REAL ACADEMIA
GALEGA DE CIENCIAS

Crecimiento exponencial de la información. Del *big data* al *small data*.

Discurso del Académico numerario electo

Sr. D. José Juan Pazos Arias

Santiago de Compostela, 29 de junio de 2016

Más fácilmente se añade lo que falta que se quita lo que sobra.

Francisco de Quevedo y Villegas.

Excmo. Sr Presidente de la Real Academia Galega de Ciencias.
Excmos. Sras. e Sres. Académicos.
Autoridades, familiares, colegas, amigos, Sras y Sres.

1. Prólogo

Permítanme que las primeras palabras de este discurso de acceso sean de agradecimiento. En primer lugar, hacia los miembros de Real Academia de Ciencias de Galicia por su generosidad al haber propuesto y apoyado mi candidatura para formar parte como Académico numerario de esta prestigiosa Institución. Es para mí un honor al que espero poder corresponder en los próximos años, al menos, con mi entusiasta colaboración a la importante labor que esta Academia viene desarrollando con el objetivo principal de impulsar la promoción social del conocimiento como medio para lograr el desarrollo socioeconómico y el bienestar social de Galicia.

Quería manifestar también mi agradecimiento a todas las personas con las que he colaborado a lo largo de todos estos años en las diferentes actividades de docencia e investigación que he venido desarrollando. En especial, este agradecimiento va dirigido hacia todos los miembros del Grupo de Servicios para la Sociedad de la Información, que puse en marcha en el año 1995 y del que soy responsable. También quiero recordar en este momento a todos los colegas de la Escola de Enxeñería de Telecomunicación de la Universidad de Vigo.

En un terreno más personal, quiero manifestar también mi agradecimiento hacia mis amigos y, especialmente, a mi mujer y mis hijos Laura y Pedro por haberme aguantado y apoyado todos estos años, que empiezan ya a ser muchos, lo que hace que sea sin duda meritorio.

Voy a dedicar este discurso de acceso a analizar un hecho del que estamos siendo testigos en los últimos años y que no es otro que el crecimiento sin límite de la información de todo tipo disponible en Internet. Enmarcaré dentro de este análisis a los sistemas recomendadores, a cuyo desarrollo me he venido dedicando los últimos años, analizaré su situación actual y también sus posibles puntos de mejora.

2. Introducción

Cuenta la leyenda que cuando el creador del juego del ajedrez ofreció su invento al rey, el monarca se sintió maravillado y se recuperó de la profunda tristeza en que le había sumido la pérdida reciente de un hijo en una batalla. Agradecido por el regalo, el rey dejó que el inventor del juego eligiese la recompensa. Éste pidió la cantidad de trigo resultante de poner un grano en el primer cuadro del tablero, dos en el segundo, cuatro en el tercero, y así sucesivamente (ver figura 1).

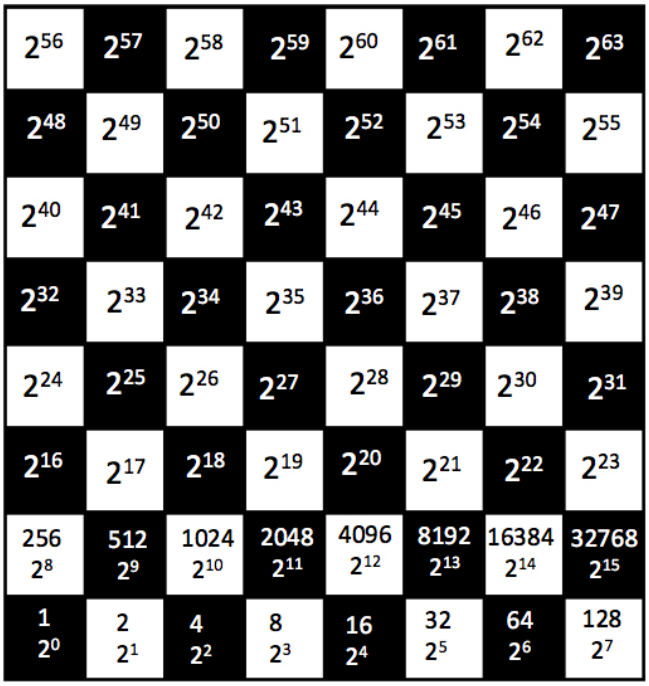

Figura 1: Tablero de ajedrez.

El monarca, aunque sorprendido con lo que consideraba una petición demasiado modesta para tan impresionante regalo, aceptó y ordenó a los expertos de la corte que calcularan la cantidad exacta de granos de trigo y se la entregaran de inmediato.

La sorpresa llegó cuando se dieron cuenta de que no disponían ni de lejos de tal cantidad de granos de trigo, toda vez que el total ascendía a más de 18 trillones de granos, concretamente a:

$$1 + 2 + 4 + 8 + \ldots + 2^{62} + 2^{63} = 18.446.744.073.709.551.615 \text{ granos de trigo}$$

cantidad suficiente para cubrir dos veces toda la superficie de la tierra, océanos incluidos. Si tenemos en cuenta que un kilogramo de trigo suele contener

aproximadamente 1.200 granos[1], podemos hacernos una idea más precisa de la cantidad de trigo de la que estamos hablando:

$$\frac{18.446.744.073.709.551.615\ granos}{1.200\ granos/kg} = 15.372.286.728.091.293\ kg$$

que, si damos como exacta la estimación de la producción mundial de trigo para la cosecha 2014-2015, se necesitarían ¡22.053,83 años! para producir tal cantidad de trigo. Y es que ya decía Albert Einstein: *el mayor problema de la humanidad es que no entiende la función exponencial.*

En los últimos años, estamos asistiendo precisamente a un crecimiento exponencial de la información disponible online. El abaratamiento de los equipos informáticos, el aumento de su capacidad de cómputo sobre todo con la aparición de la computación paralela masiva, la cada vez mayor capacidad y menor coste de las unidades de almacenamiento -se ha pasado en sólo dos décadas de un coste de 700€ por gigabyte a unos cuatro o cinco céntimos-, la conexión de casi cualquier tipo de dispositivo a Internet, y el constante aumento de la velocidad ofrecida por las redes de comunicaciones, han hecho posible este crecimiento.

En la actualidad se estima que hay en torno a 1.000 millones de sitios web [2] y una cantidad de información de 4,4 zettabytes (1 zettabyte = 10^{21} bytes), esperándose que en el año 2020 se alcancen los 44 zettabytes [31] con el impulso del Internet de las Cosas [7, 27], que conectará a la red muchos objetos cotidianos que se convertirán en nuevas fuentes de datos que también se almacenarán en la red.

Recientemente ha surgido con fuerza el término *big data* -que podríamos traducir por datos masivos o datos a gran escala- para referirse a la acumulación de grandes (enormes) cantidades de datos y también a los nuevos procedimientos usados para su procesamiento, que consisten básicamente en la búsqueda de patrones repetitivos o correlaciones dentro de esos datos. Las tecnologías de *big data* encuentran aplicación en un creciente número de ámbitos y se espera que contribuyan de manera sustancial al crecimiento económico, a través de una mejora de la productividad tanto de las empresas como de las administraciones públicas [44]. Estas tecnologías están así llamadas a cambiar la forma de hacer negocios, de gestionar recursos y, también, la forma de hacer investigación [16]. Se habla ya de un cuarto paradigma científico [29] para referirse a la ciencia intensiva en datos (*data-intensive science*). Hace miles de años, los primeros científicos describían fenómenos naturales en base a evidencias empíricas, razón por la que la ciencia de esa época la denominamos **ciencia empírica**, lo que constituye el primer paradigma. Hace cientos de años, surgió la **ciencia teórica** como el segundo paradigma, del

[1] Se pueden encontrar diferentes estimaciones, pero cualquiera de ellas da lugar a un número enorme de kilogramos de trigo.

[2] Cifra tomada del sitio web *Internet Live Stats*: http://www.internetlivestats.com

que son ejemplos muy representativos las leyes del movimiento de Newton o las leyes de Kepler. No obstante, el análisis teórico se reveló incapaz de abordar muchos fenómenos o problemas complejos, razón por la que se recurrió a simulaciones científicas, dando así lugar a la **ciencia computacional** o tercer paradigma científico. Durante buena parte del siglo XX, los datos procedentes de observaciones o experimentos científicos se almacenaron en papel o en algunos dispositivos magnéticos que, por su formato o por su deterioro, hoy son ilegibles. Por otra parte, los datos que produjeron pequeños grupos de investigación suelen ser inaccesibles, siendo muy probable que se tiren cuando termina el proyecto de investigación o cuando los investigadores abordan una nueva línea. Las posibilidades que ofrecen las TIC para almacenar, acceder y compartir grandes cantidades de datos abren la puerta a la **ciencia intensiva en datos**, que constituye el cuarto paradigma científico. El cuarto paradigma está enfocado a cómo la ciencia puede avanzar mediante la compartición de grandes cantidades de datos disponibles online, incluso entre grupos multidisciplinares. Las tecnologías de *big data* están estrechamente relacionadas con lo que se ha denominado *e-Science* [30], o ciencia intensiva computacionalmente que suele emplear sistemas de procesamiento distribuido y un conjunto de herramientas, tecnologías y plataformas que permiten el almacenamiento, compartición y análisis de cantidades muy grandes de datos. En estos entornos se incluyen también herramientas de minería, visualización y exploración de datos, así como de comunicación y de diseminación de resultados. Ejemplos de e-Science son la física de partículas, la bioinformática, las ciencias de la tierra o las simulaciones sociales.

Los términos *big data* y *big data analytics* se refieren, respectivamente, a conjuntos muy grandes de datos y a técnicas de análisis que, por su tamaño (desde terabytes a exabytes) y por su complejidad (derivada de la diversidad de su procedencia y de sus formatos) requiere de estrategias especializadas de almacenamiento, gestión, análisis, y visualizacion. La nueva situación deriva no sólo del gran volumen de datos, sino también de la velocidad con la que se crean -se estima que el 90 % de los datos que existían en el mundo en 2012 se habían creado en los dos años anteriores [28]- y de la amplia variedad de formatos y fuentes. Los modelos tradicionales de procesamiento y almacenamiento se ven superados por el volumen, por la complejidad y también por el dinamismo de estos datos. Para hacerlo todavía más difícil, una parte importante de estos datos carecen de toda definición formal o semántica -son sólo bytes-, ya sea por la alta velocidad o por los mecanismos con los que se generan. Las características del Big Data se pueden resumir así en lo que se conoce como las tres Vs (Volumen, Velocidad y Variedad) [37].

Además de la búsqueda de procedimientos e infraestructuras que permitan procesar este volumen ingente de datos, existe también un amplio debate sobre si todos los datos disponibles son útiles o si, por el contrario, comenzamos a sentir un *síndrome de Diógenes* no sólo por el almacenamiento de información inútil, sino también por la pretensión de construir infraestructu-

ras y desarrollar procedimientos para procesarlos. Con objeto de evitar esta situación, también conocida con el nombre de *fat data*, se empieza a hablar de asociar a la definición de *big data* otras dos Vs [25]:

- **Valor**. Sólo se considera útil el 22 % de la información existente en el universo digital, si bien se estima que sólo se analiza el 5 %, lo que dejaría una cantidad enorme de información sin explorar. Se espera que la cantidad de datos útiles se vea incrementada por el Internet de las Cosas, estimándose que en 2020 más del 35 % de todos los datos serán útiles.
- **Veracidad**. Con el *big data* se pasará de un entorno de pocos datos fiables, en los que los errores del análisis venían principalmente motivados por la escasez de la muestra, a un entorno donde la muestra es muy grande pero el principal problema será la falta de fiabilidad de los datos. Con datos procedentes de tan variados orígenes y en tan variados formatos, cabe esperar que la calidad de la información se resentirá.

En resumen, la cantidad de información es ciertamente ingente pero existe un amplio consenso en considerar que no toda vale y, en consecuencia, no toda aportará conocimiento. Al respecto, se plantea el importante reto de filtrar de entre toda esta enorme cantidad de información aquella que aporte valor. Según Andreas Weigend, antiguo científico de Amazon y profesor en varias universidades americanas, los datos son el nuevo petróleo y no sólo por su valor económico, sino también porque es necesario refinarlos y depurarlos para que tengan valor.

3. Procedencia y procesamiento de la información online

La información disponible proviene de fuentes muy diversas. Una parte son datos que se recogen de nuestra actividad cotidiana: llamadas telefónicas, transacciones bancarias, pagos con tarjetas, búsquedas de información, etc. Se recogen también datos sobre movimientos a través de los dispositivos GPS instalados en nuestros teléfonos móviles. Los medios digitales de audio y vídeo son también una fuente importante de datos. Otros datos los generamos voluntariamente nosotros cuando publicamos entradas en blogs, ponemos imágenes o vídeos en YouTube, o enviamos mensajes a través de redes sociales como Facebook o Twitter.

La ciencia genera también una cantidad importante de datos, especialmente en campos como la genómica, la física de partículas, la biología, las ciencias medioambientales, la astronomía o la metereología. Las grandes infraestructuras científicas como los grandes telescopios, los colisionadores de partículas, etc., son fuentes de una gran cantidad de datos. Sólo por citar

algún ejemplo, se estima que el acelerador de partículas LHC (*Large Hadron Collider*) puede generar 60 terabytes de datos en un solo día [12].

Otra cantidad importante de datos procede de sensores que monitorizan objetos, y capturan y comunican esa información a través de la red. El abaratamiento de la electrónica y la consolidación del Internet de las Cosas hará posible la inclusión de sensores en casi cualquier objeto cotidiano que, de esta manera, se convertirá también en una fuente de datos que se almacenarán en la red. Se estima que en 2020 habrá en el mundo 26.000 millones de dispositivos conectados [26] que generarán el 40 % de los datos creados [15]. Es previsible que en los próximos años las fábricas, las ciudades y el transporte se transformen en fábricas inteligentes, ciudades inteligentes y transporte inteligente, que integrarán un sinfín de sensores y software para optimizar procesos, ahorrar energía y ofrecer mejores servicios.

No obstante, cabe recordar que el almacenamiento de datos no es algo nuevo, toda vez que tanto las empresas como las distintas administraciones han estado almacenando datos de usuarios, clientes, proveedores, etc. desde hace mucho tiempo. La principal novedad surge del hecho de que hasta no hace mucho tiempo estos datos sólo se almacenaban, ya que su procesamiento era muy costoso en tiempo y en esfuerzo. El constante abaratamiento de los dispositivos de almacenamiento permite mantener online esta enorme cantidad de información y, por otra parte, el importante aumento de la capacidad de cómputo con la aparición de la computación paralela masiva hace posible su procesamiento. Atendiendo a la dificultad que supone su procesamiento, los datos se clasifican en tres grupos:

- **Estructurados**, son datos con un tipo formalmente definido y con una semántica clara, lo que facilita su procesamiento. El grueso de estos datos proviene de bases de datos relacionales, típicamente bases de datos SQL.

- **No estructurados**, estos datos son únicamente una secuencia de bits sin mayor significado y cuya semántica no está definida. Por esta razón, el procesamiento de estos datos se hace más complejo.

- **Semiestructurados**, ocupan algún lugar intermedio entre los dos tipos anteriores. Se incluyen en este apartado los datos almacenados en bases de datos NoSQL. Los datos etiquetados con XML se suelen incluir también en este apartado, toda vez que, sin poder considerarse no estructurados, suelen carecer de un tipo formalmente definido como el que tienen los datos almacenados en bases de datos convencionales.

La mayor parte de los datos que se generan pertenecen a las dos últimas categorías, estimándose que el 80-90 % de los datos generados en 2013 son no estructurados o semiestructurados [31]

El procesamiento de la información disponible online tiene su origen en los bien conocidos sistemas de gestión de bases de datos relacionales (RDBMS), que almacenan principalmente información estructurada [17]. Las técnicas de análisis empleadas, que se popularizaron en los años 90, están basadas principalmente en el análisis estadístico y en técnicas de minería de datos desarrolladas en los años 80. Consultas a las bases de datos, procesamiento analítico online y herramientas para producir diferentes tipos de gráficos permiten explorar características importantes de los datos. A principios de los 2000, el rápido crecimiento de Internet y del uso de la Web permitió recoger una gran cantidad de datos, cuyo procesamiento posterior ofrecía grandes posibilidades y oportunidades. Los primeros servidores Web (Web 1.0), entre los que cabe destacar por su relevancia los buscadores (Google, Yahoo, etc.) y grandes empresas pioneras en el comercio electrónico (Amazon, eBay, etc.), ofrecían a las empresas la posibilidad de presentar online sus negocios y de interactuar directamente con los clientes. La información registrada de tales interacciones, recogida a través de *cookies* y ficheros de registro de actividad de los servidores (ficheros *log*) pronto se reveló muy valiosa para entender los gustos y necesidades de los clientes, así como para identificar nuevas posibilidades de negocio. Posteriormente, el éxito de la Web 2.0 [48] dio lugar al desarrollo de numerosas aplicaciones como foros, blogs, redes sociales, mundos virtuales, juegos sociales, etc., que potenció el almacenamiento en la red de contenidos generados por los usuarios. Es bien conocido que las aplicaciones de la Web 2.0 pueden reunir un gran volumen de información y opiniones sobre casi cualquier tema, generado por una población muy diversa y potencialmente muy numerosa. En la actualidad, se está trabajando en el desarrollo de diferentes técnicas para procesar esta información con el objetivo de entender mejor las opiniones, necesidades y gustos de los usuarios. Más recientemente, el rápido crecimiento en el número de dispositivos móviles conectados a Internet -en 2011 el número de teléfonos móviles (*smartphones*) y tabletas superó al número de ordenadores portátiles- y el ecosistema de aplicaciones disponibles para descargar en dichos dispositivos está transformando diferentes facetas de la sociedad, desde la educación hasta el cuidado de la salud y desde el entretenimiento a la forma de relacionarse con la administración. En los próximos años, la previsible consolidación del Internet de las Cosas acentuará esta tendencia. El análisis de la cuantiosa cantidad de datos que es previsible que se genere contribuirá también de forma sustancial a este proceso de predicción de los gustos y necesidades de los usuarios.

Por último, cabe recordar que no todos estos datos tendrán el mismo valor. Según la consultora IDC, si a los datos se le asocian medidas como facilidad de acceso; disponibilidad en tiempo real; importancia o número de personas, parte de una organización o número de clientes a los que se refieren; o trascendencia de las conclusiones que se extraen de su análisis, sólo nos quedaríamos con algo más del 6 % del volumen de datos considerados

útiles, si bien se prevé que este porcentaje se eleve al 11 % en 2020. Esta cantidad se espera que crezca a medida que la información existente en la red se etiquete con metadatos -incluyendo metadatos sobre los metadatos-, lo que posibilitará su procesamiento.

4. Aprovechamiento de la información útil contenida en la red

Esta enorme y creciente cantidad de información disponible, unida al rápido desarrollo de dispositivos móviles y de nuevas redes de comunicaciones, ha venido dibujando un escenario caracterizado por la posibilidad que se le presenta a los usuarios de acceder, en cualquier momento y lugar, a una cantidad muy amplia de contenidos y servicios, con un espacio de posibilidades muy amplio: entretenimiento, noticias, contenidos educativos, teleadministración, comercio electrónico, asistencia sanitaria, etc. Este nuevo escenario, sin duda lleno de ventajas y oportunidades innegables, puso enseguida de manifiesto una situación de *saturación de información*, caracterizada por la dificultad -cuando no imposibilidad- que muchos usuarios experimentan para acceder a la información que necesitan, aún siendo conscientes de que dicha información está disponible de alguna forma en la red. La búsqueda de información se convertía así en una tarea tediosa, complicando sobremanera la toma de decisiones o, simplemente, el mantenerse informado sobre un tema determinado. Este problema, aún no satisfactoriamente resuelto, plantea enormes desafíos también a las fuerzas de mercado que esperarían obtener mayores beneficios de los contenidos y servicios que desarrollan, ya que la presencia de éstos pasa con demasiada frecuencia desapercibida para muchos usuarios, al diluirse casi por completo entre una marea inmensa y desorganizada de productos de similar naturaleza.

Hace sólo unas pocas décadas, la búsqueda de contenidos se limitaba a unos pocos canales de TV, a un puñado de periódicos, a unos cuantos mega portales de dirección conocida, y a unas cuantas emisoras de radio. El vertiginoso aumento de la oferta puso de relieve el problema de saturación de información, que se detectó por primera vez en el acceso a Internet desde ordenadores personales, al constatarse la dificultad que experimentaban los usuarios para acceder a contenidos o servicios de su interés a medida que crecía el volumen de recursos disponibles en la red. Esta situación, que suponía sin duda un freno en el empleo y adopción de estas nuevas tecnologías, se atacó desde mediados de los 90 con los denominados *buscadores* (Google, Altavista, Yahoo, Lycos, Bing, etc.), cuyo éxito e implantación es de todos conocido. Estos sistemas adoptaron diversos enfoques y algoritmos para facilitar la localización de recursos, si bien con un denominador común de que la iniciativa siempre la toma el usuario, que ha de conocer la existencia de estas herramientas y saber manejarlas, siendo capaz de introducir cadenas

de texto que describan con la mayor exactitud posible los contenidos que se buscan en cada momento. Desafortunadamente, el planteamiento de estos buscadores no es exportable a todos los escenarios, toda vez que muchos dispositivos de acceso no ofrecen interfaces adecuadas para introducir cadenas de texto (por ejemplo, los receptores de TV digital que han de manejarse a través de un mando a distancia). Por otro lado, es notorio que muchos de los nuevos dispositivos llevan asociados unos hábitos de uso radicalmente diferentes a los del PC, lo que, unido a la falta de formación tecnológica de muchos de sus usuarios, sugiere explorar nuevas vías que no presupongan mucha iniciativa por parte de éstos.

Más adelante surgió el consumo de contenidos sociales, cuya principal característica es que los usuarios consideran de interés compartir contenidos (imagen, audio y vídeo) u opiniones sobre contenidos con sus amigos, compañeros de trabajo o estudios y/o familiares. Este fenómeno ha dado lugar no sólo a un incremento notable en la producción de contenidos, sino también a su consumo. El éxito de compañías como Facebook, Twitter, etc. es una buena prueba de ello.

Otra forma de consumir contenidos es oportunista. En muchas ocasiones, un usuario no sabe qué contenido consumir, o siquiera si consumir algún contenido. En esta situación, si el usuario encuentra de forma fortuita o se le presenta algún contenido de su agrado, es posible que decida consumirlo. Este hecho entronca con la necesidad de desarrollar aplicaciones con mayores niveles de automatización y proactividad, capaces de guiar al usuario en el proceso de selección de contenidos y servicios de su interés. Con este objetivo, surgieron en los años 90 los *sistemas recomendadores* [10, 40], que implementan métodos desarrollados en el seno de la Inteligencia Artificial para seleccionar, de forma automática y personalizada, aquellos contenidos y servicios que mejor se adapten a las preferencias o necesidades de cada usuario. Reducen así lo que podría ser una decisión compleja, ya sea por las múltiples posibilidades existentes o por la dificultad para encontrar todas las opciones, a una elección entre un conjunto reducido de alternativas (recomendaciones). La calidad de estos sistemas, entendida ésta como el grado de acierto en las recomendaciones presentadas al usuario, depende fuertemente de:

- la exactitud con la que se puedan capturar los intereses, preferencias y necesidades de los usuarios, lo que se denomina *perfil de usuario*.
- los mecanismos disponibles para poder identificar el *contexto* en que se ubica el usuario.
- el etiquetado disponible para contenidos y servicios a recomendar (*metadatos*).
- las técnicas de filtrado encargadas de medir el grado de adecuación de cada contenido o servicio a un perfil de usuario concreto.

En las siguientes secciones haremos una presentación de los sistemas recomendadores y de sus dominios de aplicación más característicos, dejando para la sección de conclusiones una breve discusión de cómo las tecnologías de *big data* pueden mejorar la calidad de estas recomendaciones.

5. Sistemas recomendadores

El objetivo de los sistemas recomendadores (del inglés *recommender systems*) es seleccionar, de forma automática y proactiva, aquellos contenidos que pueden resultar de interés para un usuario o grupo de usuarios en un contexto concreto, entre una cantidad ingente de opciones disponibles. Además de impedir que puedan pasar desapercibidos contenidos de interés para un usuario o grupo de usuarios, los sistemas recomendadores permiten reducir un amplísimo espectro de contenidos (*big data*) en un conjunto más reducido (*small data*), más manejable para la toma de decisiones. Los sistemas recomendadores son también de utilidad en otros entornos en los que aún siendo pocas las opciones, la elección es también compleja por no disponer de la información suficiente sobre las mismas.

El funcionamiento de un sistema de recomendación se basa en establecer alguna medida de la adecuación de un contenido a los intereses, preferencias o necesidades de un usuario o grupo de usuarios. La recomendación se elabora así seleccionando un conjunto reducido de opciones para las cuales la adecuación estimada es mayor. De esta forma, los sistemas de recomendación permiten construir servicios de información personalizados, o adaptados a unos intereses concretos. El desarrollo de servicios de información personalizados involucra múltiples áreas de las tecnologías de la información y la comunicaciones, así como un notable cuerpo de legislación. A continuación, se describirán con mayor detalle estos aspectos.

5.1. Base tecnológica

La elaboración de recomendaciones personalizadas requiere la caracterización o modelado de los intereses del usuario o usuarios hacia los que se dirige la recomendación, la identificación del contexto, la caracterización de los contenidos a recomendar, y el diseño de un algoritmo que permita estimar la adecuación de cada contenido a los intereses del usuario o usuarios objeto de la recomendación en el contexto en el que se encuentran. En los siguientes apartados se describen cada uno de estos aspectos que conforman la base tecnológica de los sistemas recomendadores.

5.1.1. Modelado de usuarios

Para la identificación de contenidos adecuados para un usuario es preciso conocer sus intereses, preferencias y necesidades. Esta información constitu-

ye su *perfil*. En la literatura especializada [18] se pueden encontrar soluciones muy variadas para modelar y capturar el perfil de un usuario, incluyendo historiales de consumo, datos demográficos (edad, sexo, estado civil, formación, etc.), redes conceptuales semánticas o asociativas, ontologías, redes neuronales, árboles de decisión, etc. Igualmente, existen numerosos estándares *de jure* o *de facto* dependientes del dominio de aplicación para normalizar los datos que se almacenarán en los perfiles: LIP [3] (*Learner Information Packaging*) para el seguimiento de actividades educativas, *TV-Anytime* [55] para monitorizar el visionado de programas de televisión, openEHR [4] (*open Electronic Health Records*) para mantener registros de aplicaciones relacionadas con la salud, etc.

Independientemente de la representación y los datos almacenados, la inicialización de los perfiles de usuario puede realizarse de forma manual, pidiendo a cada usuario que introduzca explícitamente una descripción de sus intereses a través de cualesquiera interfaces, o automáticamente, extrayendo información a partir de las interacciones pasadas del usuario con otros sistemas (un ejemplo común es obtener información a través de historiales de navegación por la web o, más recientemente, a partir de información almacenada en redes sociales como Facebook o Twitter). Otras veces, se opta por un enfoque semiautomático a partir de *estereotipos*, definiendo categorías genéricas para clasificar a los usuarios y seleccionando una de ellas como perfil inicial una vez que el usuario ha introducido unos pocos datos relevantes.

Con independencia de cual sea la estrategia de inicialización, es necesario tener en cuenta que los perfiles de usuario no son estáticos, toda vez que las preferencias e intereses de los usuarios suelen variar con el tiempo, a la vez que aparecen nuevos contenidos y desaparecen los más viejos. En consecuencia, un sistema de recomendación debe implementar mecanismos que permitan mantener actualizados los perfiles de usuario, capturando nuevos datos (*realimentación de relevancia*) y descartando información obsoleta (*olvido gradual*).

En lo que se refiere a la realimentación de relevancia, existen principalmente dos tipos de mecanismos:

- **implícitos**, que tratan de inferir información a partir de la interacción del usuario. Por ejemplo, se puede estimar el interés que un usuario tiene en un programa de televisión midiendo si lo ve todo o sólo una parte, el número de veces que lo ve, etc. Por no requerir la participación directa del usuario, estos mecanismos no son invasivos, si bien pueden ser muy poco precisos. Por ejemplo, el hecho de que un receptor de TV tenga sintonizado el mismo canal durante la proyección de una película no implica que el usuario la haya visto en su totalidad, ya que pudo

[3] http://www.imsglobal.org/profiles/index.html
[4] http://www.openehr.org/

dormirse o simplemente dejar el televisor encendido, ni tampoco que le haya gustado.

- **explícitos**, que tienen en común solicitar a los usuarios una valoración de su interés en algunos contenidos. Ciertamente, los mecanismos explícitos son más precisos, si bien requieren la participación directa del usuario, por lo que deben dosificarse convenientemente para no causar rechazo.

En cuanto al olvido gradual, el planteamiento más habitual consiste en medir la obsolescencia de los datos almacenados en los perfiles como una función de tiempo, que generalmente dependerá del tipo de contenido. Otra limitación importante de muchos sistemas recomendadores actuales proviene de ignorar que el interés que un usuario tiene por un tipo de producto puede disminuir de forma considerable después de su adquisición y durante un tiempo variable que depende del tipo de producto y también del perfil del usuario. Por ejemplo, cuando una persona compra un coche su interés por los coches nuevos suele ser muy bajo después de la adquisición y va en aumento a medida que se va agotando la vida útil del coche, si bien es cierto que el concepto de vida útil del coche es dependiente del perfil del usuario (es bien conocido que hay usuarios que cambian el coche sólo cuando el que tienen ya no funciona o tiene averías muy frecuentes, y otros que lo cambian cada pocos años para no sufrir averías y para disfrutar siempre de un coche nuevo). Ciertamente, esta limitación se antoja de poca importancia si el objeto de la recomendación son contenidos audiovisuales o un libro, pero si el objetivo de la recomendación es cualquier tipo de producto, solucionar este problema parece una cuestión prioritaria si se quiere que las recomendaciones sean útiles. Por último, resulta difícil para un sistema recomendador elaborar recomendaciones de calidad para un usuario singular, es decir, para un usuario con gustos poco frecuentes. Por ejemplo, volviendo al ejemplo del coche, si el usuario fuera un coleccionista de coches, agradecería recomendaciones sobre coches incluso después de haber comprado uno.

5.1.2. Sensibilidad al contexto

Los intereses, gustos o necesidades de los usuarios pueden variar con el contexto en el que se encuentran. Por ejemplo, si el objeto de la recomendación es un contenido audiovisual, el formato y la longitud/duración del contenido dependerá del tiempo de que dispone el usuario para ver el contenido, del tamaño y resolución del dispositivo de acceso, etc. Por otra parte, los temas que pueden interesar al usuario en un momento dado (noticias, entretenimiento, restaurante, consejo médico, etc.) suelen depender también de su localización geográfica, hora del día, rutinas diarias, etc.

El objeto de la sensibilidad al contexto en los sistemas recomendadores es recabar información sobre la situación física y social en que se ubican

los dispositivos que utiliza el usuario para tratar de maximizar la calidad de las recomendaciones que se le presentan. El conocimiento del contexto se sumará a la información contenida en el perfil para guiar la selección de los contenidos de interés que se incluirán en la recomendación [2].

Históricamente, la primera posibilidad explorada en la sensibilidad al contexto fue la ubicación geográfica, desarrollándose aplicaciones dependientes de la localización para dispositivos móviles, capaces de mostrar diferentes piezas de información siguiendo los movimientos de los usuarios en interiores (por ejemplo, museos) o exteriores (por ejemplo, rutas turísticas personalizadas) [8]. Posteriormente, se fueron añadiendo otras como la temática (por ejemplo, inferida en base a palabras o imágenes representadas en la pantalla), infraestructura (por ejemplo, dispositivos de entrada/salida disponibles o recursos de comunicaciones existentes en el entorno), y condiciones físicas (por ejemplo, ruido ambiente o iluminación). Los trabajos más recientes en *computación emocional* (del inglés *affective computing*) [59], unidos a teorías de las emociones y la cognición, posibilitan tener en cuenta también el estado anímico del usuario (humor, estrés, etc.).

No obstante, a pesar de los avances citados, hasta la fecha la investigación en sensibilidad al contexto se ha centrado más en su identificación que en su explotación, de modo que la integración de estas tecnologías en sistemas recomendadores es un aspecto en gran medida no resuelto.

5.1.3. Caracterización de contenidos

La automatización de la selección de contenidos hace necesario disponer de una caracterización precisa de los mismos, es decir, de un conjunto de *metadatos* que los describan con la precisión suficiente. Se han elaborado diferentes estándares de especificación de metadatos (MPEG7 [43], TV-Anytime [55], SCORM [1], etc.), que tienen como característica común estar basados en un mecanismo controlado de clasificación y etiquetado de productos para determinados ámbitos, en el que la responsabilidad de la inclusión de metadatos recae exclusivamente en el creador o proveedor del producto.

Con el despliegue de los primeros sistemas recomendadores surgieron dudas sobre la conveniencia de adoptar este modelo, habida cuenta de los problemas de índole práctico y social que su uso conlleva. De índole práctico porque en muchas ocasiones los metadatos que acompañan a los contenidos son muy escasos o inexistentes, y en algunos dominios de aplicación no parece fácil motivar a los fabricantes o proveedores de contenidos a que proporcionen descripciones más precisas y completas. De índole social porque las normas imponen un vocabulario controlado, casi siempre fijado a priori, que sólo permite expresar la perspectiva de los creadores o de los proveedores de los productos y no de sus consumidores. Estas limitaciones se pueden paliar en gran medida incorporando la filosofía de la Web 2.0 para posibilitar la participación de los usuarios en la catalogación de los contenidos. Otra

posibilidad complementaria en la que se ha estado trabajando en los últimos años es la extracción automática de metadatos de contenidos en formato texto, audio y vídeo.

5.1.4. Algoritmos de filtrado

En última instancia, la elaboración de recomendaciones se consigue cotejando de alguna forma la información almacenada en el perfil de un usuario, incluyendo el contexto actual, con los metadatos que caracterizan los contenidos susceptibles de recomendación. Una primera clasificación ampliamente aceptada divide los sistemas recomendadores en dos grupos:

- **Basados en memoria**: Estos métodos emplean una matriz con las valoraciones de los usuarios para cada producto a recomendar y definen una métrica de similitud que permite calcular la distancia entre dos usuarios o entre dos productos. Cada vez que un usuario hace una valoración de un producto (implícita o explícita), este dato se introduce en la matriz. Este tipo de algoritmos tiene como principal problema el aumento del coste computacional que se produce a medida que aumentan el número de usuarios y/o el número de productos susceptibles de recomendación.
- **Basados en modelo**: Estos métodos usan las valoraciones de los usuarios para crear un modelo que permite generar las recomendaciones. Ejemplos de métodos basados en modelo son los clasificadores bayesianos, las redes neuronales, los sistemas *fuzzy* o los algoritmos genéticos. Es importante destacar que cada vez que se obtiene una nueva valoración de usuario es necesario cambiar o actualizar el modelo.

En función de la estrategia seguida para estimar la adecuación de contenidos a perfiles, se pueden encontrar en la literatura cuatro tipos de algoritmos de filtrado:

1. **Demográfico:** [35] Emplea las características personales de los usuarios (edad, sexo, estado civil, estudios, profesión, nacionalidad, aficiones, etc.) para valorar la adecuación de un determinado producto a un usuario. Por su naturaleza, este método modela las preferencias de usuario como un vector de características demográficas y recurre a la técnica de estereotipos para inicializar sus perfiles. Todos aquellos usuarios que pertenezcan al mismo estereotipo recibirán las mismas recomendaciones.

   Por considerar sólo características demográficas, este filtrado suele producir recomendaciones generales e imprecisas. Además, no es fácil incorporar en los perfiles los cambios en las preferencias de los usuarios. No obstante, esta técnica puede ser una estrategia útil si se combina con otros métodos.

2. **Basado en contenidos:** [51, 39] Del inglés *content-based*, consiste en recomendar contenidos *similares* a los que ya gustaron al usuario en el pasado. Se trata de una estrategia muy precisa y relativamente sencilla de implementar y ejecutar. No obstante, sufre de *sobreespecialización*, toda vez que tiende a recomendar productos excesivamente similares a los que el usuario ya conoce, e incluso demasiado similares entre sí. Las soluciones existentes en la literatura únicamente combaten el excesivo parecido entre los productos recomendados, para así tratar de evitar que el usuario rechace toda la recomendación.

3. **Colaborativo:** [53, 3, 14] No evalúa únicamente el perfil del usuario objeto de las recomendaciones, sino también los perfiles de otros usuarios con intereses similares (lo que se ha dado en llamar *vecinos* de dicho usuario). En la literatura se pueden encontrar dos técnicas diferentes dentro de este tipo de estrategias [52]:

   - **Basado en usuario:** Recomienda a un usuario contenidos que han interesado a sus vecinos. Se considera que dos usuarios tienen intereses similares (son vecinos) si han clasificado los mismos productos en sus perfiles y les han asignado valoraciones similares. Otra posibilidad para identificar vecinos es mediante el empleo de alguna métrica de similitud entre perfiles de usuario.
   - **Basado en item:** Recomienda a un usuario un producto si es similar a los definidos en su perfil personal. Se considera que dos productos son similares si los usuarios que han valorado uno de ellos, tienden también a valorar el otro con una calificación similar a la de éste. Esta técnica suele dar mejores resultados que la anterior cuando el número de productos disponibles en el sistema es inferior que el número de usuarios.

   Por su propia naturaleza, este enfoque resuelve la falta de diversidad de los métodos basados en contenido, pero tiene un importante problema de dispersión (*sparsity problem*), consistente en que a medida que crece el número de productos disponibles se vuelve más improbable encontrar usuarios con valoraciones similares para unos mismos productos. Para atacar este problema, en métodos basados en memoria se recurre a técnicas de factorización de matrices [34, 41, 42]. Otro problema importante de los algoritmos de filtrado colaborativo surge de las denominadas *usuarios singulares* (del inglés *grey-sheep problem*) o usuarios cuyas preferencias son distintas a las del resto de usuarios, lo que dificulta o imposibilita la creación de su vecindario. Por otra parte, este tipo de algoritmos tienen también el problema de los *falsos vecinos* [57, 58] (del inglés *faked neighbours*), caracterizado porque los vecinos de un usuario dado para un tipo de productos pueden no serlo

para otro tipo de productos. Así, es frecuente que dos personas que tienen los mismos gustos para la lectura no lo tengan, por ejemplo, para los coches.

Otro inconveniente importante de este tipo de algoritmos de filtrado es su mayor latencia derivada de su también mayor coste computacional.

4. **Híbridos:** Para neutralizar las desventajas de las técnicas de filtrado anteriores y aunar sus ventajas, muchos autores han propuesto el empleo de *enfoques híbridos* [13], siendo una posibilidad común la de recomendar contenidos o servicios similares a los que figuran en el perfil del usuario, como en las estrategias basadas en contenido, pero considerando dos recursos similares si los usuarios que muestran interés en uno tienden a interesarse también en el otro (*vecindarios de recursos*). Estos modelos permiten superar la sobreespecialización de las técnicas de filtrado basadas en contenido y, en cierta medida, los efectos del *sparsity problem*.

Un problema común de todos los algoritmos de filtrado es el arranque inicial (del inglés, *cold start*) que se manifiesta por no ser posible hacer recomendaciones de calidad al carecer inicialmente de la información necesaria. Ciertamente, cuando se pone en marcha el sistema recomendador no se dispone de las valoraciones necesarias para poder construir una matriz o para desarrollar un modelo fiable que permita generar las recomendaciones. Existen tres tipos de problemas de arranque: *nueva comunidad*, *nuevo usuario* o *nuevo item*. El problema de nueva comunidad [36] se refiere a la dificultad de elaborar recomendaciones de calidad en el arranque de un sistema recomendador por no disponer de valoraciones. Existen dos estrategias para atacar este problema: animar a los usuarios a que introduzcan valoraciones y emplear un filtrado colaborativo cuando existan valoraciones suficientes. El problema de nuevo item [49] surge cuando aparece un nuevo producto susceptible de recomendación. Al no disponer inicialmente de valoraciones para él, es improbable que se recomiende, por lo que puede pasar mucho tiempo inadvertido para una gran parte de los usuarios del sistema recomendador. Una solución común a este problema es disponer de alguna estrategia para incentivar a un grupo de usuarios para que hagan valoraciones sobre los nuevos productos que llegan al sistema. Por último, una de las mayores dificultades que tienen los sistemas recomendadores es el problema del nuevo usuario [56], ya que, la información sobre sus gustos o preferencias es escasa. Las principales soluciones a este problema fueron comentadas en la sección 5.1.1.

Otro problema común al que se enfrentan los algoritmos de filtrado es su escalabilidad, toda vez que, a medida que aumentan los usuarios y/o el número de productos, aumenta también el coste computacional y, por tanto, su latencia. Ciertamente, tiempos de latencia por encima de cierto umbral

hacen que la experiencia de usuario sea pobre.

Como una cuestión ortogonal al filtrado, los primeros sistemas de personalización se apoyaron en técnicas de similitud sintáctica, basadas en comparar cadenas o patrones de texto. Toda vez que este planteamiento depara una mayor similitud entre, por ejemplo, "*caballo*" y "*caballa*" que entre "*caballo*" y "*yegua*", salta a la vista que de este modo resulta imposible descubrir relaciones semánticas entre preferencias de usuario y recursos disponibles, limitando enormemente la calidad de la personalización. Una solución a esta limitación es la aplicación de las técnicas desarrolladas en el seno de la *web semántica* [5], que habilitan procesos de razonamiento capaces de penetrar en el significado de palabras y frases para relacionar eficazmente unos recursos con otros. El razonamiento se realiza sobre *metadatos* que describen contenidos multimedia, ya sea documentos de texto, imágenes, material audiovisual o programas interactivos. En este proceso se han usado *ontologías* como un método para representar formalmente el conocimiento en forma de conceptos y relaciones entre conceptos.

5.2. Evaluación de la calidad de los sistemas recomendadores

Para poder realizar comparaciones entre diversas soluciones a un mismo problema, se han definido métricas y también entornos de evaluación de sistemas recomendadores [20, 11]. A continuación se describen las métricas más usadas.

5.2.1. Calidad de las predicciones

Su objetivo es medir la exactitud de los resultados ofrecidos por un sistema recomendador. Las más usadas son la *media del error absoluto*, *error cuadrático medio*, y la *cobertura* o capacidad para hacer predicciones de un sistema recomendador.

5.2.2. Calidad del conjunto de recomendaciones

La calidad de experiencia que un usuario obtiene de un sistema recomendador no depende directamente de la exactitud del conjunto de posibles predicciones. Esta calidad de experiencia se incrementa cuando el usuario está de acuerdo con un subconjunto de las recomendaciones hechas por el sistema. Al respecto, las tres métricas de calidad más usadas son:

- *Precisión* o proporción de items relevantes del total de items recomendados.
- *Exhaustividad* (*Recall*) o proporción de items relevantes recomendados del total de items relevantes.
- *F1* o media armónica de precisión y recall.

5.2.3. Rango de la recomendación

Cuando el número de items recomendados no es pequeño, los usuarios conceden mayor importancia a los primeros items de la lista de recomendación. Así, los errores que se cometan en los primeros items son más serios que los que se cometan en los últimos items de la lista. El objetivo de estas medidas es cuantificar esta situación, siendo las más importantes:

- *Vida media* que supone un decremento exponencial en el interés del usuario cuando se mueve hacia abajo en la lista de recomendación.
- *Ganancia acumulada descontada* que supone un decremento logarítmico en el interés del usuario.

5.2.4. Novedad y diversidad

La novedad mide el grado de diferenciación entre los items recomendados y los items conocidos por un usuario. Por su parte, la diversidad mide el grado de diferenciación entre los items recomendados.

5.2.5. Estabilidad

La estabilidad en las predicciones y en las recomendaciones tiene incidencia en la confianza que el sistema recomendador le merece al usuario. Un sistema recomendador es estable si las predicciones que proporciona no cambian significativamente en un corto periodo de tiempo.

5.2.6. Confiabilidad

La confiabilidad de una predicción o de una recomendación es una medida de la seriedad con la que un usuario se debe tomar dicha predicción o recomendación. El objetivo de esta medida es cuantificar el hecho de que cuando un recomendador otorga una nota a una predicción, lo hace únicamente en función de los datos de que dispone. Ciertamente, no es lo mismo que los datos sean muchos o que sean pocos, toda vez que cabe esperar que la predicción sea mejor cuanto mayor sea el número de datos manejados para su elaboración.

6. Sistemas recomendadores y redes sociales

Ningún sistema recomendador podrá elaborar recomendaciones de calidad si no dispone de información suficiente y precisa, tanto de los productos a recomendar como de los gustos, preferencias y necesidades de los usuarios consumidores de estas recomendaciones. En consecuencia, el éxito de estos

sistemas estará fuertemente ligado a la capacidad de disponer de una información lo más completa posible sobre cada producto a recomendar, y a trabajar con versiones actualizadas de los perfiles de cada usuario.

En los últimos años se han desarrollado numerosos trabajos de investigación con el objetivo de poder extraer automáticamente metadatos de vídeos, audios y textos, lo que sin duda permite mejorar la descripción de los productos a recomendar. Por otra parte, el éxito de la conexión social en Internet ha producido una cantidad enorme de datos principalmente en redes sociales. Los usuarios han proporcionado todos estos datos de forma natural, conformando una información en bruto. La aplicación de técnicas de minería de datos permite obtener información útil, tanto para la descripción de productos como para los perfiles de usuario, lo que permite mejorar la calidad de las recomendaciones elaboradas por los sistemas de recomendación. Por otra parte, la naturalidad con la que los usuarios participan en las redes sociales permite adivinar que la información que de ellas se extraiga estará menos sesgada.

En este contexto, se están realizando numerosos trabajos con el objetivo de incorporar a los sistemas de recomendación información social: usuarios de confianza, usuarios no confiables, usuarios seguidos, usuarios seguidores, listas de amigos, correos, blogs y etiquetas. Esta información social permite aliviar también el problema de la dispersión (*sparsity*), inherente a las técnicas de filtrado basadas en modelo, toda vez que los usuarios conectados mediante una red de confianza o amistad exhiben una similitud notablemente mayor que los usuarios no conectados.

Los trabajos de investigación en el ámbito de la información social se han dirigido hacia tres objetivos: (i) mejorar la calidad de las predicciones y recomendaciones; (ii) construir nuevos sistemas de recomendación; y (iii) identificar las relaciones más significativas entre información social y procesos de recomendación colaborativos.

La confianza y la reputación son dos conceptos importantes para el empleo de información social en sistemas de recomendación. Al respecto, se han definido métricas de confianza que calculan la credibilidad de un usuario a partir de información explícita aportada por otros usuarios o a través de información implícita contenida en redes sociales. También se establecieron medidas de la reputación de un producto a través de las valoraciones hechas por usuarios o estudiando cómo los usuarios trabajan con esos productos.

Una parte importante de la información social la constituyen las etiquetas que los usuarios asocian a los productos. Con el término *folksonomía* se identifica el conjunto de triplas <usuario, producto, etiqueta> que se usan para la creación de sistemas de recomendación de etiquetas y para enriquecer los sistemas de recomendación. Por otra parte, la tendencia de los usuarios a etiquetar productos y a establecer enlaces sociales ha permitido mejorar también la calidad de las recomendaciones realizadas con técnicas de filtrado basadas en contenidos.

7. Recomendación grupal

Existen diferentes situaciones en las que un conjunto de usuarios realizan conjuntamente una actividad común. El objetivo de la recomendación grupal [32] es elaborar sugerencias para un grupo de usuarios a partir de los perfiles individuales de cada uno de ellos. Tiene interés en diferentes dominios de aplicación como la recomendación de películas, música, páginas web, eventos, destinos turísticos, etc. También tiene interés desde el punto de vista de un proveedor para identificar qué producto o servicio tendrá mejor aceptación para un grupo de usuarios dado.

Existen dos estrategias dominantes para la elaboración de recomendaciones grupales [4, 9]: (i) crear un pseudo-usuario que representa los intereses del grupo, para el que se elaboran las recomendaciones como si de un usuario individual se tratara; y (ii) elaborar las recomendaciones para cada usuario individual del grupo y fusionar o combinar los resultados obtenidos para elaborar las recomendaciones que se presentarán al grupo de usuarios.

La principal ventaja de la segunda estrategia es que se puede implementar como una extensión de los sistemas recomendadores clásicos, toda vez que es únicamente necesario añadir un mecanismo que permita agregar las predicciones realizadas para cada usuario individual y obtener una predicción para el grupo de usuarios, a partir de la cual se elaborarán las recomendaciones para el grupo. El principal inconveniente de esta estrategia es que, a diferencia de lo que ocurre con la recomendación individual, no se puede garantizar que las recomendaciones así elaboradas sean satisfactorias o suficientemente buenas para el grupo. Parece claro que, aunque la recomendación para un usuario individual maximiza la satisfacción de este usuario, cuando esta recomendación se usa para el grupo, la probabilidad de producir recomendaciones no satisfactorias es elevada. Aunque se propusieron algunas soluciones a este problema, esta estrategia de recomendación a grupos no se suele usar. [33]

Para implementar la primera estrategia se suele emplear una función de agregación con el objeto de elaborar una *valoración de consenso* para cada producto a recomendar, a partir de las valoraciones individuales de cada usuario que estima el recomendador para dicho producto. Existen varias posibilidades para construir esta función de agregación, siendo las siguientes las más usadas:

- **Promedio**: Todos los usuarios del grupo se consideran iguales y la valoración para el grupo de un producto se calcula como la media aritmética de las valoraciones individuales para ese producto de todos los miembros del grupo.

- **Menor rechazo**: La valoración para el grupo de un producto es el mínimo de las valoraciones individuales. De esta forma, se busca no

causar rechazo en ninguno de los miembros del grupo. Es equivalente a dar derecho de veto a los usuarios del grupo.

- **Promedio sin rechazo**: Trata de evitar el veto de un usuario del grupo si la valoración que hace de un producto cae por debajo de cierto umbral δ. Así, la valoración que se hace para el grupo de un producto es igual al mínimo de las valoraciones individuales, excluyendo aquellas que son menores o iguales que δ.

- **Mayor satisfacción**: La valoración que para el grupo se hace de un producto es igual al máximo de las valoraciones individuales de los miembros del grupo.

- **Dictadura**: Esta función de agregación busca primar a ciertos usuarios en ocasiones especiales (cumpleaños, etc.) y calcula la valoración para el grupo de un producto como la valoración que hace uno de los usuarios del grupo.

- **Votación**: En este caso, cada miembro del grupo vota y se seleccionan los productos que obtienen un mayor número de votos.

En el caso de emplear filtrado colaborativo, existen cuatro estrategias principales para fusionar la información de los usuarios que forman el grupo con la finalidad de obtener datos para el grupo que permitan realizar la recomendación: (i) emplear una métrica de similitud en el conjunto de valoraciones de los usuarios individuales del grupo, proporcionando así un conjunto de vecinos para el grupo de usuarios; (ii) construir un vecindario para el grupo a partir de los vecindarios de los usuarios individuales, típicamente mediante intersección de conjuntos; (iii) fusionando las recomendaciones obtenidas para cada uno de los miembros del grupo; y (iv) por agregación de las predicciones hechas para cada uno de los miembros del grupo.

Por otra parte, el rápido despliegue de las redes sociales ha puesto de relieve una nueva aplicación de los sistemas recomendadores en el ámbito de lo que se ha denominado *inteligencia grupal*, término con el que la comunidad investigadora se refiere a un modo particular en el que los usuarios, fuertemente condicionados por el grupo al que pertenecen, conforman su opinión de acuerdo a la opinión general del grupo, llegando incluso a censurar sus propias convicciones cuando éstas entran en contradicción con las del grupo.

Para la elaboración de este tipo de recomendaciones, un aspecto fundamental a valorar es el grado de heterogeneidad del grupo. Algunas soluciones fusionan los perfiles individuales para obtener el perfil de un usuario virtual representativo del grupo, al que aplican directamente algoritmos de recomendación existentes. Este enfoque plantea inconvenientes cuando el grado de heterogeneidad del grupo es elevado, toda vez que las características individuales contrapuestas acaban anulándose mutuamente, lo que produce una

pérdida significativa de conocimiento. En estas circunstancias, la generación previa de recomendaciones individuales seguida de un proceso de fusión de las mismas para maximizar la satisfacción del grupo es un enfoque más apropiado.

Las redes sociales introducen peculiaridades que las distinguen de un grupo más tradicional. Así, si se considera deseable incentivar la participación y el debate, el anonimato no tiene por qué ser una característica imprescindible (entendiendo por anonimato el hecho de que al intercambiar todas las valoraciones de dos usuarios, no cambien los resultados); al contrario de lo que ocurre con muchos grupos tradicionales, la falta de equidad no tiene por qué ser algo negativo, toda vez que el grado de participación e influencia de todos los usuarios no es habitualmente el mismo. Si se consigue identificar de forma precisa los patrones que caracterizan a los líderes del grupo, a aquellos miembros que alimentan el debate y, en general, a los participantes que más contribuyen a mantener *viva* la red social, resulta razonable otorgarles mayor peso a la hora de elaborar recomendaciones para el grupo.

Pero aún más importante es el hecho diferenciador de que los miembros de una red social no están obligados a consumir los recursos seleccionados, ni tampoco a respetar la ordenación sugerida por el sistema. En la mayor parte de los sistemas existentes, los servicios recomendados están destinados a su consumo en grupo y resultan, por tanto, en una imposición para todos los miembros del grupo. Por ello, evitar una insatisfacción profunda de algunos usuarios es una meta importante a perseguir. En consecuencia, los inconvenientes de algunas estrategias que permiten la posibilidad de insatisfacción extrema de algunos miembros del grupo se atenúan cuando se aplican a redes sociales.

8. Aspectos legales

La personalización típicamente choca de frente con la privacidad de los usuarios. De hecho, la efectividad de la primera depende de acumular la mayor cantidad posible de información sobre un usuario, mientras que la segunda aspira a limitar el manejo de información personal por parte de entidades comerciales o administrativas. El juego de equilibrios entre ambas posturas está siendo objeto de cierta controversia y suscita también un debate intenso en la comunidad investigadora.

En primer lugar, parece pertinente preguntarse de quién son los datos que se analizan. Si bien en algunos casos (datos demográficos, datos médicos, etc.) la respuesta parece clara, surgen más dudas en otros escenarios: Los datos que se recogen sobre nuestras llamadas telefónicas, ¿nos pertenecen a nosotros o al operador telefónico que nos proporciona el servicio?; o los datos que se recogerán a través de sensores instalados en nuestros coches, ¿nos pertenecen a nosotros como propietarios del coche o a la compañía fabricante

del vehículo?; etc. En segundo lugar, gran parte del valor de los datos suele proceder de su empleo en objetivos distintos a aquellos para los que fueron recogidos inicialmente, existiendo el riesgo de que datos personales y privados sean usados con fines que sus propietarios ni conocen ni aprueban. En tercer lugar, la naturaleza digital de los datos personales permite que puedan hacerse de forma sencilla copias y distribuirlas globalmente. Por último, los datos se pueden reutilizar y aumentar su valor cuando se relacionan con otros datos.

En la medida en que la privacidad es un derecho en sí misma, existe un marco legal que ha de tenerse en cuenta al implementar aplicaciones que procesen estos datos. Globalmente, este marco está delimitado por pautas de la OCDE [24], que se erigen en el núcleo de leyes más específicas para los distintos países alrededor del mundo, incluida la *Ley Orgánica de Protección de Datos* que rige en España (Ley Orgánica 15/1999 de 13 de diciembre, publicada en el BOE del 14 de diciembre de 1999). Estas leyes pueden ser predominantemente específicas de sector, como en los Estados Unidos, o ser comunes a todos los sectores, como sucede en la Unión Europea. Las pautas de la OCDE plantean restricciones en aspectos tales como la recolección de información personal, los niveles necesarios de consentimiento del usuario, el intercambio y agregación de datos procedentes de diversas fuentes, los límites para el acceso a datos por parte de terceros y para el mercadeo con dichos datos, el requisito de mantener al usuario informado de para qué se utilizarán sus datos, etc.

Otro aspecto importante es la *transparencia*, entendida como el acceso de los usuarios a los datos que se recogen sobre ellos. Proporcionando acceso a los datos se consigue una mayor conciencia de los usuarios, toda vez que pueden entender mejor su comportamiento y realizar elecciones de contenidos o productos más informadas. Otra discusión que se plantea es cuál es el límite sobre el tipo de inferencias o conclusiones que se pueden extraer del análisis de nuestros datos, toda vez que las recomendaciones especializadas pueden provocar que nuestra *identidad personal* se diluya en la de un colectivo con características comunes y que las empresas lleguen a saber de los usuarios más que los propios usuarios modelando su identidad a su antojo. Por otra parte, la capacidad de este tipo de análisis para medir la probabilidad de un comportamiento no deseado, puede provocar que determinados usuarios sufran discriminación o una especie de *castigo anticipado*, lo que, evidentemente, contraviene valores tan importantes como la presunción de inocencia, la capacidad de decisión o el libre albedrío. Por último, estos tipos de servicios pueden suponer un *riesgo de exclusión* para todas aquellas personas de las que no es posible recabar la información suficiente y que, en consecuencia, se quedan fuera de las ventajas que ofrecen este tipo de sistemas.

En general, la experiencia muestra que vivimos un proceso en el que la investigación y las fuerzas de mercado habitualmente apuran la regulación existente hasta el extremo de revelar nuevos vacíos legales. De ahí se sigue que

el debate sobre legislación no está ni mucho menos cerrado por mucho que las pautas generales no se hayan revisado en varias décadas. Una discusión detallada sobre estos y otros aspectos en cuanto a cómo afectan al desarrollo de servicios de información personalizados puede encontrarse en [19].

9. Algunos dominios de aplicación

Esta sección del discurso está dedicada a presentar el trabajo realizado por el Grupo de Servicios para la Sociedad de la Información de la Universidad de Vigo en el campo de los sistemas recomendadores para diferentes dominios de aplicación.

9.1. Recomendador para televisión

Con la finalidad de validar las métricas de similitud definidas en nuestro trabajo de investigación, desarrollamos un recomendador de programas de televisión que permitía la elaboración de guías electrónicas de programación personalizadas. El objetivo era reemplazar el tradicional modelo inspirado en los buscadores, donde el usuario lanzaba la guía y navegaba por la oferta de canales y programas, por otro basado en recomendadores que identifican automáticamente los contenidos potencialmente más interesantes para cada usuario. La principal contribución de esta implementación fue la definición de una arquitectura de múltiples agentes (ver figura 2) para realizar todas las tareas de personalización, normalizando sus interfaces y delimitando los flujos de información entre agentes.

Este recomendador fue usado en el marco de otro proyecto como núcleo de un sistema para la creación de canales personalizados de televisión para dispositivos móviles usando las tecnologías de difusión DVB-H [23]. El empleo de recomendadores en este tipo de entornos se hace todavía más necesario, toda vez que el visionado de televisión sobre dispositivos móviles se realiza durante cortos intervalos de tiempo (viajes en transporte urbano, estancia en salas de espera, etc.), por lo que los canales generalistas no son adecuados.

9.2. Plataforma de teleenseñanza ATLAS y tutor inteligente T-MAESTRO

El uso de la televisión con fines educativos no era algo nuevo, pero sus posibilidades fueron tradicionalmente muy limitadas hasta la llegada de la televisión digital interactiva, por su capacidad para transmitir ficheros y aplicaciones junto con el audio y vídeo de los programas. El término *t-learning* se acuñó precisamente para referirse a los enfoques que trataban de sacar provecho de estas capacidades. Por ser la televisión un dispositivo de comunicación prácticamente universal y habida cuenta de las diferencias técnicas de estas tecnologías con respecto a las de acceso a Internet a través de PC,

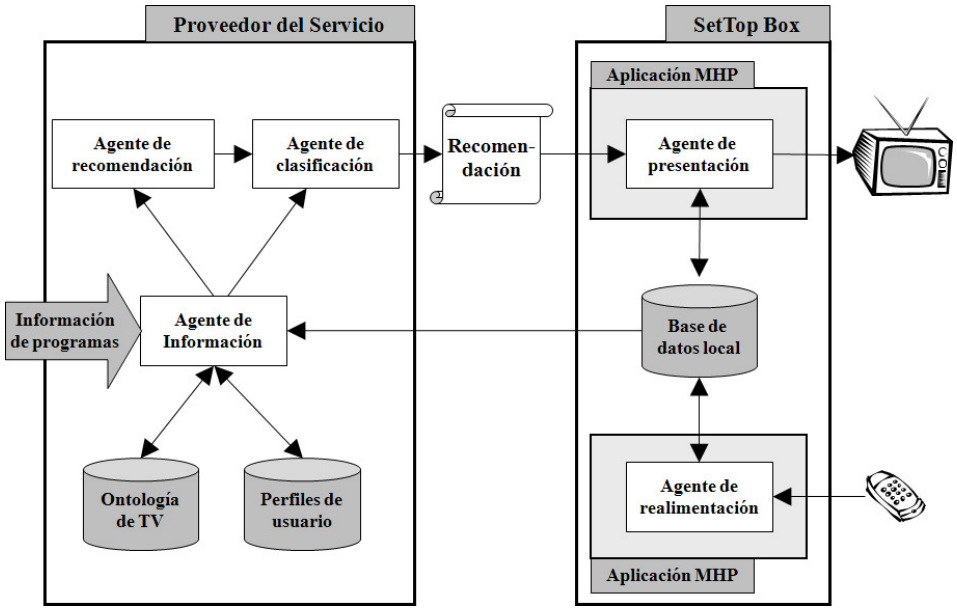

Figura 2: La arquitectura multiagente del sistema AVATAR.

pronto se vio la necesidad de dar al *t-learning* un enfoque diferenciado al *e-learning* y no construir soluciones para el primero como meras traslaciones de aplicaciones existentes en el segundo.

Con estos antecedentes, diseñamos la plataforma ATLAS (*Arquitecture for T-Learning interActive Services*) como un marco para el desarrollo y despliegue de servicios educativos adaptados a las características técnicas de la televisión digital, con una orientación informal basada en el predominio de los contenidos audiovisuales sobre el texto y las imágenes estáticas, y donde las funcionalidades interactivas pudieran manejarse únicamente a través del mando a distancia. ATLAS [6] se construyó enteramente siguiendo el conjunto de estándares que constituyen la norma SCORM. Con estos mimbres, desarrollamos una especialización del sistema recomendador para televisión con el objetivo de recomendar cursos de *t-learning* razonando sobre contenidos educativos, dependencias entre cursos a modo de prerrequisitos y características de los usuarios tales como su nivel formativo. Sobre este sistema recomendador se diseñó un sistema inteligente (ver figura 3), denominado T-MAESTRO (T-*learning* M*ultimedia* A*daptive* E*ducational* S*ystem based on* R*eassembling TV* O*bjects*), capaz de crear cursos personalizados para cada usuario y adaptados a distintos tipos de receptores y a deficiencias visuales o auditivas que pudieran tener los usuarios. El objetivo de T-MAESTRO [38] fue la construcción de experiencias educativas *a la carta* que combinasen contenidos educativos y contenidos de televisión tradicionales, actuando estos últimos como elemento motivador y potenciador del aprendizaje.

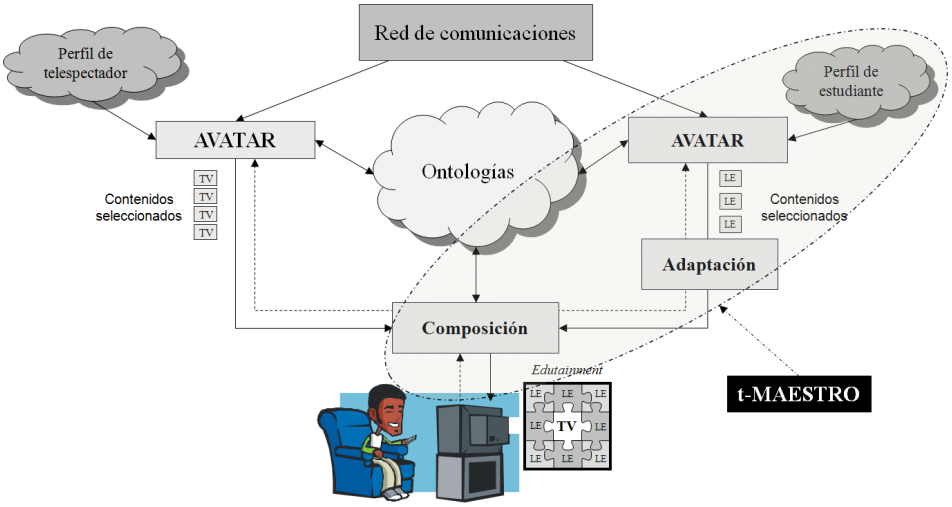

Figura 3: Esquema operacional del sistema T-MAESTRO.

9.3. El sistema de publicidad MiSPOT

El declive en la efectividad de las técnicas tradicionales de publicidad en la televisión deriva principalmente del hecho de presentar los mismos productos a todos los televidentes de un modo que, además, interfiere con el disfrute de los programas. El desarrollo de técnicas de publicidad menos invasivas y la personalización son elementos cruciales para convertir la publicidad en una fuente útil de información para cada teleespectador.

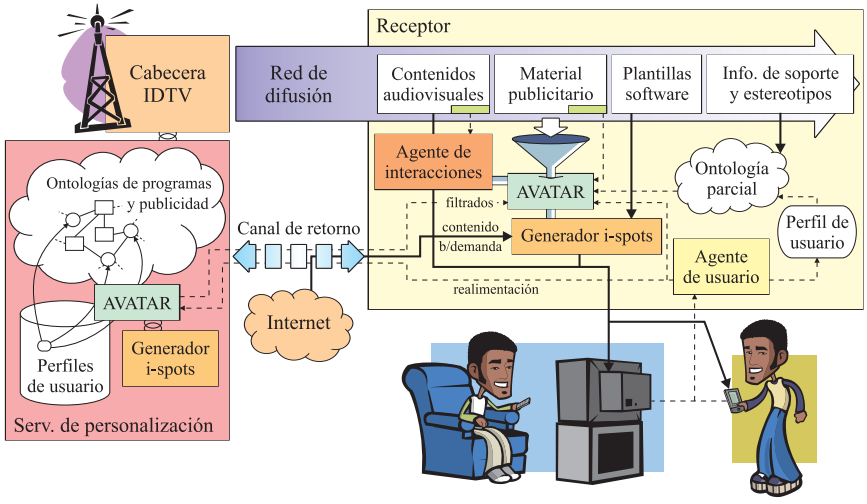

Figura 4: Esquema operacional del sistema MiSPOT.

En este contexto, se planteó el sistema MiSPOT (PlataforMa para la provisión de servicios interactivos perSonalizados de Publicidad nO intru-

siva para TV) con el objetivo de introducir nuevos modelos de publicidad aprovechando el desarrollo de las tecnologías de televisión digital interactiva para receptores fijos y móviles, y además habilitar mecanismos de venta en línea sin precedentes en el medio. Concretamente, MiSPOT [46] (ver figura 4) propuso dos ideas principales:

- *Posicionamiento dinámico de producto*: Con el propósito de ser menos invasivo, el modelo tradicional de publicidad integra el material publicitario con la acción que se ve por pantalla, sin interrumpirla ni ocultarla. No obstante, en este modelo la publicidad se inserta de una vez y para siempre en tiempo de producción de los contenidos, lo que da lugar a pérdidas de eficacia por falta de localidad espacial y temporal. Así, sucede habitualmente que se inserta publicidad de productos que no se comercializan en zonas donde viven muchos teleespectadores o, en contenidos con cierta antigüedad, que los productos se han dejado de comercializar, han cambiado de nombre o han experimentado una evolución en su estética. Al respecto, la propuesta de MiSPOT fue un modelo de publicidad no invasiva y personalizada basada en trasmitir por separado los programas de televisión y los contenidos publicitarios, identificar la publicidad más adecuada para cada teleespectador en el contexto definido por el programa que está viendo, integrar automáticamente la publicidad seleccionada en la pantalla para que se perciba como un elemento más de la acción, y componer automáticamente servicios interactivos (denominados *i-spots*) que ofrecen servicios personalizados de comercio electrónico relacionados con los productos anunciados.

- *Publicidad espontánea*: Ofrece funcionalidades de comercio electrónico asociadas a la semántica de los hechos, objetos o personas que llaman la atención del teleespectador en la pantalla, que pueden precisarse interactuando con menús del mando a distancia o indicarse directamente si se utilizan pantallas táctiles.

9.4. Recomendador de contenidos y destinos turísticos

Los contenidos y destinos turísticos son otro dominio donde la oferta es muy grande y diversa, por lo que los sistemas de recomendación encuentran una aplicación clara. El hecho de que la actividad turística afecte a un porcentaje elevado de la población y que en el caso de nuestro país suponga una actividad económica importante (España es el primer destino mundial en turismo vacacional y el segundo en número de llegadas), lo convierte además en especialmente atractivo. Cuando un turista se encuentra de vacaciones, o bien planificando un viaje, se enfrenta a una oferta de servicios (alojamientos, monumentos, actividades, ...) muy elevada y además desconocida para

él. En este entorno, los sistemas de recomendación pueden ser de gran ayuda, localizando servicios turísticos adaptados a sus preferencias.

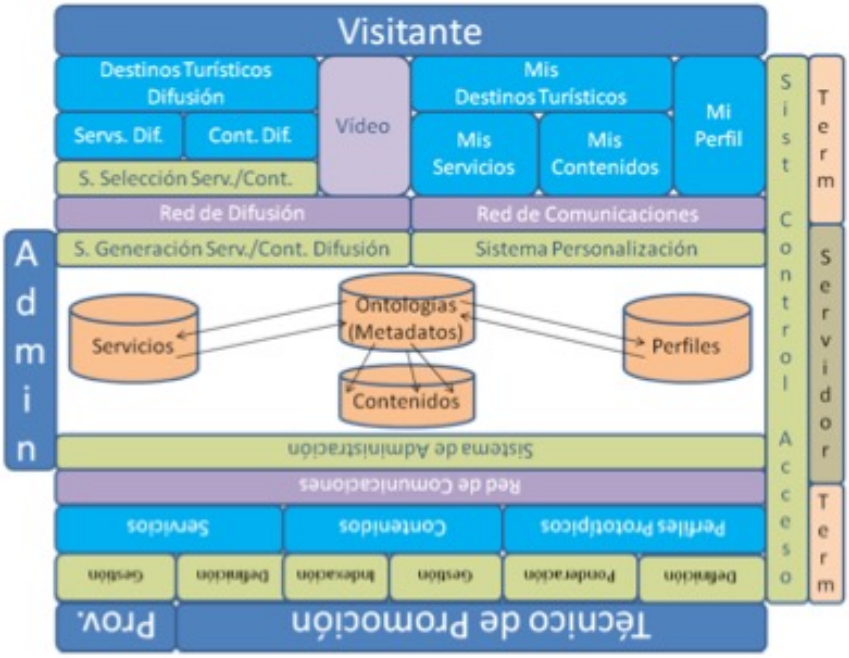

Figura 5: Recomendador de contenidos y destinos turísticos.

En este contexto, nuestro grupo de investigación participó en diversos proyectos en cuyo marco se desarrolló un entorno de recomendación con sensibilidad al contexto [21, 22], que además de recomendaciones individuales de servicios turísticos, permite seleccionar contenidos multimedia turísticos (ver figura 5), de tal forma que en un entorno turístico determinado se generan automáticamente canales de difusión de contenidos turísticos que pueden visualizase en dispositivos móviles y que están adaptados a los gustos del conjunto de turistas presentes en dicho entorno en cada momento mediante técnicas de recomendación grupal (ver sección 7).

Este entorno de recomendación fue utilizado también, en el marco de otros proyectos, para adaptar los contenidos multimedia turísticos que se visualizan en medios de transporte públicos a los gustos del conjunto de usuarios concreto que los estén utilizando en cada momento, y al contexto del viaje, como puede ser el destino o la duración del trayecto, y para generar paquetes turísticos completos para un usuario, que contemplen todas las actividades y servicios que componen el plan de un viaje.

9.5. Redes sociales esporádicas

Las redes sociales tienen un gran número de usuarios que diariamente emplean una cantidad importante de tiempo interactuando activamente con sus contactos, amigos, seguidores, etc. Estas interacciones suelen circunscribirse al mundo virtual y concentrarse en las personas incluidas en el círculo social de cada cual, provocando una des-socialización que aísla al individuo de su entorno introduciéndolo, quizás no conscientemente, en una burbuja de comunicación, dejando en un segundo plano la relación personal con personas que se encuentran físicamente a su alrededor. Este fenómeno social está siendo ampliamente estudiado, aunque sus efectos a medio y largo plazo son todavía una incógnita. Algunos autores hablan de FOMO (del inglés F*ear* O*f* M*issing* O*ut*) para referirse al temor de los usuarios a perderse algún evento, noticia, o situación que ocurre en su red social y que les lleva a estar permanentemente atentos a su dispositivo móvil.

El objetivo de este trabajo fue aplicar los sistemas de recomendación para habilitar nuevas formas de interacción social fuera de la mencionada burbuja, explorando las posibilidades de los contextos que forma cada individuo con las personas (conocidas o no) y los recursos presentes en su entorno físico en un momento dado, en distintos niveles de alcance (estancia, edificio, ciudad, etc.) Con este propósito, se diseñó una plataforma para articular Redes Sociales Esporádicas (RSE) que permiten a cada individuo sacar el máximo partido de esas personas y recursos. La propuesta encuentra aplicación en diversas áreas, desde la formación de grupos de vida corta y la orquestación de actividades en torno a eventos donde se congregan personas con intereses afines (museos, conciertos, campings,...), hasta la provisión de nuevas aplicaciones que amplíen las posibilidades de comunicación entre vehículos (e.g. para optimizar flujos de tráfico, habilitar comunicaciones más ricas que las permitidas por el claxon o distribuir selectivamente publicidad de lugares cercanos) y avances en la visión de las ciudades inteligentes, relacionados con la planificación de la movilidad personal, los bancos de tiempo o la celebración de juegos basados en localización (*urban games*), entre otras.

En primera instancia, las RSE se apoyan sobre redes ad-hoc (sin infraestructura) tendidas entre los dispositivos móviles próximos, si bien es posible también recurrir a la infraestructura proporcionada por conexiones tipo 3G o Wi-Fi para ofrecer funcionalidades mejoradas gracias al acceso a Internet (ver figura 6). En este punto se ha adoptado el paradigma incipiente de la computación en nube móvil para simplificar el acceso a la información y para superar limitaciones de batería y/o capacidad de cómputo de los dispositivos móviles. Por encima de esta capa, se ofrecen funcionalidades avanzadas de acceso a contenidos, explotando conceptos de web semántica, minería de datos y sistemas recomendadores para personalizar la información ofrecida a individuos o a grupos.

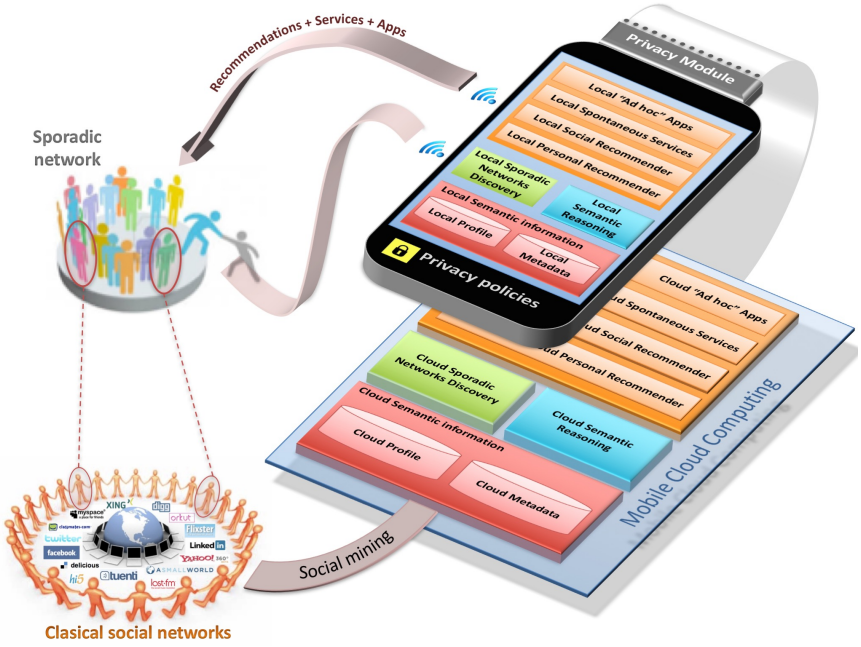

Figura 6: Redes sociales esporádicas.

9.6. SPELTA

La adquisición de habilidades de lenguaje y comunicación son uno de los factores más importantes para el desarrollo de nuestro cerebro. Es de la máxima importancia poder dotarse de herramientas para expresar necesidades, para aprender, para relacionarse con el entorno, y, en general, para tener la oportunidad de desarrollarse uno mismo como miembro activo de la sociedad. Según datos publicados en 2014 por el *National Institute of Deafness and Other Communications Disorders* [47] del *Department of Health & Human Services* de EEUU, 15 millones de personas en el mundo son tartamudas, 6 millones de personas en EEUU tienen alguna discapacidad en el lenguaje, y una de cada 10 personas ha tenido alguna discapacidad de comunicación. El problema es más importante en la educación especial en la que en torno al 70 % de niños tienen problemas de aprendizaje derivados de discapacidades en el habla y en el lenguaje [50].

Los especialistas evalúan, diagnostican y diseñan tratamientos personalizados para una amplia variedad de discapacidades en el habla y en el lenguaje, que en muchas ocasiones están asociadas con otras patologías como la parálisis cerebral, el autismo, etc. Los terapeutas suelen trabajar con los pacientes con una frecuencia semanal, en sesiones con una duración aproximada de una hora en la que se trabajan distintas áreas, realizando ejercicios desde un nivel básico (respiración, tragado, control de la lengua, etc.) a un nivel más avanzado (construcción de frases, ejecución de órdenes comple-

jas, etc.), registrando la actividad y progreso del paciente. Al final de cada sesión, el especialista genera o actualiza el plan de terapia personalizado, tomando elementos de un conjunto formado por cientos o miles de actividades. A la dificultad que para el especialista plantea el manejo de un conjunto tan grande, se une el hecho de que un especialista suele trabajar con varios pacientes en un mismo día y que el conjunto de actividades disponibles no es estático sino creciente. En estas condiciones, al especialista le resulta muy difícil tener en mente este amplísimo conjunto de posibilidades para diseñar una terapia personalizada para cada uno de sus pacientes, por lo que suele emplear únicamente un reducido subconjunto de las actividades disponibles, lo que da lugar a que las terapias diseñadas sean subóptimas.

Con estos antecedentes, el objetivo de nuestro trabajo es diseñar un sistema de recomendación que ayude a los especialistas en la generación y actualización de los planes de terapia personalizados, identificando actividades en cinco áreas: oído, estructura y función oral, formulación lingüística, lenguaje y articulación expresiva, y lenguaje receptivo. Concretamente, hemos desarrollado el sistema SPELTA (SPE*ech* and L*anguage* T*herapy* A*ssistant*) con el objetivo de generar automáticamente dos tipos de planes terapéuticos:

- Un *plan maestro* conteniendo las actividades generales -por ejemplo, *realizar ejercicios para mejorar la fuerza de soplado*- en las cinco áreas citadas y que se elabora para realizarse con cada paciente durante un período de seis meses. El conjunto de actividades para un área específica constituye un *subplan*.

- Un *plan específico* con el conjunto de actividades diarias a realizar durante los seis meses del plan maestro, distribuidas a lo largo de un cierto número de sesiones terapéuticas con duración de 1 ó 2 horas cada una de ellas. Estos ejercicios se eligen de acuerdo a las actividades generales establecidas en el programa maestro, en base a criterios de diversidad, duración, esfuerzo, etc. Por ejemplo, para la actividad general citada en el punto anterior, dos ejercicios específicos serían: *soplar confeti 10 veces durante 2 segundos* o *inflar una pelota en no más de 6 exhalaciones*.

SPELTA (ver figura 7) es una parte de un ecosistema de herramientas diseñadas para dar soporte a la elaboración y seguimiento de terapias, que proporciona un entorno integrado para médicos, estudiantes, terapeutas, pacientes y familiares.

10. Conclusiones

El objetivo principal de los sistemas recomendadores es ofrecer de forma automática y proactiva los contenidos y servicios que mejor se adaptan a las

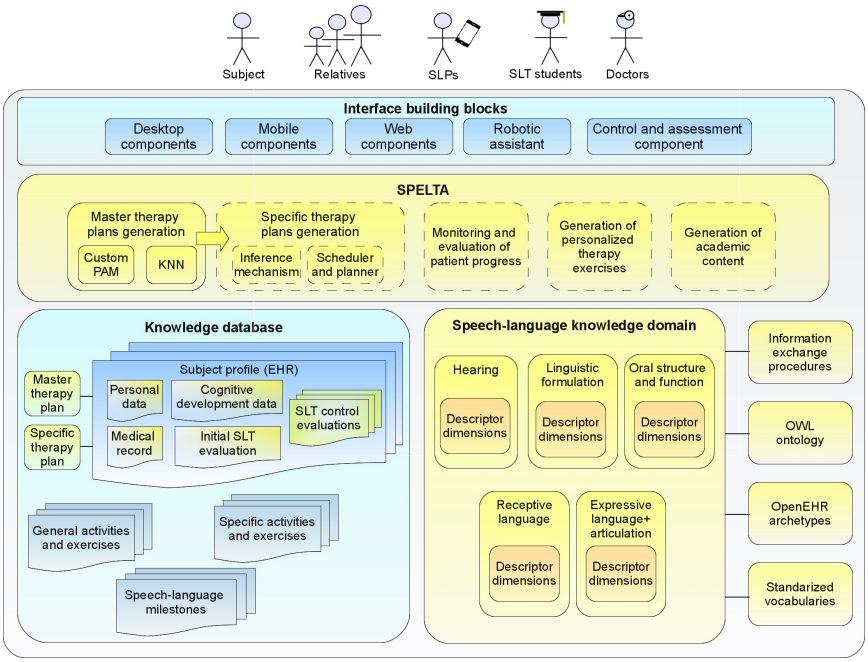

Figura 7: Diseño de alto nivel y componentes principales del entorno SPELTA.

preferencias o necesidades de cada usuario, proporcionando así la tecnología base para la construcción de aplicaciones personalizadas. La consecución de este objetivo plantea importantes problemas de escalabilidad, toda vez que a medida que aumenta el número de productos a recomendar y/o el número de usuarios destinatarios de estas recomendaciones, los recursos informáticos necesarios -principalmente memoria y capacidad de cómputo- se tornan muy elevados, pudiendo provocar retrasos importantes en la elaboración de las recomendaciones, que hacen que la experiencia de usuario sea pobre e incluso que en algunas ocasiones las sugerencias elaboradas sean inútiles.

Por su propia naturaleza, los resultados de los sistemas recomendadores no son exactos, sino únicamente sugerencias, no exentas de fallos. Los fallos provienen de la falta de información sobre los productos a recomendar y también sobre la incompleción e inexactidud sobre los gustos o necesidades de los usuarios reflejados en su perfil. Así, resultará imposible elaborar recomendaciones de calidad si no se dispone de la información suficiente y precisa, tanto de los productos a recomendar como de los gustos, preferencias y contexto de los usuarios consumidores de estas recomendaciones.

Otra dificultad añadida para los sistemas recomendadores la constituyen aquellos dominios de aplicación en los que las preferencias de los usuarios pueden variar rápidamente como suele ocurrir con las modas, o aquellos en que los productos a recomendar provocan reacciones extremas como el amor

y el odio, toda vez que puede resultar muy difícil acertar y, en caso de no hacerlo, el fallo es estrepitoso.

Por otra parte, la implementación, despliegue y mantenimiento de los sistemas recomendadores suele correr a cargo de algún proveedor que de esta forma trata de mejorar la experiencia de usuario, a la vez que asegurar que ningún producto de interés para un cliente pasa desapercibido. Esta situación provoca que los dos inconvenientes citados pongan a las pequeñas empresas en desventaja por no disponer de la cantidad de información necesaria, ni tampoco de la capacidad de procesamiento requerida, de las que si disponen las empresas grandes (Amazon, etc.).

El desarrollo y despliegue de los primeros sistemas recomendadores puso pronto de manifiesto que se podían usar con dos objetivos diferentes, que en algunas ocasiones eran difíciles de conciliar. Por una parte, ayudar a los usuarios a superar el problema de saturación de información seleccionando los productos que mejor se adaptan a sus necesidades o preferencias; y, por otra parte, a proporcionar a las empresas responsables de su despliegue información relevante para aumentar sus ventas o para incrementar el uso de los servicios que ofrecen. Por encargarse de su despliegue y mantenimiento algún proveedor, los sistemas recomendadores actuales se inclinan sobre todo a conseguir un incremento de ventas a corto plazo u otros objetivos de negocio, descuidando en mayor medida la elaboración de recomendaciones de calidad con las que contribuir a que los usuarios superen el problema de saturación de información [45]. Así, no cabe esperar que un sistema recomendador de películas de una plataforma comercial sugiera el visionado de un programa de otro canal que se emite en abierto o por otra plataforma, o que un sistema de recomendación de libros sugiera que no se compre un libro porque se le puede pedir prestado a un amigo que lo adquirió recientemente, o que se deje de comprar música porque este mes se han hecho muchas compras y podría resentirse la economía del usuario, etc.

En la actualidad, no hay ninguna razón técnica por la que un sistema recomendador no pueda centrar la elaboración de sus recomendaciones en los intereses de los usuarios mediante el empleo de fuentes de información más amplias que las que resultan del almacenamiento de datos por parte del proveedor.

Además de los aspectos técnicos citados, los sistemas recomendadores pueden plantear también inconvenientes de tipo social. La generalización en su empleo no está exenta de riesgos. Existe un peligro de crear una sobreespecialización de los gustos, creando grupos estancos de usuarios que comparten los mismos o parecidos intereses y necesidades. Ciertamente, esta situación dará lugar a una menor diversidad y a una mayor polarización de la sociedad, que puede devenir en grupos cerrados, incapaces de comprender los gustos o preferencias de otros grupos y en los que los comportamientos dogmáticos, o incluso intransigentes, lejos de considerarse un defecto a corregir se consideran una virtud.

El empleo de tecnologías de *big data* puede ser de gran ayuda para solucionar o atenuar los principales problemas de los sistemas recomendadores. Por un lado, evitarán los problemas de escalabilidad producidos por el inevitable aumento de los usuarios y de los productos o servicios a recomendar, reduciendo los retardos debidos a la elaboración de un número cada vez más elevado de recomendaciones mediante la paralelización de los algoritmos de filtrado [54]. Por otro lado, la capacidad de procesamiento de grandes volúmenes de datos que ofrecen estas tecnologías permitirá mejorar la calidad de las recomendaciones al poder disponer de una más rica y precisa descripción de los productos y servicios a recomendar, así como de ofrecer mecanismos para capturar y actualizar automáticamente los perfiles de usuario.

Referencias

[1] *Advanced Distributed Learning (ADL). Sharable Content Object Rreference Model (SCORM)*, 2004. http://www.adlnet.org.

[2] G. Adomavicius and A. Tuzhilin. *Recommender Systems Handbook*, chapter Context-Aware Recommender Systems, pages 217–253. Springer, October 2010. ISBN: 978-0-387-85819-7.

[3] Gediminas Adomavicius and Alexander Tuzhilin. Toward 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):734–749, June 2005.

[4] Sihem AmerYahia, Senjuti Basu Royz, Ashish Chawlayz, Gautam Dasz, and Cong Yuy. Group recommendation: Semantics and efficiency. In *Proceedings of the VLDB Endowment*, volume 2, pages 754–765, August 2009.

[5] Grigoris Antoniou and Frank van Harmelen. *A semantic web primer*. The MIT Press, 2004.

[6] José J. Pazos Arias, Martín López Nores, Jorge García Duque, Alberto Gil Solla, Manuel Ramos Cabrer, Yolanda Blanco Fernández, Rebeca P. Díaz Redondo, and Ana Fernández Vilas. ATLAS: A framework to provide multiuser and distributed t-learning services over MHP. *Software, Practice and Experience*, 36(8):845–869, July 2006.

[7] Luigi Atzoria, Antonio Ierab, and Giacomo Morabito. The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805, October 2010.

[8] Matthias Baldauf, Schahram Dustdar, and Florian Rosenberg. A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2(4):263–277, 2007.

[9] Shlomo Berkovsky and Jill Freyne. Group-based recipe recommendations: analysis of data aggregation strategies. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 111–118. ACM, 2010. ISBN: 978-1-60558-906-0.

[10] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, July 2013.

[11] Jesús Bobadilla, Antonio Hernando, Fernando Ortega, and Jesús Bernal. A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38(12):14609–14623, December 2011.

[12] Geoff Brumfiel. High-energy physics: Down the petabyte highway. *Nature*, 469:282–283, January 2011.

[13] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[14] Laurent Candillier, Frank Meyer, and Marc Boullé. Comparing state-of-the-art collaborative filtering systems. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Computer Science*, pages 548–562. Springer, July 2007. ISBN: 978-3-540-73499-4.

[15] Capgemini. The Internet of Things: are organizations ready for a multi-trillion dollar prize? Capgemini Consulting. http://www.capgemini-consulting.com/resource-file-access/resource/pdf/the-internet-of-things.pdf, 2014.

[16] C.L. Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275(10):314–347, August 2014.

[17] Hsinchum Chen, Roger H. L. Chiang, and Veda C. Storey. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 2012.

[18] Cristina Conati, Kathleen McCoy, and Georgios Paliouras, editors. *11th International Conference on User Modeling*, volume 4511 of *Lecture Notes in Computer Science*. Springer, July 2007.

[19] Teresa Rodríguez de las Heras Ballell. Legal aspects of recommender system in the web 2.0: Trust, liability and social networking. In José J. Pazos Arias, Ana Fernández Vilas, and Rebeca P. Diaz Redondo, editors, *Recommender System for the Social Web*, volume 32 of *Intelligence Systems Reference Library*, pages 43–62. Springer, 2012.

[20] Félix Hernández del Olmo and Elena Gaudioso. Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, 35(3):790–804, November 2008.

[21] Yolanda Blanco Fernández, Martín López Nores, José J. Pazos Arias, Jorge García Duque, and Manuela I. Martín Vicente. TripFromTV+: Targeting personalized tourism to interactive digital TV viewers by social networking and semantic reasoning. *IEEE Transactions on Consumer Electronics*, 57(2):953–961, 2011.

[22] Yolanda Blanco Fernández, Martín López Nores, José J. Pazos Arias, Alberto Gil Solla, and Manuel Ramos Cabrer. Exploiting digital TV usersś preferences in a tourism recommender system based on semantic reasoning. *IEEE Transactions on Consumer Electronics*, 56(2):1–9, 2010.

[23] Yolanda Blanco Fernández, Martín López Nores, Alberto Gil Solla, Manuel Ramos Cabrer, and José J. Pazos Arias. User-generated contents and reasoning-based personalization: Ingredients for a novel model of mobile TV. *Expert Systems with Applications*, 38(5):5289–5298, May 2011.

[24] Organization for Economic Co-operation and Development. OECD guidelines on the protection of privacy and transborder flows of personal data. http://www.oecd.org, 1980.

[25] J. Gantz and D. Reinsel. Extracting value from chaos, 2011. IDC iView, pp. 1-12.

[26] Gartner. Gartner says the Internet of Things will transform the data center. Gartner News Room. http://www.gartner.com/newsroom/id/2684616, 2014.

[27] Jayavardhana Gubbia, Rajkumar Buyyab, Slaven Marusica, and Marimuthu Palaniswami. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1465–1660, September 2013.

[28] Susan Gunelius. The Data Explosion in 2014 Minute by Minute - Infographic. http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/, July 2014.

[29] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009.

[30] Tony Hey and Anne E. Trefethen. The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*, 18(8):1017–1031, October 2002.

[31] IDC. The digital universe of opportunities: rich data and the increasing value of the Internet of Things. http://www.emc.com/leadership/digital-universe/2014iview/index.htm, 2014.

[32] Anthony Jameson and Barry Smyth. Recommendation to groups. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 596–627. Springer, 2007. ISBN: 978-3-540-72078-2.

[33] Michal Kompan and Maria Bielikova. Group recommendations: Survey and perspectives. *Computing and Informatics*, 33(2):446–476, 2014.

[34] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, August 2009.

[35] Bruce Krulwich. LIFESTYLE FINDER: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37–45, 1997.

[36] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, pages 208–211, February 2008.

[37] D. Laney. 3-d data management: controlling data volume, velocity and variety, 2001. META Group Research Note.

[38] Marta Rey López, Rebeca P. Díaz Redondo, Ana Fernández Vilas, José J. Pazos Arias, Martín López Nores, Jorge García Duque, Alberto Gil Solla, and Manuel Ramos Cabrer. T-MAESTRO and its authoring tool: Using adaptation to integrate entertainment into personalized t-learning. *Multimedia Tools and Applications*, 40(3):409–451, December 2008.

[39] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender systems handbook*, pages 73–105. Springer, 2011. ISBN: 978-0-387-85820-3.

[40] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments: A survey. *Decision Support System*, 74:12–32, June 2015.

[41] Xin Luo, Yunni Xia, and Qingsheng Zhu. Incremental collaborative filtering recommender based on regularized matrix factorization. *Knowledge-Based Systems*, 27:271–280, March 2012.

[42] Xin Luo, Yunni Xia, and Qingsheng Zhu. Applying the learning rate adaptation to the matrix factorization based collaborative filtering. *Knowledge-Based Systems*, 37:154–164, January 2013.

[43] B. S. Manjunath, Philippe Salembier, and Thomas Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, April 2002. ISBN: 978-0-471-48678-7.

[44] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big

data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation, June 2011.

[45] Francisco J. Martin, Justin Donaldson, Adam Ashenfelter, Marc Torrens, and Rick Hangartner. The Big Promise of Recommender Systems. *AI Magazine*, 32(3):19–27, 2011.

[46] Martín López Nores, José J. Pazos Arias, Jorge García Duque, Yolanda Blanco Fernández, Manuela I. Martín Vicente, Ana Fernández Vilas, Manuel Ramos Cabrer, and Alberto Gil Solla. MiSPOT: Dynamic product placement for digital TV through MPEG-4 processing and semantic reasoning. *Knowledge and Information Systems*, 22(1):101–128, January 2010.

[47] National Institute of Deafness and Other Communications Disorders. Statistics on voice, speech and language. https://www.nidcd.nih.gov/health/statistics/pages/vsl.aspx, May 2014.

[48] Tim O'Reilly. What is web 2.0? design patterns and business models for the next generation of software. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html, September 2005.

[49] Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 21–28. ACM, 2009. ISBN: 978-1-60558-435-5.

[50] R. Parrilla and C. Sierra. Trastornos del lenguaje. *Boletín de la SPAO*, 7, 2013.

[51] Michael J. Pazzani and Daniel Billsus. Content-based recommendation systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer, 2007. ISBN: 978-3-540-72078-2.

[52] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001. ISBN:1-58113-348-0.

[53] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 291–324. Springer, 2007. ISBN: 978-3-540-72078-2.

[54] Sebastian Schelter, Christoph Boden, and Volker Markl. Scalable similarity-based neighborhood methods with MapReduce. In *Proceedings of the 6th ACM conference on Recommender systems*, pages 163–170, September 2012. ISBN: 978-1-4503-1270-7.

[55] Alberto Gil Solla and Rafael G. Sotelo Bovino. *TV-Anytime: paving the way for personalized TV*. Springer, 2013.

[56] Le Hoang Son. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Information Systems*, In press, 2014.

[57] Manuel I. Martín Vicente, Alberto Gil Solla, Manuel Ramos Cabrer, Yolanda Blanco Fernández, and Martín López Nores. A semantic approach to avoiding fake neighborhoods in collaborative recommendation of coupons through digital TV. *IEEE Transactions on Consumer Electronics*, 56(1):54–62, February 2010.

[58] Manuela I. Martín Vicente, Alberto Gil Solla, Manuel Ramos Cabrer, Jose J. Pazos Arias, Yolanda Blanco Fernández, and Martín Lopez Nores. A semantic approach to improve neighborhood formation in collaborative recommender systems. *Expert Systems With Applications*, 41(17):7776–7788, December 2014.

[59] Alessandro Vinciarelli and Gelareh Mohammadi. A survey on Personality Computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.