24–32, 1999.

[27] Cotter P. and Smyth B. A personalized TV listing service. *Communications of the ACM*, 43(8):107–111, 2000.

[28] Cunningham P., Bergmann R., Schmitt S., Traphöner R., Breen S. and Smyth B. WebSell: Intelligent Sales Assistants for the World Wide Web. In *Proceedings of eBusinees and eWork Conference and Exhibition (E-01)*, pages 31–41, 2001.

[29] Davies J., Fensel D. and van Harmelen F. *Towards the Semantic Web: Ontology-driven Knowledge Management*. John Wiley & Sons, 2003.

[30] Demiriz A. An association mining-based product recommender. In *Proceedings of INFORMS Miami 2001 Annual Meeting Cluster: Data Mining*, pages 34 – 44, 2001.

[31] Diestel R. *Graph Theory*. Springer-Verlag, 2000.

[32] Dima M., Vogiatzis D., Paliouras G. and Stamatopoulos P. Expert based prediction of user preferences. In *5th International Workshop on Semantic Media Adaptation and Personalization*, pages 112–117, 2001.

[33] Ding L., Steil D., Dixon B., Parrish A. and Brown D. A relation context oriented approach to identify strong ties in social networks. *Knowledge-Based Systems*, 24(8):1187–1195, 2011.

[34] Domingos P. and Richardson M. Mining the network value of customers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[35] Ebay. *Disponible en http://www.ebay.com.*

[36] eCla@ss. *Disponible en http://www.eclass.de.*

[37] Efron M. and Geisler G. Is it all About Connections? Factors Affecting the Performance of a Link-Based Recommender System. In *SIGIR Workshop on Recommender Systems*, pages 34–43, 2001.

[38] eOTD. *Disponible en http://www.eccma.org.*

[39] Epinions. *Disponible en http://www.epinions.com.*

[40] Facebook. *Disponible en http://www.facebook.com.*

[41] FilmAffinity. *Disponible en http://www.filmaffinity.com.*

[42] Funakoshi K. and Ohguro T. Evaluation of integrated content-based collaborative filtering. In *Proceedings of ACM SIGIR Workshop on Recommender Systems*, pages 42–51, 2001.

[43] Ganesan P., Garcia-Molina H. and Widom J. Exploiting Hierarchical Domain Structure to Compute Similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.

[44] Golbeck J. Generating predictive movie recommendations from trust in social networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3986:93–104, 2006.

[45] Golbeck J. and Hendler J. Accuracy of metrics for inferring trust and reputation in semantic web-based social networks. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 33257:116–131, 2004.

[46] Golbeck J. and Hendler J. Inferring binary trust relationships in web-based social networks. *ACM Transactions on Internet Technology*, 6(4):497–529, 2006.

[47] Goldberg D., Nichols D., Oki B. and Terry D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[48] Good N., Schafer J., Konstan J., Borchers A., Sarwar B., Herlocker J. and Riedl J. Combining Collaborative Filtering with Personal Agents for Better Recommendations. In *17th National Conference on Artificial Intelligence (AAAI-99)*, pages 439–446, 1999.

[49] Google+. *Disponible en http://plus.google.com*.

[50] Google. *Disponible en http://www.google.com*.

[51] Gruber T. R. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[52] Hayes C. and Cunningham P. Smart Radio - a Proposal. In *Technical Report, Trinity College Dublin, Computer Science, TCD-CS-1999-24*, 1999.

[53] Hayes C. and Cunningham P. Smart radio - community based music radio. *Knowledge-Based Systems*, 14(3):197–201, 2001.

[54] Hepp M. Products and services ontologies: A methodology for deriving owl ontologies from industrial categorization standards. *International Journal on Semantic Web & Information Systems*, 2(1):72–99, 2006.

[55] Hepp M., Leukel J. and Schmitz V. A quantitative analysis of product categorization standards: Content, coverage and maintenance of ecl@ss, unspsc, eotd, and the rosettanet technical dictionary. *Knowledge and Information Systems*, 13(1):77–114, 2007.

[56] Hill W., Stead L., Rosenstein M. and Furnas G. Recommending and Evaluating Choices in a Virtual Community of Use. In *Proceedings of International Conference on Human Factors in Computing Systems (CHI-95)*, pages 194–201, 1995.

[57] Huberman B. and Kaminsky M. Beehive: A System for Cooperative Filtering and Sharing of Information. In *Technical Report, Dynamics of Computation Group*, 1996.

[58] Jennings A. and Higuchi H. A User Model Neural Network for a Personal News Service. *User Modeling and User-Adapted Interaction*, 3(1):1–25, 1993.

[59] Jensen F. V. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[60] Joachims T., Freitag D. and Mitchell T. WebWatcher: A Tour Guide for the World Wide Web. In *16th International Joint Conference on Artificial Intelligence (IJCAI-97)*, pages 770–777, 1997.

[61] Josang A., Ismail R. and Boyd C. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[62] Kamba T., Bharat K. and Albers M.C. The Krakatoa Chronicle - an Interactive, Personalized Newspaper on the Web. In *4th International World Wide Web Conference (WWW-95)*, pages 159–170, 1995.

[63] Kamba T., Sakagami H. and Koseki Y. ANATAGONOMY: A Personalized Newspaper on the World Wide Web. *International Journal of Human-Computer Studies*, 46(6):789–803, 1997.

[64] Kempe D., Kleingberg J. and Tardos É. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[65] Kiss C. and Bichler M. Identification of influencers – measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008.

[66] Kobsa A., Koenemann J. and Pohl W. Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review*, 16(2):111–155, 2001.

[67] Konstan J., Miller B., Maltz D., Herlocker J., Gordon L. and Riedl J. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[68] Krulwich B. LifeStyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. *AI Magazine*, 18(2):37–45, 1997.

[69] Krulwich B. and Burkey C. Contact Finder: Extracting indications of expertise and answering questions with referrals. In *Technical Report, FS-95-03. The AAAI Press*, 1995.

[70] Krulwich B. and Burkey C. Learning User Information Interests through Extraction of Semantically Significant Phrases. In *AAAI Spring Symposium on Machine Learning Learning in Information Access*, pages 22–27, 1996.

[71] Lang K. NewsWeeder: Learning to Filter News. In *12th International Conference on Machine Learning (ICML-95)*, pages 331–339, 1995.

[72] Last.fm. *Disponible en http://www.last.fm.*

[73] Lee J., Kim M. and Lee Y. Information Retrieval based on Conceptual Distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.

[74] Leskovec J., Adamic L. A. and Huberman B. A. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(5), 2007.

[75] Li Y., Lu L., Xuefeng L. A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in e-commerce. *Expert Systems with Applications*, 28:67–77, 2004.

[76] Lieberman H. Letizia: An Agent that Assists Web Browsing. In *14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 924–929, 1995.

[77] Lieberman H., Van Dyke N. and Vivacqua A. S. Let's Browse: A Collaborative Web Browsing Agent. In *International Conference on Intelligent User Interfaces (IUI-99)*, pages 924–929, 1999.

[78] Lin D. An Information-Theoretic Definition of Similarity. In *15th International Conference on Machine Learning (ICML-98)*, pages 296–304, 1998.

[79] Lin W., Álvarez S. and Ruiz C. Collaborative recommendation via adaptive association rule mining. In *Proceedings of ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD-00)*, pages 40 – 47, 2000.

[80] Ma H., King I. and Lyu M. R. Learning to recommend with explicit and implicit social relations. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

[81] Mak H., Koprinska I. and Poon J. INTIMATE: A Web-Based Movie Recommender Using Text Categorization. In *Proceedings of IEEE/WIC International Conference on Web Intelligence (WI-03)*, pages 602 – 612, 2003.

[82] Maloof M. and Michalski R. Selecting Examples for Partial Memory Learning. *Machine Learning*, 41(1):27–52, 2000.

[83] Martín-Vicente M. I., Gil-Solla A., Ramos-Cabrer M., Blanco-Fernández Y. and López-Nores M. 4th International Workshop on Semantic Media Adaptation and Personalization, SMAP'09. pages 9–14, 2009.

[84] Martín-Vicente M. I., Gil-Solla A., Ramos-Cabrer M., Blanco-Fernández Y. and López-Nores M. A semantic approach to avoiding fake neighborhoods in collaborative recommendation of coupons through digital tv. *IEEE Transactions on Consumer Electronics*, 56(1):54–62, 2010.

[85] Martín-Vicente M. I., Gil-Solla A., Ramos-Cabrer M., Blanco-Fernández Y. and López-Nores M. 6th International Workshop on Semantic Media Adaptation and Personalization, SMAP'11. pages 9–14, 2011.

[86] Martín-Vicente M. I., Gil-Solla A., Ramos-Cabrer M., Blanco-Fernández Y. and López-Nores M. Improving collaborative recommendation of coupons through digital tv by semantic inference of users'reputation. *IEEE Transactions on Consumer Electronics*, 57(1):178–186, 2011.

[87] Martín-Vicente M. I., Gil-Solla A., Ramos-Cabrer M., Blanco-Fernández Y. and López-Nores M. Semantic inference of user's reputation and expertise to improve collaborative recommendations. *Expert Systems with Applications*, 39(9):8248–8258, 2012.

[88] Massa P. and Avesani P. Trust-aware collaborative filtering for recommender systems. *Lecture Notes in Computer Science*, 3290:492–508, 2004.

[89] Massa P. and Avesani P. Trust metrics on controversial users: balancing between tyranny of the majority and echo chambers. *International Journal on Semantic Web and Information Systems*, 3(1):39–64, 2007.

[90] McDonald D. W. Evaluating expertise recommendations. In *Proceedings of the International ACM Conference on Supporting Group Work*, pages 214–223, 2001.

[91] Melville P., Mooney R. and Nagarajan R. Content-Boosted Collaborative Filtering for Improved Recommendations. In *18th National Conference on Artificial Intelligence (AAAI-2002)*, pages 187–192, 2002.

[92] Middleton S. *Capturing knowledge of user preferences with recommender systems*. PhD thesis, Department of Electronic and Computer Science. University of Southampton, 2003.

[93] Miller B., Albert I., Lam S., Konstan J. and Riedl J. Movielens unplugged: experiences with an occasionally connected recommender system. In *ACM Conference on Intelligent User Interfaces*, pages 263–266, 2003.

[94] Minio M. and Tasso C. User Modeling for Information Filtering on Internet Services: Exploiting an Extended Version for the UMT Shell. In *Workshop on User Modeling for Information Filtering on the World Wide Web*, pages 45–55, 1996.

[95] Mitchell T., Caruana R., Freitag D., McDermott J. and Zabowski D. Experience with a Learning Personal Assistant. *Communications of the ACM*, 37(7):81–91, 1994.

[96] Mitchell T. M., Mahadevan S. and Steinberg L. Leap: A Learning Apprentice for VLSI Design. In *4th International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 573–580, 1985.

[97] Mladenic D. Personal WebWatcher: Implementation and Design. In *Technical Report IJS-DP-7472, Department of Intelligent Systems*, 1996.

[98] Mobasher B., Cooley R. and Srivastava J. Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8):142–151, 2000.

[99] Mobasher B., Dai H., Luo T., Nakagawa M., Sun Y., Wiltshire J. Discovery of Aggregate Usage Profiles for Web Personalization. In *Proceedings of ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD-00)*, pages 61 – 68, 2000.

[100] Montaner M., López B. and De la Rosa J. A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review*, 19(4):285–330, 2003.

[101] Montaner M., López B. and De la Rosa J. L. Opinion-based filtering through trust. In *Proceedings of the International Conference on Autonomous Agents*, pages 304–305, 2002.

[102] Mooney R. and Roy L. Content-based book recommending using learning for text categorization. In *Workshop on Recommender Systems: Algorithms and Evaluation*, pages 195–204, 1999.

[103] Morita M. and Shinoda Y. Information filtering based on user behavior analysis and best match text retrieval. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 272–281, 1994.

[104] Moukas A. Amalthea: Information Filtering and Discovery Using in a Multiagent Evolving System. *Journal of Applied AI*, 11(5):437–457, 1997.

[105] Moviefinder Recommender System. Disponible en *http://www.moviefinder.com*.

[106] O'Donovan J. and Smyth B. Mining trust values from recommendation errors. *International Journal on Artificial Intelligence Tools*, 15(6):945–962, 2006.

[107] Orwant L. J. Heterogeneous Learning in the Doppelganger User Modelling System. *User Modeling and User-Adapted Interaction*, 42(2):107–130, 1995.

[108] O'Sullivan D., Smyth B., Wilson D. and McDonald K. Improving the Quality of the Personalized Electronic Program Guide. *User Modeling and User-Adapted Interaction*, 14(1):5–36, 2004.

[109] Pazzani M. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5):393–408, 1999.

[110] Pazzani M. and Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, 27(3):313–331, 1997.

[111] Pazzani M., Muramatsu J. and Billsus D. Syskill & Webert: Identifying Interesting Web Sites. In *13th National Conference on Artificial Intelligence (AAAI-96)*, pages 54–61, 1996.

[112] Quan T. K., Fuyuki I., Shinichi H. Improving accuracy of recommender system by clustering items based on stability of user similarity. In *Proceedings Of the International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, pages 61–61, 2007.

[113] Quillian R. Semantic memory. In *Semantic Information Processing*. The MIT Press, 1968.

[114] Quinlan J. R. Learning Efficient Classification Procedures and their Application to Chess End Games. In *Machine Learning: An Artificial Intelligence Approach*, pages 463–482, 1983.

[115] Rada R., Mili H., Bicknell E. and Blettner M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[116] Rafter R., Bradley K and Smyth B. Personalized retrieval for online recruitment services. In *22nd Annual Colloquium on Information Retrieval*, pages 58–69, 1999.

[117] Replay TV. Disponible en *http://www.replaytv.com*.

[118] 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.

[119] Rich E. User Modeling via Stereotypes. *International Journal of Cognitive Science*, 3(1):329–354, 1979.

[120] Riordan A. and Sorensen H. An intelligent agent for high-precision text filtering. In *4th International Conference on Information and Knowledge Management (CIKM-95)*, pages 205–211. ACM Press, 1995.

[121] RosettaNet. *Disponible en http://www.rosettanet.org*.

[122] Salton G. and McGill M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[123] Sarwar B., Karypis G., Konstan J. and Riedl J. Item-based collaborative filtering recommendation algorithms. In *10th International WWW Conference (WWW-01)*, pages 285–295, 2001.

[124] Schmitz C., Hotho A., Jäschke R. and Stumme G. Content Aggregation on Knowledge Bases using Graph Clustering. In *3rd European Semantic Web Conference (ESWC-06)*, pages 61–75, 2006.

[125] Schwab I., Kobsa A. and Koychev I. *Learning user interests through positive examples using content analysis and collaborative filtering*. Draft from Fraunhofer Institute for Applied Information Technology, Germany, 2001.

[126] Shardanand U. *Social Information Filtering for Music Recommendation*. PhD thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, 1994.

[127] Shardanand U. and Maes P. Social Information Filtering: Algorithms for Automating "Word of Mouth". In *Proceedings of International Conference on Human Factors in Computing Systems (CHI-95)*, pages 210–217, 1995.

[128] Sheth B. and Maes P. Evolving Agents for Personalized Information Filtering. In *9th International Conference on Artificial Intelligence for Applications (AIA-93)*, pages 331–339, 1993.

[129] Smyth B. and Cotter P. Surfing the digital wave: generating personalized tv listings using collaborative, case-based recommendation. In *3rd International Conference on Case-Based Reasoning (CBR-99)*, pages 561 – 571, 1999.

[130] Smyth B. and Cotter P. A personalized TV listing service for the Digital TV age. *Journal of Knowledge-Based Systems*, 13(2):53–59, 2000.

[131] Smyth B. and McClave P. Similarity vs. diversity. In *4th International Conference on Case-Based Reasoning (ICCBR-01)*, pages 347–361, 2001.

[132] Sorensen H. and McElligott M. PSUN: A Profiling System for Usenet News. In *Workshop on Intelligent Information Agents*, pages 121–127, 1995.

[133] Staab S. and Studer R. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

[134] Stefani A. and Strapparava C. Personalizing Access to Web Sites: The SiteIF Project. In *2nd Workshop on Adaptive Hypertext and Hypermedia (HYPERTEXT-98)*, pages 69–75, 1998.

[135] Tatemura J. Virtual Reviewers for Collaborative Exploration of Movie Reviews. In *ACM Intelligent User Interfaces Conference (IUI-2000)*, pages 272–275, 2000.

[136] TiVo, Inc. TiVo: TV your way. Disponible en *http://www.tivo.com*, year = 2002.

[137] Twitter. *Disponible en http://twitter.com*.

[138] UNSPSC. *Disponible en http://www.unspsc.org*.

[139] Webb G. and 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.

[140] Weigand L. A Multilingual Ontology-based Lexicon for News Filtering - The TREVI Project. In *15th International Joint Conferences on Artificial Intelligence (IJCAI-97)*, pages 160–165, 1997.

[141] Widmer G. and Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23(1):69–101, 1996.

[142] World Wide Web Consortium (W3C). *OWL 2 Web Ontology Language Document Overview. Disponible en http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.*

[143] Wu P., Liu W. and Jin C. A novel recommender system fusing the opinions from experts and ordinary people. In *ACM International Conference Proceeding Series*, pages 41–44, 2010.

[144] Yan T. and García-Molina H. Sift - a Tool for Wide-Area Information Dissemination. In *USENIX Technical Conference*, pages 177–186, 1995.

[145] Yuan W., Guan D., Lee Y. K. and Hur S. J. Improved trus-aware recommender system using small-worldness of trust networks. *Knowledge-Based Systems*, 23(3):232–238, 2010.

[146] Zhou X., Xu Y., Li Y., Josang A. and Cox C. The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review*, in press, 2011.

[147] Ziegler C.-N. and Golbeck J. Investigating interactions of trust and interest similarity. *Decision Support Systems*, 43(2):460–475, 2007.

[148] Ziegler C. N. and Lausen G. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4–5):337–358, 2005.

[149] Ziegler C. N., Lausen G. and Konstan J. A. On exploiting classification taxonomies in recommender systems. *AI Communications*, 21(2–3):97–125, 2008.

[150] Ziegler C. N., Schmidt-Thieme L. and Lausen G. Exploting Semantic Product Descriptions for Recommender Systems. In *2nd ACM SIGIR Semantic Web and Information Retrieval Workshop (SWIR-04)*, pages 25–29, 2004.

[151] Zimmerman J., Kurapati K., Buczak A., Schafer D., Gutta S. and Martino J. TV Personalization System: Design of a TV Show Recommender Engine and Interface. In *Personalized Digital Television: Targeting Programs to Individual Viewers*, pages 27 – 51. Kluwer Academics Publishers, 2004.