UNIVERSITAT POLITÈCNICA DE CATALUNYA

Departament de
Llenguatges i Sistemes Informàtics

Master and PhD in Computing

TESI DE MÀSTER

Sistemas de recomendación para webs de información sobre la salud

ESTUDIANT: Miguel Seguido Font

DIRECTOR: Carles Farré

DATA: 23 de Junio del 2009

"*Silent gratitude isn't much use to anyone.*"
Gladys Bronwyn Stern.

A mi padre, por el apoyo mostrado durante este tiempo; Gracias.

Índice

1 Introducción

La presente tesis de máster representa un estado del arte de los sistemas de recomendación en el hábito computacional. Mediante un profundo estudio de la literatura se ha desarrollado un análisis de los diferentes sistemas de recomendación existentes así como su clasificación, bondades y defectos. Este estudio del estado del arte de los sistemas de recomendación se ha llevado a cabo con el fin de obtener una idea clara de las posibles soluciones (sistemas de recomendación) a implementar para un proyecto del "Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED)" llamado "Calidad en los sitios del área de la salud". El proyecto "calidad en los sitios del área de la salud" consiste en la creación de una aplicación web dedicada a temas de salud.

1.1 Motivación

A menudo es necesario tomar decisiones sin tener la experiencia necesaria o suficiente sobre las alternativas a elegir. Por ello, en el día a día, las personas confían y tienen en cuenta las recomendaciones de terceros para poder llevar a cabo decisiones de mayor provecho o interés.

Esta falta de experiencia o conocimiento sobre las alternativas, así como el gran incremento de la información que genera la misma sociedad, hace que de cada día vaya cobrando más importancia el hecho de tener métodos automáticos o semiautomáticos de filtrado y/o selección de información para la toma de decisiones.

Así como comenta Hanani et al. en [1], hoy en día se crea y entrega al usuario una gran cantidad de contenido electrónico tal que los usuarios se saturan por ese flujo de información. La utilización de herramientas para abordar la saturación de los usuarios se ha vuelto irremediablemente necesaria. Este tipo de herramientas, que proveen algún tipo de recomendación, ayudan a los usuarios a entender mejor la información necesaria para hacer un uso más efectivo de ella [2].

La combinación del crecimiento desmesurado de la información disponible y la falta de experiencia en sectores determinados por parte del usuario, hacen evidente la incitación para la búsqueda de sistemas automáticos o semiautomáticos capaces de

ayudar a los usuarios en la toma de decisiones mediante la aportación de sugerencias o indicaciones.

1.2 Objetivo

El objetivo de esta tesis es la creación de un estado del arte de sistemas de recomendación con el fin de aportar conocimiento útil para la creación de un sistema de recomendación basado en tecnología web y enfocada en el área de salud. Así pues, se pueden distinguir tres sub-objetivos a abordar:

- Proveer de un comprensivo y profundo análisis sobre el estado del arte de sistemas de recomendación mediante el estudio de la literatura correspondiente.
- Evaluar cuales de las soluciones existentes, obtenidas mediante el estudio del estado del arte de los sistemas de recomendación, se ajustan mejor con los requerimientos de un proyecto de creación de una página web de sistemas de salud.
- Proponer un esquema para la adopción o implementación de un sistema de recomendación como solución particular al proyecto de creación de una página web de sistemas de salud.

1.3 Estructura de la tesis

Teniendo presentes los objetivos anteriores, la tesis se divide en un primer apartado teórico y dos apartados con un enfoque más práctico. En primer lugar se aborda una revisión de la literatura de los sistemas de recomendación mediante la creación del estado del arte de los sistemas de recomendación; En segundo lugar, se realiza la aplicación práctica del estudio realizado sobre los sistemas de recomendación en donde se pueden distinguir los siguientes apartados:

1. La evaluación de las soluciones existentes sobre los sistemas de recomendación para los requerimientos de una proyecto de página web de sistemas de salud y,
2. Proposición de un esquema que se adapte a los requerimientos del proyecto de página web de sistemas de salud.

1º Acto: Obertura

Estado del arte

El arte es la filosofía que refleja un pensamiento

2 Sistemas de recomendación

A continuación se presentan los aspectos generales de los sistemas de recomendación. En primer lugar en el apartado "2.1 Historia" se describen los primeros pasos que se dieron para la creación de sistemas de recomendación así como se nombran y describen los primeros sistemas de recomendación que aparecieron en la literatura. A continuación, en el apartado "2.2 Definición y característica", se exponen diferentes definiciones que estos sistemas han ido tomando a lo largo del tiempo así como sus características más significativas. Para finalizar, en el apartado "2.3 Clasificación", se realiza una categorización general de los sistemas de recomendación desde diferentes puntos de vista.

2.1 Historia

A principios de la década de los 90 empezaron a surgir dentro de los servicios de grupos de noticias (*newsgroups*), servicios de filtrado de noticias que permitían a su comunidad de usuarios acceder exclusivamente a aquellas noticias que potencialmente podían ser de su interés [3,4]. No obstante, el primer sistema de recomendación que apareció fue el llamado "*Tapestry*" [5], desarrollado por XeroX PARC. Tapestry es un sistema que permite almacenar el feedback de los usuarios sobre los artículos o noticias que éstos han leído y posteriormente ser utilizado por otros usuarios que aún no han leído el artículo o noticia, para establecer si la información del documento es relevante o no. En un principio este tipo de sistemas fue adoptado con el nombre de filtro colaborativo (*collaborative filter*) dado que permite que los usuarios creen filtros a través de sus ítems de interés (en el caso de Tapestry, artículos o noticias), y colaborativo pues los usuarios añaden las anotaciones con las opiniones sobre los documentos. Las opiniones añadidas pueden ser utilizadas para las búsquedas de otros usuarios. Ya en el 1997, Resnick y Varian proponen llamar a los sistemas con el nombre de "sistemas de recomendación" (*recommender systems*), dado que por esa fecha estos sistemas no sólo se limitaban al filtro de información y habían aparecido nuevos sistemas en el que no se utilizaban las opiniones de otros usuarios [6].

Otra de las primeras formas de filtrado de información electrónica apareció con el trabajo de Housman y Kaskela [7] en el que se diseñó un método que de forma automática se pudiera mantener a los científicos informados sobre nuevos documentos publicados en sus áreas de trabajo o especialización. El método se basaba en la creación de un perfil de usuario que contenía ciertas palabras clave, relevantes para el usuario que son utilizadas para buscar coincidencias entre estas palabras clave y los nuevos documentos o artículos con el fin de intentar predecir qué artículos o documentos serian del interés de los científicos. Esta primera aproximación a los sistemas de recomendación fue implementada, pero su uso fue mucho menor de lo esperado [8].

A partir de entonces aparecieron nuevas ideas como la propuesta por Allen [9] en la que se consideraba la creación de modelos de usuario, con el objetivo de predecir qué nuevas publicaciones de artículos los científicos se prestarían a leer. El método, pese a mejorar los resultados de Packer y Soergel [8], era más bueno prediciendo preferencias de usuario en categorías generales que de artículos específicos [10].

Otra aproximación a los sistemas de recomendación se dio con "*The information Lens system*" [11,12] en donde, basado en el contexto del correo electrónico, se les permitía a los usuarios crear reglas para filtrar los correos electrónicos. Así pues, los usuarios podían, por ejemplo, crear una regla para etiquetar todos los mensajes provenientes de cierta persona o correo electrónico. El problema de "*The information Lens system*" apareció con las personas o usuarios cuyo conocimiento en el ámbito de la informática era elemental o casi nulo. Estos usuarios eran incapaces de crear reglas de filtrado para priorizar o filtrar los correos recibidos [12].

2.2 Definición y característica

Desde los inicios de los sistemas de recomendación su definición ha ido cambiando y evolucionando a lo largo del tiempo. A continuación se muestran diferentes definiciones que se le ha dado a los sistemas de recomendación a lo largo de los años [13]:

1992 [5]: “Collaborative filtering simply means that people collaborate to help one another perform filtering by recording their reactions to documents they read.”

1994 [3]: “Collaborative filters help people make choices based on the opinions of other people.”

1997 [6]: “In a typical recommender system people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients.”

2001 [14]: “Recommender systems use product knowledge—either hand-coded knowledge provided by experts or ‘mined’ knowledge learned from the behavior of consumers—to guide consumers through the often-overwhelming task of locating products they will like.”

2001 [15]: “The term ‘collaborative filtering’ describes techniques that use the known preferences of a group of users to predict the unknown preferences of a new user; recommendations for the new users are based on these predictions. Other terms that have been proposed are ‘social information filtering’ and ‘recommender system’.”

2004 [16]: “Recommender systems use the opinions of a community of users to help individuals in that community more effectively identify content of interest from a potentially overwhelming set of choices.”

2004 [17]: “Recommender systems—a personalized information filtering technology used to either predict whether a particular user will like a particular item (prediction problem) or to identify a set of N items that will be of interest to a certain user (top-N recommendation problem).”

2004 [18]: “Recommender system is a system that helps users to find their wanted items by making recommendations based on either the content of the recommended items (content-based filtering), or ratings of similar users on the recommended items(collaborative filtering).”

2005 [19]: “A recommender system is a typical software solution used in e-commerce for personalized services. It helps customers find the products they would like to purchase by providing recommendations based on their preferences, and is partially useful in e-commerce sites that offer millions of products for sale.”

2005 [20]: “A recommender system is the information filtering that applies data analysis techniques to the problem of helping customers find the products they would like to purchase by producing a predicted likeness score or a list of recommended products for a given customer.”

Como se observa, a lo largo de los años la definición de los sistemas de recomendación ha ido evolucionando y siempre ha ido ligada con el avance de las nuevas técnicas o ideas que iban surgiendo en la literatura de los sistemas de recomendación. Así pues, a partir de 1997 empieza a aparecer un cambio significativo en la forma de denominar estos sistemas, pasando de ser llamados “sistemas de filtrado colaborativo” a “sistemas de recomendación”. Ya en el 2004 las definiciones contemplan los sistemas de recomendación mixtos. Por otro lado, cabe destacar que en los últimos años, los sistemas de recomendación han sido ampliamente utilizados en el ámbito del comercio electrónico, aspecto que queda plasmado en la definición dada por Kim et al., en el 2005 [19].

Según afirma Burk en [21], los sistemas de recomendación se distinguen por dos criterios fundamentales: Por un lado se encuentran aquellos que producen recomendaciones individualizadas en forma de "output" y, por otro, aquellos que tienen el efecto de guiar al usuario de forma personalizada para los intereses de éste dentro de un dominio con grandes cantidades de ítems posibles a elegir. Teniendo esto en cuenta, se puede afirmar que los sistemas de recomendación son un tipo específico de filtro de información cuyo objetivo es mostrar ítems (libros, artículos, películas, imágenes, web-sites, etc) al usuario que le sean relevantes o de interés. Se entiende por filtro de información un sistema que elimina información inadecuada o no deseada de un flujo de información de forma automática o semiautomática para ser presentada a los usuarios [1].

Un aspecto relacionado con los sistemas de recomendación es la "recuperación de información" o "búsqueda de información" (*information retrieval*). Ambos conceptos se asemejan en el hecho de que intentan proporcionar información relevante al usuario pero se distinguen por ciertas singularidades [22]:

- **Frecuencia de uso:** Los sistemas de búsqueda son enfocados por uso concreto y puntual del usuario mientras que los sistemas de recomendación están diseñados para un uso prolongado y de forma repetitiva.
- **Representación de las necesidades de información:** En sistemas de búsqueda la información requerida se expresa en forma de pregunta (*query*) mientras que en los sistemas de recomendación la información es descrita en los perfiles del usuario.
- **Objetivo:** Los sistemas de búsqueda seleccionan los ítems de la base de datos que coinciden con la pregunta (*query*) mientras que los sistemas de recomendación eliminan la información irrelevante de flujos de entrada de información o reúnen información relevante de diferentes repositorios de acuerdo al perfil del usuario.
- **Base de datos:** Los sistemas de búsqueda trabajan con bases de datos relativamente estáticas mientras que los sistemas de recomendación trabajan con información dinámica.
- **Tipo de usuarios:** En los sistemas de búsqueda no se tiene porque tener información sobre los usuarios que lo utilizan mientras que en los sistemas de recomendación se necesita saber o tener información sobre los usuarios.
- **Ámbito social:** Los sistemas de recomendación están interesados en aspectos sociales de modelado y privacidad del usuario mientras que los sistemas de búsqueda no.
- **Acción:** El proceso de filtrado o recomendación se relaciona con la acción de "eliminar" información, mientras que el proceso de búsqueda se relaciona con la acción de "encontrar" información.

2.3 Clasificación general

Los sistemas de recomendación se pueden clasificar, dependiendo del tipo de información que utilizan para realizar sus recomendaciones, en diferentes tipos. Tradicionalmente existen dos paradigmas para la selección de elementos o filtrado de acuerdo con la manera de realizar las recomendaciones [23-25]:

- **Basados en contenido (*Content based*):** Generan recomendaciones equiparando las preferencias del usuario (expresadas por éste de forma implícita o explícita) con las características utilizadas en la representación de los ítems ignorando la información relativa de otros usuarios. En otras palabras, se le recomendará al usuario un ítem similar al que el mismo usuario haya elegido anteriormente. Por ejemplo, si un usuario es afín a los libros de historia, el sistema centrará sus recomendaciones en cualquier libro etiquetado con el tema de historia. [26] "*InfoFinder*" y [27] "*NewsWeeder*" ofrecen diferentes ejemplos basados en este tipo de sistema de recomendación. En el apartado "3 Sistemas basados en el contenido", se realiza un estudio exhaustivo sobre este tipo de sistemas de recomendación.
- **Filtrado colaborativo (*Collaborative filter*):** También llamados sistemas de recomendación sociales. Son un tipo de sistemas de recomendación para los que se crea un conjunto de usuarios que tienen características o gustos similares entre sí. Si a un conjunto del grupo de usuarios les gusta un determinado ítem, es de suponer que ese mismo ítem guste al resto de usuarios de ese grupo; En otras palabras, se le recomendará al usuario un ítem que a las personas con mismos gustos y preferencias les haya gustado en el pasado. Algunos ejemplos de sistemas de filtro colaborativo los podemos encontrar en [3] "*GroupLens*", [28] "*Bellcore*" y [29] "*Ringo*". El apartado "4 Sistemas de filtrado colaborativo", profundiza sobre este tipo de sistemas de recomendación.
- Por otra parte, Malone et al., propone en [30] un tercer sistema basado en factores económicos el cual genera recomendaciones basándose en elementos de coste. También Resnick et al., en [3] se refiere a este sistema que busca las relaciones entre diferentes conceptos para así proporcionar una medida

equitativa de recomendación. Por ejemplo, la relación entre calidad y precio de un ítem o en el caso de las telecomunicaciones: ancho de banda y el tamaño del fichero a descargar. El ejemplo que usa Malone et al., [30] concierne a las decisiones basadas en el coste-valor que toman las personas a la hora de procesar un cierto ítem. En el caso de un documento, Malone afirma que uno de los principales costes de este tipo de ítem viene determinado por la longitud del documento. Pese a que este sistema es utilizado instintivamente por las personas en su día a día, su uso en sistemas de recomendación es muy marginal.

Otros tipos de sistemas de recomendación se han propuesto en la literatura. R.Burke en [21] distingue los siguientes:

1. **Recomendaciones demográficas (*Demographic*):** Clasifican a los usuarios de acuerdo a su perfil y hacen las recomendaciones basándose en clases demográficas. Un ejemplo enmarcada en la búsqueda de libros en una biblioteca se encuentra en [31]. Según HJ Ahnlas, las recomendaciones demográficas son similares a las recomendaciones basadas en el contenido con la excepción de que las similitudes están calculadas a partir de la utilización de información demográfica en lugar de valoraciones de los ítems [32].
2. **Recomendaciones basadas en el conocimiento (*Knowledge based*):** Las sugerencias de los ítems se basan en inferencias sobre las necesidades de los usuarios y sus preferencias. Para ello se utiliza conocimiento en donde se tiene información sobre cómo un ítem específico responde a una necesidad en particular del usuario y, por lo tanto, la razón sobre la relación entre la necesidad y una posible recomendación. Varios ejemplos se encuentran en [33]. En el apartado "5 Recomendaciones basadas en el conocimiento" se profundiza sobre este tipo de sistemas de recomendación.
3. **Recomendaciones basadas en la utilidad (*Utility based*):** Son sistemas de recomendación que crean una función de utilidad para cada ítem la cual interviene directamente en el proceso de recomendación. La ventaja de este método es que permite evaluar elementos no atribuibles al producto o ítem en sí. Aspectos como la fiabilidad de un proveedor o la disponibilidad de un ítem

estarían representados en la función de utilidad. Un ejemplo de su uso es el "*EQO: European Quality Observatory*" [34]. En cierto modo, este tipo de sistemas se puede considerar un subgrupo de los sistemas de recomendación basados en el conocimiento ya que ambos utilizan la misma base de conocimiento. Por ello, en el apartado "5 Recomendaciones basadas en el conocimiento" se encuentra "5.2 Recomendaciones basadas en la utilidad" en donde se amplía la información de este tipo de sistemas de recomendación.

Por otro lado, se encuentran los sistemas de recomendación semánticos (*semantic recommender systems*) que mejoran y enriquecen la representación de la información mediante la aplicación de tecnologías de Web Semántica [35]. Se pueden clasificar en tres tipos:

1. Sistemas basados en ontologías o esquemas de conceptos: Las ontologías (esquemas conceptuales de la información correspondientes a un dominio en concreto), son utilizadas para representar la información o modelado de los ítems así como para la modelación de usuarios o perfiles de usuario.
2. Sistemas adaptables al contexto: Toman en consideración diferentes factores del usuario tales como temporales, de lugar, nivel de experiencia, dispositivo que se está utilizando en el momento de recibir la recomendación, etc. para inferir el contexto en que se encuentra el usuario y adaptar las recomendación a esas circunstancias.
3. Sistemas basados en redes de confianza: Añaden filtros de información adicionales a los sistemas adaptables al contexto.

En el apartado "6 Sistemas de recomendación semánticos" se ha realizado un estudio en profundidad de este tipo de sistemas de recomendación.

Para finalizar, existen los sistemas de recomendación híbridos (*Hybrid Recommender Systems*), es decir, que combinan diversos tipos de sistemas de recomendación. El sistema híbrido más común utiliza las recomendaciones basadas en el contenido y además, el filtrado colaborativo. Como consecuencia, se atenúan las desventajas de cada una de ellos y se combinan los beneficios de las recomendaciones de ambos métodos [23,36]. En este campo también se contemplan los Meta-Recomendadores,

herramientas que permiten al usuario personalizar las sugerencias del sistema mediante la combinación de varias fuentes de recomendación y utilizando diversas técnicas de filtrado [37]. En el apartado “7 Sistemas híbridos” se amplía la información referente a la combinación de diferentes sistemas de recomendación.

2.3.1 Clasificación exhaustiva

En el apartado anterior “2.3 Clasificación general” se han mostrado los diferentes sistemas de recomendación de la literatura ordenados según su característica principal, en otras palabras, según son conocidos en la jerga. No obstante, existen otros matices o factores a tener en cuenta que permiten una clasificación más apurada de los sistemas de recomendación [16,18,21,38]. Además, también se utilizan aspectos concretos como pueden ser el comercio electrónico [14,39] o el uso de agentes [40] para enfatizar la clasificación (ver “6.1.2 Agentes” para una explicación más detallada sobre el término de agentes).

Según las referencias expuestas anteriormente, la clasificación (exhaustiva) de los sistemas de recomendación se enmarca en tres categorías: Enfoque lógico (*rationale*), Enfoque de aproximación (*approach*) y operaciones (*operation*). A continuación se muestran los detalles de cada una de estas categorías:

2.3.1.1 Enfoque lógico

El “Enfoque lógico” se puede dividir en dos grandes grupos [16]: “Tareas soportadas” y “Decisiones al problema”. A continuación se explican cada una de ellas:

1. **Tareas soportadas (*Supported tasks*):** Hace referencia a la distinción de los sistemas de recomendación de acuerdo a las tareas de usuario que el sistema está destinado a soportar. Se pueden distinguir los siguientes casos:
    a. **Anotación en el contexto (*Annotation in context*):** El sistema de recomendación se diseña para ser integrado en un trabajo ya existente del usuario para producir algún tipo de información adicional. Por ejemplo, en el caso de una página web, mostrarle al usuario, en forma de recomendación, links de otras páginas de su interés [41].

b. **Encontrar ítems válidos (*Find good items*):** Consiste en recomendar un ítem a un determinado usuario. Este es el caso más general de los sistemas de recomendación [25]. Existen variantes de esta tarea que en lugar de realizar la recomendación de un ítem, recomiendan los N mejores ítems al usuario.
c. **Encontrar todos los ítems válidos (*Find all good items*):** Este caso se da en situaciones muy determinadas en las cuales el usuario quiere todos los ítems que le pueden ser de interés. Se da en campos como la medicina o casos legales, donde es muy importante el escrutinio de todos los casos potencialmente útiles [42].
d. **Recibir secuencia de ítems (*Receive sequence of items*):** Esta tarea ocurre en escenarios donde una secuencia relacionada con ítems es recomendada al usuario. Típicamente se da en dominios de educación o entretenimiento. Por ejemplo: Los sistemas de educación muestran una secuencia de temas que el usuario debe estudiar o aprender (antes de aprender a multiplicar, es recomendable aprender a sumar y restar) [43].

2. **Decisiones al problema (*problem decision*):** Hace referencia a como elegir uno o varios ítems del conjunto de ítems candidatos para la recomendación [44]:
    a. **Elección (*choice*):** Supone elegir un elemento del conjunto de posibles candidatos. Es el caso más simple, de un modo u otro, se elige un ítem del grupo de ítems candidatos.
    b. **Ordenación (*sort*):** Se muestran los ítems en forma de clasificación por categorías. En el caso de una librería se clasifican los libros que se pretenden recomendar en diferentes categorías.
    c. **Valoración (*ranking*):** Se presenta una lista de ítems ordenados en función de su valoración. Por ejemplo, la lista de los libros más vendidos.
    d. **Descripción (*description*):** Supone describir todos los ítems en función de su comportamiento.

2.3.1.2 Enfoque de aproximación

El enfoque de aproximación se refiere a la manera en que las diferentes partes del sistema de recomendación se adaptan a las necesidades o preferencias de cada usuario. Se pueden distinguir las siguientes categorías:

1. **Modelo de usuario (*user model*):** Es la forma en que las características, preferencias o datos del usuario son tratados dentro del sistema de recomendación. Se distinguen tres sub apartados:
    a. **Generación (*generation*):** Es la fase inicial que se da al entrar un nuevo usuario al sistema. Para ello, el sistema necesita de cierta información sobre el usuario (modelo de usuario) para poder procesar las recomendaciones en base a sus preferencias. Esta creación del modelo del usuario puede ser realizada de diferentes formas: (A) El sistema va adquiriendo datos del usuario mientras éste trabaja con él. (B) El usuario provee al sistema de esta información directamente mediante el uso de formularios o cuestionarios. (C) El usuario se relaciona con un estereotipo que lo define [40].
    b. **Actualización (*update*):** Dado que las personas van cambiando de hábitos y gustos a lo largo del tiempo, la actualización de sus preferencias o modelo de usuario es un aspecto a tener en cuenta. Esta se puede dar de forma (A) implícita [45,46]: El usuario no interviene directamente en la valoración de un ítem, sino que el sistema monitoriza al usuario y a partir de sus acciones deduce cierta información. Por ejemplo, la cantidad de veces que se escucha una canción, el tiempo que transcurre leyendo un artículo o los links visitados son aspectos que al inferir en ellos, se pueden sacar valoraciones sobre los contenidos relacionados; O de forma (B) explícita [47,46]: El usuario aporta directamente información sobre su valoración personal del ítem. Por ejemplo, al final de una noticia, el usuario puede valorar si ésta ha sido de su agrado: o (C) híbrida: Cuando se toman en cuenta ambos factores, implícito y explícito, de recogida de datos. Las técnicas usadas incluyen: (I) que el usuario cambia directamente los

datos de un modelo, (II) el sistema va añadiendo información al modelo mediante uso del usuario del sistema o bien el sistema puede (III) gradualmente ir eliminado la información del modelo más antiguo y reemplazarla por otra más nueva [30,48].

c. **Representación (*representation*):** Determina la manera en que los modelos de usuario son representados en el sistema. Diferentes posibilidades son: Modelos basados en la historia [5]; vector espacial [49], redes semánticas y asociativas [50]; modelos basados en clasificadores [19] y ontologías [51].

2. **Modelo del dominio (*domain model*):** Del mismo modo que la información del usuario es representada en el "modelo del usuario", la información relativa a los ítems es representada en el "modelo del dominio". Al igual que el "modelo del usuario", se pueden distinguir dos sub-apartados (el sub-apartado "actualización" no se contempla en este caso dado que los ítems no varían con el paso del tiempo, es decir, mientras que un usuario va progresivamente cambiando de hábitos o preferencias, un ítem no difiere de aspecto o modifica sus propiedades con el transcurso del tiempo) :

   a. Generación: Para la generación de información referente a los ítems se utilizan diferentes técnicas dependiendo del contexto y tipo de ítem. Las diferentes formas de establecer información de un determinado ítem pueden ser: (A) Manual, donde cada ítem es etiquetado por uno o varios usuarios; o (B) automático, en donde procesos computacionales, a partir del estudio del ítem, determinan atributos del mismo. En el apartado "3 Sistemas basados en el contenido", se explican diferentes formas automáticas de generación de información.

   b. Representación: Las diferentes formas de representar los ítems son: (A) mediante listas indexadas (todos los elementos del dominio se encuentran al mimo nivel) [52]; mediante la utilización de (B) taxonomías de ítems [53]; o finalmente mediante la utilización de (C) ontologías donde se definen relaciones más complejas entre los diferentes ítems del dominio [51].

3. **Personalización (*personalization*):** Hace referencia a la forma en que el sistema describe la información de sus recomendaciones. Esta puede darse de cuatro formas distintas:
    a. **Nivel (*degree*):** Dependiendo del nivel de personalización del sistema de recomendación, éste se puede clasificar en: (A) No personalizable, es decir, el sistema de recomendación es el mismo para todos los usuarios [54]; (B) Recomendaciones precarias que están basadas en pocas preferencias del usuario [47] y (C) recomendaciones persistentes en donde las preferencias de los usuarios se tienen en cuenta [5].
    b. **Método (*method*):** Se corresponde a la forma general de clasificación de los sistemas de recomendación expuestos anteriormente en "2.3 Clasificación". Además, también se pueden añadir, en menor medida, los siguientes: (A) Las recomendaciones de ítems son parte de un proceso de búsqueda en donde no hay ninguna personalización en el proceso de recomendación. (B) Selección manual de las recomendaciones. Éste es el caso donde intervienen expertos en el tema o también cuando se dan las opiniones de líderes o famosos [55].
    c. **Algoritmo (*algorithm*):** Hace referencia a la manera de inferir en el conocimiento para realizar las recomendaciones. Se pueden distinguir dos vertientes:
        i. **Según el tipo:** En este apartado se encuentran los basados en: (A) modelo (ver "4.2.1.2 Basados en la memoria" para más información): Sistemas que no utilizan toda la información disponible, sino que crean modelos estadísticos a partir de una cierta cantidad de datos. Esto les permite procesar recomendaciones de forma más rápida pero comprometiendo la precisión del resultado [17,46,56]. (B) Memoria o heurísticos (ver "4.2.1.1 Basados en la memoria" para más información): Utilizan todos los datos disponibles para generar las recomendaciones. Con ello, los sistemas de recomendación generan resultados más precisos, pero requieren de mayor tiempo de proceso [3,57-59]. Los sistemas de recomendación de

filtro colaborativo son típicamente divididos en estos dos subgrupos. En el apartado “4.2.1 Clasificación” del tema “4 Sistemas de filtrado colaborativo” se extiende la explicación sobre estos conceptos.

ii. **Según la técnica:** Los diferentes algoritmos que existen dependiendo de la técnica son [14]: (A) Basados en atributos: Se trata de los sistemas en donde las recomendaciones se basan en las propiedades de los ítems y los intereses de los usuarios. Para entenderlos mejor, se puede decir que el sistema más simple de este tipo es un "buscador". Por ejemplo, si una persona está mirando la sección de libros de terror, el sistema de recomendación le sugerirá un libro de terror. (B) Correlación de ítem a ítem: El sistema identifica ítems que se pueden asociar a otros ítems por los que el usuario haya expresado interés. Por ejemplo, en un supermercado, el hecho de comprar una caja de cereales, puede estar asociado a la compra de leche. (C) De usuario a usuario: Las recomendaciones vienen determinadas por el parentesco entre usuarios. El principio de este modelo es el siguiente: Si a diversos usuarios de una comunidad les gusta el último álbum de un artista en concreto, es muy probable que al resto de usuarios de esa comunidad también les guste ese mismo disco.

d. **Resultado (*output*):** Típicamente el resultado de un sistema de recomendación es una (A) sugerencia (por ejemplo, “pruebe este ítem”); (B) valoraciones que otros usuarios establecen sobre un ítem en concreto o (C) predicciones de las valoraciones que un usuario realizará de las recomendaciones presentadas por el sistema.

2.3.1.3 Operaciones

En este apartado se encuentran aspectos relacionados con las operaciones de los sistemas de recomendación, es decir, donde se realizan las recomendaciones o quién

realiza el primer paso para empezar la ejecución del sistema recomendación. La categoría está distribuida en tres apartados:

1. **Arquitectura (*architecture*):** Dependiendo del diseño del sistema, se pueden considerar los siguientes casos:
    a. **Centralizados:** Toda la información perteneciente al sistema de recomendación se encuentra en un mismo sitio o servidor.
    b. **Distribuidos:** Cuando la información del sistema de recomendación está distribuida en diferentes lugares [60]. Los sistemas distribuidos son muy frecuentes entre los sistemas de recomendación semánticos (ver "6 Sistemas de recomendación semánticos").
2. **Localización (*location*):** Hace referencia a donde (lugar o máquina) las recomendaciones son producidas y mostradas. Se distinguen tres casos:
    a. En la fuente de información: Este es el caso más habitual de los sistemas de recomendación. Tanto el perfil del usuario, como el proceso de las recomendaciones, se realizan en el mismo sitio donde se encuentra la fuente de información [41].
    b. En un servidor de recomendación: Las recomendaciones se realizan por terceras partes, refiriéndose éstas a diferentes fuentes de información externas. Por ejemplo, cuando un restaurante es recomendado por los usuarios interesados de un sistema de recomendación independiente [5,47,58].
    c. En el lugar del usuario: El perfil del usuario se almacena en el ordenador del usuario y las recomendaciones se procesan en el mismo ordenador. El filtrado de correos electrónicos es un típico caso de esta categoría.
3. **Modo (*mode*):** Indica quién inicia el proceso de recomendación. Se distinguen dos casos:
    - Modo pasivo [61]: Las recomendaciones generadas son válidas para todos los usuarios del sistema. Por ejemplo, en el comercio electrónico, mostrar los 10 ítems más valorados como parte regular del sistema operacional.

- Modo activo [62,63]: Las recomendaciones se establecen teniendo en cuenta el historial de recomendaciones de los usuarios. Se pueden distinguir dos grupos [64]:
    - Modelo "*Push*" de información. Se da cuando una consulta al sistema es definida de forma implícita a través de las preferencias definidas en el perfil del usuario [5].
    - Modelo "*Pull*" de información. Se da en el momento en que el usuario realiza una consulta al sistema (de forma explícita) para recibir la información o recomendación [23,58].

2.4 Problemas generales

Existen una serie de problemas que son de ámbito general e independientes del sistema de recomendación que se utilice [65]:

- **Carencia de información (*Lack of data*):** Los sistemas de recomendación necesitan de mucha información para hacer efectivas sus recomendaciones. Así pues, para el funcionamiento de los algoritmos de recomendación, se necesita información de los factores que intervienen en las recomendaciones, es decir, productos (ítems) y usuarios. No siempre es necesario saber de ambos componentes, pero sí como mínimo de uno de ellos. Cuanta más información, más precisa será la recomendación. Este aspecto evoca al concepto de economía de escala, es decir, cuanto más grande sea el conjunto de usuarios o ítems con información, más probabilidades habrá de encontrar coincidencias entre perfiles o gustos de un usuario en concreto.
- **Información cambiante (*Changing data*):** Hace referencia a los contenidos que su uso, modo o costumbre, están en alza durante algún tiempo, o en una determinada región, es decir, productos "de moda" o "*fashion*". Estos productos representan preferencias de usuarios en espacios de tiempo muy puntuales. El problema de este tipo de ítems radica en que cuando un determinado ítem está "de moda", recibe gran cantidad de valoraciones muy positivas. Por otro lado, cuando se termina el período de tiempo en que el ítem está "de moda", los usuarios no desean recomendaciones del ítem; no

obstante, el sistema de recomendación los sigue recomendado ya que tiene una gran cantidad de votos positivos del ítem (que fueron realizados en el periodo de tiempo en que el ítem estaba "de moda"). Este es un problema habitual en el mundo de la ropa. Por ejemplo, las tendencias de ropa de un año a otro varían completamente. En una temporada se pueden llevar los pantalones de pitillo, mientras que para la temporada siguiente se llevan los pantalones acampanados.

- **Cambio de preferencias de usuario (*Changing user preferences*):** Usualmente los usuarios buscan recomendaciones para ellos mismos, pero eventualmente, estos mismos usuarios pueden buscar recomendaciones para otros usuarios de diferente perfil. Por ejemplo: Un usuario determinado "A", del cual se tiene establecido un perfil con sus preferencias y gustos, le puede interesar buscar un cierto ítem para una persona "B" (por ejemplo para hacer un regalo de cumpleaños) con unas preferencias y gustos totalmente diferentes a los suyos.
- **Ítems impredecibles (*Unpredictable items*):** Existen una serie de ítems que es difícil recomendarlos ya que la reacción del usuario hacia ellos puede ser diversa e impredecible. Por ejemplo, éste es el caso de la película "Napoleon Dynamite". Para la gente, este ítem tiende a ser o muy buena recomendación o muy mala, pero no un término medio.
- **Voto pronto y a menudo (*Vote early and often*):** Según comenta Resnick y Varian en [6], si cualquiera puede hacer recomendaciones, los propietarios de un ítem pueden hacer recomendaciones positivas de ese ítem y recomendaciones negativas para los ítems de sus competidores. Este aspecto se puede considerar como un ataque a los sistemas de recomendación ya que se intenta sesgar las recomendaciones de cierto(s) ítem(s). En el apartado "8 Seguridad en los sistemas de recomendación" se realiza un estudio más exhaustivo de aspectos relacionados con la ataques a los sistemas de recomendación.
- **Complejidad computacional (*Computer complexity*):** Los sistemas de recomendación tienen una mayor exactitud de sus predicciones cuando cuentan con una mayor cantidad de información disponible. No obstante, a mayor cantidad de información en el sistema, más costosos (en tiempo) son los

cálculos para obtener los resultados. Una de las formas más habituales para reducir estos cálculos es el proceso off-line de los datos, es decir, en el momento de realizar una recomendación, ya se tienen datos que han sido calculados posteriormente (de forma off-line). Otra manera de abordar el problema consiste en la creación de grupos o “clusters” de elementos parecidos de tal forma que los elementos similares se agrupan y se computan como si fueran el mismo.

3 Sistemas basados en el contenido

En este capítulo se explican los sistemas de recomendación basados en el contenido (*content based recommender systems*). Su modus operandi es captado en el apartado "3.1 Funcionamiento". Los problemas generales que presenta los sistemas basados en el contenido están enumerados en la sección "3.2 Problemas generales". Para finalizar, en el apartado "3.3 Tendencias generales" se muestra la directriz actual de estos sistemas de recomendación.

3.1 Funcionamiento

Los sistemas de recomendación basados en el contenido generan recomendaciones teniendo en cuenta la valoración que el usuario ha hecho de los ítems en el pasado. Así pues, el sistema tiene en cuenta las valoraciones que el usuario realiza de los ítems y las utiliza para recomendar nuevos ítems que tienen características similares a los ítems evaluados positivamente por el usuario.

	Pedro	Juan	Carlos
Saw IV (terror)	Ø	3	2
El fujitivo (acción)	5	Ø	3
Amor eterno (romántica)	2	Ø	Ø
La gran evasión (acción)	8	9	Ø
Los otros (terror)	9	3	2
Titanic (romántica)	3	2	8
El exorcista (terror)	10	1	6

Tabla 1 Ejemplo recomendación basada en el contenido

Por ejemplo, tómese como contexto el alquiler de películas. En la siguiente tabla "Tabla 1 Ejemplo recomendación basada en el contenido" se muestra una relación entre la valoración de diferentes usuarios para diferentes películas (ítems). Las valoraciones van desde uno (peor puntuación posible) hasta diez (máxima puntuación posible). Además, en la tabla aparece el símbolo "Ø", el cual indica que el correspondiente ítem no ha sido valorado por el usuario en concreto. Por otra parte, al lado de cada película (entre paréntesis) apare el género al que pertenece el cual será utilizado para establecer las recomendaciones, es decir, en este caso las

recomendaciones basadas en el contenido serán determinadas por el atributo "género" de los ítems (películas).

Teniendo en cuenta los datos de la tabla anterior, se observa que Pedro no ha realizado ninguna valoración para la película Saw IV, que forma parte del género de terror. El sistema de recomendación basado en el contenido determina, teniendo en cuenta el historial de valoraciones del usuario uno de las películas de terror, que sería apropiado recomendarle esta película ya que las valoraciones de las otras películas de terror del sistema (Los otros y el exorcista) han sido muy positivas (valoradas con 9 y 10 respectivamente). El mismo razonamiento puede ser aplicado para Juan y Carlos, siendo el resultado esperado que Juan prefiera películas de acción mientras que Carlos prefiera las películas de género romántico.

3.1.1 Representación de los ítems

Para poder inferir sobre el conocimiento (atributos o cualidades de los ítems) y realizar recomendaciones, es necesario almacenar de un modo u otro las características de cada ítem. Por lo general, este tipo de sistemas se suelen basar en el uso de una estructura bien definida para la anotación de las características y los valores de los correspondientes ítems. Por ejemplo, en un sistema de recomendación de viajes de vacaciones una posible estructura para representar las características de los viajes podría constar de campos como: precio, duración, localización, modo de transporte, tipo de alojamiento, etc. No obstante, no siempre es posible contar con una estructura de características y valores bien definida (por ejemplo en el caso de una crítica de cine o un artículo sobre un restaurante nuevo) por lo que se han propuesto diferentes algoritmos para analizar el contenido de los documentos (sin estructura o de texto plano) y encontrar regularidades en el contenido del documento que pueden servir como base para realizar recomendaciones.

Una primera aproximación al análisis de texto sin formato es ver cada palabra como un atributo de tipo boleano o como un número entero, que se va incrementando cada vez que aparece la palabra en el documento y así se obtiene un listado de las palabras más nombradas en un documento [66]. El mecanismo es perfeccionado mediante el uso del algoritmo de "Stemming" [67]. Este algoritmo trata de concebir las palabras con la

misma raíz o lexema como la misma palabra, por ejemplo: marinero y marino son dos palabras distintas que comparten el mismo lexema, por lo tanto, ambas palabras, en relación al algoritmo de "Stemming", serían la misma. La importancia de una palabra en un documento es determinada con un peso que puede ser definido de diferentes formas. En [68] se proponen diferentes algoritmos para el análisis de documentos no estructurados. Uno de los métodos más conocidos para asignar un peso a las palabras de un documento es el "term frequency–inverse document frequency" (TF-IDF) [68].

El sistema FAB [23] es un ejemplo que representa los documentos con las 100 palabras que contengan el índice TF-IDF más alto. En el caso de "Syskill & Webert" [69] representan los documentos con las 128 palabras más informativas, por ejemplo, las palabras que están más asociadas a un tipo de documento que otro.

EL problema de este tipo de representación es que no se tiene en cuenta el contexto. Por ejemplo, si en la descripción de un computador aparece la frase: "El ordenador es apto para diseñadores, pero no para jugadores", el sistema podría entender, si no se tiene en cuenta el contexto, que al aparecer la palabra "diseñador" y "jugador", el ordenador es apto para los dos tipos de usuarios: diseñadores y jugadores.

3.1.2 Perfil de usuario

El perfil del usuario es utilizado en los sistemas de recomendación basados en el contenido para almacenar información referente al usuario. Esta información puede ser proporcionada directamente por el usuario de forma explícita (por ejemplo, mediante un formulario) o deducida por el sistema (por ejemplo, mediante el feedback de los ítems que el usuario proporciona al sistema). El tipo de información que se puede contener en el perfil del usuario es:

1. **Preferencias del usuario:** Contiene una relación de los tipos de ítems en que el usuario está interesado (o que no está interesado).
2. **Historial del usuario:** Se almacenan las interacciones que el usuario ha tenido con el sistema. Por ejemplo, la valoración de un ítem. Dicho aspecto es importante ya que el sistema puede utilizarlo para aprender gustos y preferencias del usuario.

3.1.3 Inferencia sobre el conocimiento

La obtención de datos es necesaria en el sistema de recomendaciones basada en el contenido pero no sirve de nada sin un mecanismo o algoritmo que los procese y extraiga conclusiones o, en este caso, recomendaciones. Estos algoritmos crean una función que modela los intereses de los usuarios. Por consiguiente, dado un modelo de usuario y un ítem nuevo, la función predice si el ítem será del agrado del usuario. A continuación se presentan diferentes formas de inferir en el conocimiento almacenado en el sistema:

- **Personalización del usuario:** El sistema proporciona una interface donde el usuario puede aportar las preferencias de sus gustos. Una vez que el usuario ha proporcionado sus preferencias, el sistema simplemente busca los elementos que encajan con los parámetros indicados por el usuario. El método implica que el usuario debe implicarse y rellenar un formulario o cuestionario para que el sistema pueda realizar las recomendaciones. Además, e método no provee de ninguna forma para determinar el orden en que se presentan los resultados obtenidos por el sistema.
- **Basadas en reglas (*Rule based*)**: El sistema tiene una regla para recomendar productos basados en su historial. Por ejemplo, si el usuario ha adquirido los capitulos dos y tres de una determinada serie, el sistema recomendaría el capitulo cuatro de esa serie.
- **Reglas de asociación (*Associative rules*):** Se intenta descubrir relaciones entre los ítems, que luego pueden ser utilizadas para recomendación. Los ítems se comparan en base al comportamiento pasado de los usuarios respecto a ellos. Por ejemplo, la compra de cereales puede relacionarse con la compra de leche. En [70] muestra la utilización del método por parte de "*Amazon*".
- **Árboles de decisión (*Decision tree*):** Los árboles de decisión se basan en una estructura en forma de árbol la cual reparte u organiza la información a lo largo de ésta. Las tres partes de que consta la misma son: nodos (contienen atributos), Arcos (contienen valores posibles del nodo padre) y Hojas (Nodos que clasifican el ejemplo). Para elaborar recomendaciones, los algoritmos parten del nodo padre y evalúan los atributos de éste para seleccionar un arco.

Este paso se repite iterativamente hasta alcanzar una hoja, la cual, es el resultado del algoritmo que identifica las recomendaciones posibles a tener en cuenta. El método es recomendable en los casos donde haya un número reducido de atributos [66]. Un ejemplo de árboles de decisión es el algoritmo de ID3 [71]. En [72] se muestra un ejemplo del sistema de recomendación basado en árboles de decisión.

- **Método del vecino más cercano:** Permite clasificar un nuevo dato basándose en observaciones conocidas o pasadas. En el contexto de sistemas de recomendación basados en el contenido, se usa para determinar valores de un ítem sin valorar. Para ello, el nuevo ítem, que carece de valoración, se compara (mediante el uso de los atributos del ítem) con los otros ítems ya valorados. Mediante una función de similitud, se obtienen los vecinos más cercanos al nuevo ítem. El uso de las etiquetas de los vecinos más cercanos, se utiliza para obtener o derivar una(s) etiqueta(s) para el nuevo ítem. Diferentes funciones son utilizadas para establecer el vecino más cercano dependiendo del tipo o estructura de datos con la que se trabaja. Las más usadas son: Distancia euclidiana [73] y modelo de espacio vectorial [74]. Diferentes ejemplos se encuentran en [75].
- **Feedback relevante:** Método que propone obtener información sobre los ítems recomendados al usuario cuando éste ha realizado una búsqueda. Por ejemplo, páginas web que ofrecen ayuda o soporte de sus programas, piden al usuario que valore la información aportada por el sistema conforme a lo que estaba buscando. Mediante la apreciación del usuario, el sistema conoce qué información ha sido útil para la búsqueda que se ha realizado y la tiene en cuenta para posteriores búesquedas, iguales o relacionadas, a la valorada por el usuario. El algoritmo de "*Rocchio*" [76] ha sido ampliamente utilizado en este contexto. En [77] se muestra un ejemplo del feedback relevante.
- **"Clústering":** Dado que muchas veces es una tarea costosa tanto en tiempo de cálculo como de requisitos, el obtener una recomendación para un individuo en concreto, una alternativa es agrupar usuarios en categorías en base a comportamientos pasados de otros consumidores. Cada grupo o "*cluster*" tiene preferencias que son típicas en base a sus miembros. Los usuarios dentro de

cada grupo recibirán recomendaciones calculadas a nivel grupal. La ventaja del "*clustering*" es que trabaja sobre los datos agregados, es decir, más eficientes. Esta técnica se suele usar como un primer paso para reducir las búsquedas del filtrado colaborativo. Por otro lado las recomendaciones (a nivel grupal) son menos relevantes o precisas que en el filtrado colaborativo (a nivel individual).

- **Redes neuronales artificiales (*Artificial neural network*):** Las redes neuronales simulan las propiedades observadas en los sistemas neuronales de animales a través de modelos matemáticos recreados mediante mecanismos artificiales. El objetivo es conseguir que las máquinas (o algoritmos) den respuestas similares a las que es capaz de dar el cerebro que se caracterizan por su generalización y su robustez. Las redes neuronales están compuestas de muchos elementos sencillos que operan en paralelo. El diseño de la red mayoritariamente viene determinado por las conexiones entre sus elementos (igual que las conexiones de las neuronas cerebrales); en otras palabras, las redes neuronales representan funciones usando redes de elementos con cálculo aritmético sencillo y utilizan métodos para aprender esa representación a partir de ejemplos. La gran diferencia del empleo de las redes neuronales en relación con otras aplicaciones de la computación, radica en que no son algorítmicas, esto es que no se programan haciéndoles seguir una secuencia predefinida de instrucciones sino que ellas mismas generan sus propias "reglas" para asociar la respuesta a su entrada; es decir, aprende por ejemplos y de sus propios errores.
- **Bayesiano naive (*Naïve Bayes*):** Método de clasificación probabilístico. Se utiliza para clasificar un nuevo ítem dentro de las preferencias del usuario. Por ejemplo, dado un nuevo ítem y un conjunto de usuarios, se calcula la probabilidad de que el ítem tenga cabida dentro de cada grupo de usuarios dependiendo de sus preferencias. Es un método importante ya que no sólo ofrece un análisis cualitativo de los atributos y valores que pueden intervenir en el problema, sino que además da a entender la importancia cuantitativa de esos atributos. El aspecto cualitativo permite representar cómo se relacionan esos atributos (de forma causal o señalando la correlación que existe entre esos atributos). De forma cuantitativa dan una medida probabilística de la

importancia de las variables en el problema. Esta es una de las diferencias de las redes bayesianas con respecto a los métodos de árboles de decisión y las redes neuronales. Por último, cabe destacar que el aprendizaje basado en redes bayesianas es especialmente adecuado en entornos donde es necesaria la clasificación de textos [78].

3.2 Problemas generales

Los sistemas de recomendación basados en el contenido presentan los siguientes inconvenientes:

- **Limitación por el análisis del contenido:** El sistema se ve limitado por las propiedades que están explícitamente asociadas al ítem. Por consiguiente, es necesario tener las suficientes propiedades de cada objeto para un correcto funcionamiento. Mientras que para ciertos tipos de ítems (por ejemplo, libros, revistas, artículos) es relativamente fácil la extracción de información de forma automática. Existen otros ítems, por ejemplo los contenidos multimedia, donde la asignación de propiedades debe hacerse de forma manual dada la dificultad de hacerlo de forma automática. Además, otro problema relacionado con la limitación del análisis del contenido, aparece cuando dos ítems diferentes son representados por el mismo conjunto de características, en cuyo caso, ambos ítems son indistinguibles entre sí ya que presentan las mismas características o propiedades.
- **Sobre-especialización:** La sobre-especialización aparece cuando el sistema sólo muestra al usuario ítems similares a los que ya ha visto antes. Por ejemplo, a una persona que nunca haya tenido una experiencia con restaurantes de comida japonesa, nunca se le recomendará un restaurante japonés.
- **Un usuario nuevo:** Un usuario tiene que haber valorado algunos ítems antes de que el sistema pueda saber de sus gustos y preferencias para poderle recomendar.

3.3 Tendencias generales

En los últimos años y especialmente a finales de los años 90, con la aparición de internet, la cantidad o volumen de información a la que los usuarios tienen acceso ha

ido en aumento de forma vertiginosa. Este hecho conlleva un mayor número de ítems y, por lo tanto, una gran cantidad de atributos que deben ser asignados a cada ítem.

Dependiendo del entorno es factible la asignación manual de atributos a los ítems, no obstante, en la mayoría de situaciones, el trabajo manual es inapropiado e imposible. Por consiguiente, se ha manifestado una tendencia hacia la automatización del proceso de asignación de atributos a los ítems para recomendar. Para paliar la situación, se están usando técnicas como TF-IDF [68] para poder realizar valoraciones de atributos de documentos de forma automática. Este método es válido en contextos limitados, donde el mismo ítem es el propio texto. Por ejemplo: noticias, papers, libros, artículos, revistas, etc… Por otro lado, la utilización de mecanismos o formas automatizadas de asignación de atributos a los ítems en otros entornos como por ejemplo, el multimedia, es muy pobre o poco desarrollada. No obstante se han realizado diversos trabajos para, por ejemplo, poder extraer información o atributos en el ámbito multimedia. Éste es el caso de [79,80], que de modo análogo a un texto, permite analizar una canción y extraer de la misma características, es decir, atributos. En los casos de ámbito multimedia, es común el uso de atributos como el autor, género u otros atributos relacionados con el tema.

Por otra parte, existen ítems que su naturaleza no permite la obtención de datos sobre éstos. Es el caso de chistes o poemas, donde la escasa información implícita del ítem (por ejemplo, la longitud o cantidad de palabras de un chiste suele ser muy pequeña) no permite extraer atributos sobre éste de forma automática. Así pues, el uso de puros sistemas de recomendación basados en el contenido es insuficiente para realizar buenas recomendaciones. Estos sistemas deben ser concedidos como una herramienta de filtrado extra sobre los resultados o recomendaciones obtenidas de otros sistemas de recomendación, como pueden ser los basados en el filtrado colaborativo.

Como conclusión a lo expuesto anteriormente se puede deducir que:

- Los sistemas de recomendación basados en el contenido no son una solución en sí misma al problema de las recomendaciones, sino más bien sirven de ayuda o complemento a otros sistemas de recomendación. En el apartado “7 Sistemas híbridos” se muestran diferentes formas de combinar los sistemas de

recomendación para reducir los problemas de los mismos. Como el capítulo explica, es habitual la combinación de los sistemas de recomendación basados en el contenido con los sistemas de filtrado colaborativo (ver “4 Sistemas de filtrado colaborativo”).

- Existe una tendencia a generar atributos o meta-información sobre los ítems de forma automática o semi-automática dado la creciente cantidad de información e ítems que hay hoy en día.

4 Sistemas de filtrado colaborativo

A continuación se describen los sistemas de filtrado colaborativo. La sección empieza con una explicación sobre sus orígenes, "4.1 Historia", para a continuación, en el apartado "4.2 Definición y funcionamiento", dar a conocer su definición y su forma de proceder para realizar las recomendaciones. Además se realiza una clasificación de este tipo de sistemas de recomendación. En el apartado "4.3 Dominios afines del sistema", se consideran las características propias de los sistemas de filtro colaborativo para determinar qué dominios son más convenientes para su aplicación. En el apartado "4.4 Problemas", se dan a conocer, en el ámbito general, sus inconvenientes o limitacions. Por último, en el apartado "4.5 Tendencias generales", se establecen las directrices que están siendo investigadas hoy en día.

4.1 Historia

Los sistemas de recomendación basados en el filtrado colaborativo aparecen después de los sistemas de recomendación basados en el contenido (ver "3 Sistemas basados en el contenido"). En un principio los sistemas de recomendación basados en el contenido fueron de gran utilidad para una primera aproximación de filtrado de información, pero presentaban un inconveniente: No podían medir la calidad de las recomendaciones que realizaban. Frente a este problema, a principios de los noventa aparecieron dos posibles alternativas:

a. Esperar que técnicas de inteligencia artificial mejorasen la clasificación automática de los documentos o ítems.
b. Introducir la opinión de las personas en el proceso de las recomendaciones.

Con la adopción del punto "a", se mejoraban los algoritmos de recomendación basados en el contenido, mientras que el desarrollo del punto "b" produjo la creación de sistemas de recomendación social o de filtrado colaborativo.

El primer sistema que permitió añadir opiniones de los usuarios fue el llamado "*Tapestry*" [5]. El sistema permite almacenar opiniones o anotaciones de los usuarios sobre los contenidos de mensajes como un tipo de meta información. Por su parte, el

sistema brindaba a los usuarios la posibilidad de realizar búsquedas sobre el contenido de un documento así como de la meta información producida por los usuarios. Otro de los primeros desarrollos de este tipo de sistema, el cual utilizó “estereotipos” como mecanismo para construir modelos de usuario, fue el llamado “Grundy” [81]. Usando estereotipos, el sistema “Grundy” era capaz de crear modelos de usuario individuales y usarlos para recomendar libros a cada usuario. Más adelante aparecieron los primeros sistemas de recomendación colaborativa en usar algoritmos de autómata de predicción. GroupLens [3] y Ringo [29] fueron los primeros ejemplos en utilizar del autómata.

4.2 Definición y funcionamiento

Los sistemas de filtrado colaborativo son sistemas de recomendación que utilizan la información que los usuarios aportan sobre los ítems para realizar las recomendaciones. Así pues, tratan de predecir las valoraciones de un (varios) ítem(s) para un usuario en particular (llamado usuario activo) mediante el uso de ítems valorados por otros usuarios, es decir, teniendo en cuenta la opinión de los otros usuarios. Como consecuencia, estos sistemas se basan en las tendencias marcadas por grupos de personas con los mismos gustos o preferencias que el usuario activo para realizar las recomendaciones o sugerencias. Para un individuo o usuario cuyos gustos o tendencias son similares a los de un grupo de personas determinado, se establece que este usuario pasa a formar parte de este grupo; Por consiguiente, si una o varias persona del grupo valoran un nuevo ítem de forma positiva, el sistema de filtrado colaborativo establece que es muy probable que los demás miembros del grupo al que pertenece(n), también valoren positivamente el nuevo ítem.

Los sistemas de filtrado colaborativo, por lo general, se adaptan al siguiente funcionamiento:

- Los usuarios elaboran valoraciones sobre ítems.
- Se analizan las valoraciones realizadas por los usuarios con tal de establecer grupos o vecinos cercanos de usuarios con preferencias similares.

- Una vez obtenidos los grupos de vecinos para un usuario, se realizan las recomendaciones al usuario activo teniendo en cuenta sus vecinos más cercanos y sus correspondientes valoraciones.

Por ejemplo, la siguiente tabla ("Tabla 2 Ejemplo filtro colaborativo") muestra los valores de diferentes usuarios para diferentes ítems. Las valoraciones se comprenden entre los valores uno al diez, siendo diez la máxima puntuación que puede percibir un ítem. Además, en la tabla aparece el símbolo nulo "Ø" para poder indicar la ausencia de valoraciones del ítem correspondiente por parte del usuario.

	Lenguado	Merluza	Perca	Rape	Marisco	Dorada
Pedro	2	6	Ø	10	8	Ø
Juan	7	4	9	Ø	Ø	2
María	3	5	5	9	8	9
Andrés	Ø	9	Ø	Ø	2	6
Raúl	9	8	2	2	1	5

Tabla 2 Ejemplo filtro colaborativo

Teniendo en cuenta la tabla anterior, se fija como usuario activo a Pedro. Se quiere establecer si la perca (ítem que no ha sido valorado) es un pescado candidato a recomendar. Para ello, se buscan usuarios de la tabla que se asemejen en preferencias a Pedro, es decir, tengan las mismas valoraciones para los mismos ítems, en este caso, pescados. En el ejemplo, se observa que María muestra una gran coincidencia con Pedro, el usuario activo (por lo que ambos pasan a formar parte de un grupo de vecinos cercanos), además de que María ha valorado la perca. Con la siguiente información se puede establecer que si se le recomienda a Pedro una perca, existe una gran probabilidad de que la perca sea valorada (por parte del usuario activo, es decir, Pedro) con una puntuación que ronda el valor de cinco (valoración del usuario vecino, María). Por lo tanto, la perca no sería una opción a recomendar a Pedro ya que presenta una valoración de su grupo de vecinos cercanos muy pobre. No obstante, el efecto es el contrario si se toma como posible candidato a recomendar la dorada.

4.2.1Clasificación

Típicamente los sistemas de filtrado colaborativo se han clasificado en dos tipos [56]:

- **Basados en la memoria (o heurísticos):** Son algoritmos que realizan sus predicciones teniendo en cuenta todos los elementos evaluados, es decir, utilizan toda la información disponible en el sistema.
- **Basado en el modelo:** Son un tipo de sistemas que usa un conjunto de valoraciones para crear un "modelo". Este modelo es utilizado a posteriori para crear predicciones de valoraciones.

El principal problema de los métodos heurísticos es la escalabilidad debido a la gran cantidad de información a procesar. Como consecuencia al problema de la escalabilidad, se utilizan técnicas para reducir el tiempo de proceso de la información como es el cálculo off-line de diferentes variables o constantes. Por otra parte, los algoritmos basados en el modelo utilizan conceptos probabilísticos que simulan la información presente en el sistema, pero que no es utilizada directamente. A continuación se describen diferentes técnicas usadas en los dos sistemas para alcanzar su objetivo.

4.2.1.1 Basados en la memoria

La predicción de votos de usuario activo en los sistemas de filtrado colaborativo basados en memoria se fundamenta en la medida de un conjunto de pesos calculado mediante el uso de la base de datos de votaciones. Los pesos pueden representar distancia, correlación o similitud entre el usuario activo y los diferentes usuarios de la base de datos. La distinción entre los diferentes sistemas de filtrado colaborativo basado en memoria viene determina por la forma de calcular estos pesos [56].

En los casos más simples se utilizan conjuntos cuya medida es representada por un promedio o promedio ponderado. Estas metodologías asumen que las valoraciones de los ítems utilizan la misma escala, en caso contrario, sería necesario utilizar medidas correctoras antes de aplicar las técnicas mencionadas anteriormente. Otra medida utilizada para establecer la fuerza y/o la dirección de una relación lineal entre dos valoraciones de un ítem es la de correlación de dos variables. Para finalizar, el uso de vectores de similitud está también presente en el contexto de sistemas de filtro colaborativo.

Por otro lado, se han propuesto y analizado diferentes ampliaciones o mejoras que incrementan la calidad de los resultados obtenidos por los resultados de los algoritmos anteriormente expuestos[56]. A destacar:

- **Voto por defecto (*default voting*):** Se aplica en el caso de los algoritmos de correlación en donde existen pocos votos, tanto por parte del usuario activo, como de los otros usuarios. Se ha demostrado empíricamente que la precisión de los resultados (en los algoritmos de correlación) es mejorada si se asumen valoraciones por defecto en los ítems no valorados [56].
- **Frecuencia inversa del usuario (*inverse user frequency*):** La idea es aplicar el mismo principio que "*inverse document frequency*" (IDF) [68] a los usuarios de la base de datos. Es decir, reducir los pesos para las valoraciones de los ítems valorados más frecuentemente con el fin de identificar los ítems más relevantes. El sentido de este método es que los ítems de gusto general no capturen la similitud entre usuarios.
- **Amplificación de casos (*case amplification*):** Se trata de enfatizar los pesos que están más cerca los unos de los otros y al mismo tiempo castigar los pesos menores. Para ello, se aumentan de forma proporcional los pesos de las valoraciones próximas y se les resta valor a las valoraciones que se encuentran alejadas.

4.2.1.2 Basados en el modelo

Los algoritmos, basados en el modelo, tienen un enfoque probabilístico. Los diferentes algoritmos empleados son:

- **"Redes Bayesianas":** Una red Bayesiana se basa en la creación de un grafo el cual relaciona los diferentes nodos de éste de forma probabilística. En el contexto del filtro colaborativo basado en el modelo, cada nodo corresponde a un ítem y su estado es determinado por la posibilidad de que ese nodo/ítem sea votado por los otros ítems del grafo. A este grafo se le aplican algoritmos para encontrar dependencias entre nodos y así poder obtener para cada ítem, un conjunto de ítems padres que corresponden a las mejores predicciones de sus votos.

- **"Clustering":** Se trata el problema colaborativo como si fuera un problema de clasificación por lo que crea grupos de usuarios similares (o "*cluster*") y en lugar de procesar usuarios, procesa "*clusters*". Los algoritmos, en lugar de comparar el usuario activo con todos los usuarios del sistema, comparan el usuario activo con grupos de usuarios, es decir, los clústeres.
- **Redes neuronales artificiales (*Artificial neural network*)** [82]**:** Al igual que en los sistemas de recomendación basados en el contenido, las redes neuronales, también tienen cabida en los sistemas de recomendación de filtrado colaborativo. Ver "Redes neuronales artificiales" en "3.1.3 Inferencia sobre el conocimiento".
- **Redes basadas en grafos (*graph-theoretic approach*)** [83]**:** La idea del método es la formación de un grafo dirigido cuyos nodos corresponden a los usuarios y a sus relaciones se les asigna un peso y dirección. Cuando un usuario predice a otro usuario, se crea una transformación lineal que es construida con las valoraciones "trasladadas" de un usuario a otro. De este modo, la predicción de recomendación de un ítem para un usuario se realiza mediante la media de pesos de las direcciones de las relaciones, antes mencionadas, de los diferentes usuarios.
- **Popularidad del impacto de proximidad (*Proximity Impact Popularity*)** [32]**:** El enfoque, propuesto por Hyung Jun Ahn, intenta solventar el problema del arranque en frío. La popularidad del impacto de proximidad es una medida heurística que se basa en tres factores: (A) La proximidad: Se refiere a la distancia entre las valoraciones y representa la diferencia aritmética entre dos valoraciones. (B) El impacto: Representa la fuerza con la que las valoraciones son del agrado o desagrado de los usuarios y (C) la popularidad que incrementa el valor de las similitudes entre valoraciones que estén lejos de la media (de los ítems ya valorados).

4.3 Dominios afines del sistema

Debido a la naturaleza de los filtros colaborativos, no siempre es posible la aplicación de éstos para realizar recomendaciones. Así como sin la existencia de un grupo de usuarios vecinos al usuario activo no es posible realizar recomendaciones precisas,

existen dominios donde estos sistemas se muestran más propensos para el trabajo con la información ya que presentan mayor efectividad en la precisión de los resultados [84]. A continuación se muestra una clasificación de los diferentes dominios y las características que permiten una mayor (o menor) influencia en la precisión de los resultados para los sistemas de filtrado colaborativo [84]:

- **Distribución de la información:** Se refiere al número y forma que toma la información. Obsérvese que existe una relación directa entre el enfoque del dominio y el problema general "Carencia de información" de los sistemas de recomendación (ver "2.4 Problemas generales"). Los diferentes aspectos que se pueden valorar de la distribución de la información son:
    - Hay muchos ítems: Si hubieran pocos ítems en el sistema, el usuario no tendría la necesidad de utilización de este sistema ya que él mismo podría conocerlos todos y valorar por sí mismo el grado de utilidad que éstos le aportan.
    - Hay muchas valoraciones por ítem: Si no hubieran suficientes valoraciones de un ítem, el sistema sería incapaz de proporcionar recomendaciones precisas pues estaríamos ante un problema de falta de información.
    - Hay más usuarios valorando que ítems: A menudo es necesario que haya más usuarios que número de ítems para así el sistema ser capaz de realizar recomendaciones a los usuarios. Por otra parte, normalmente la mayoría de usuarios de los sistemas suelen realizar pocas valoraciones en comparación a los ítems del sistema. Por consiguiente es deseable que existan, en proporción, muchos más usuarios que ítems en el sistema.
- **Secreto del significado (underlying Meaning):** Se refiere a las propiedades que se esconden debajo del significado de la información:
    - Para cada usuario de la comunidad, hay otros usuarios con necesidades o gustos similares: Este sistema de recomendación es factible porque la gente tiene preferencias o gustos parecidos. Como consecuencia, las recomendaciones a usuarios con preferencias poco comunes serán

difíciles de realizar. Obsérvese que el apartado se refiere a una variante de los problemas de los sistemas de recomendación de filtro colaborativo llamado “arranque en frío” (ver “4.4 Problemas”).

  - La valoración de los ítems requiere que éste haya sido probado: Los filtros colaborativos añaden valor a las apreciaciones de ítems de naturaleza subjetiva. Los ítems que no necesitan de la opinión de las personas, pueden ser procesados computacionalmente sin la intervención de los usuarios. Por lo tanto, los dominios en que sus ítems tengan una posible valoración más subjetiva son más afines a ser utilizados en los sistemas de filtrado colaborativo.
  - Los ítems son homogéneos: Si los ítems son similares y la diferencia viene marcada por el criterio subjetivo, los sistemas de recomendación son apropiados porque mediante la subjetividad de los usuarios, permiten la diferenciación de ítems semejantes. Por ejemplo, los álbumes de música tienen precios y duraciones similares. Además dentro de un mismo estilo musical, los ritmos y sonidos son muy semejantes. En este contexto, las opiniones subjetivas aportan un valor añadido que permite la distinción de ítems semejantes.

- **Precisión de la información:** Se refiere a cúan relevantes son las propiedades de los ítems.
  - Persistencia de los ítems: Existen ítems donde su recomendación, además de por sus características, también viene determinada por otros aspectos como puede ser el tiempo. Por ejemplo, las noticias en los periódicos son recomendables en un espacio de tiempo relativamente reciente. Pasados unos días u horas, estas carecen de importancia. En los contextos en donde el tiempo en que un ítem es susceptible a ser recomendado, es relativamente corto o escaso, los filtros colaborativos se topan con un requerimiento difícil de solventar.
  - Persistencia del gusto: Al igual que el caso anterior, si las preferencias de los usuarios cambian rápidamente, las valoraciones antiguas pierden valor proporcionalmente a la velocidad del cambio. El entorno de la moda, en el mundo de la ropa, es un ejemplo donde las tendencias o

preferencias de un año no son validas para el año siguiente. Los filtros colaborativos trabajan mejor en entornos donde los gustos de los usuarios son persistentes a lo largo del tiempo.

4.4 Problemas

Los sistemas de filtrado colaborativo presentan los siguientes problemas:

- **Nuevo usuario:** Cuando un usuario llega al sistema, no es posible hacerle recomendaciones hasta que su perfil sea lo suficientemente completo para encontrarle a su grupo de vecinos cercanos. Existen diferentes técnicas que buscan paliar el problema intentando determinar el mejor ítem para los nuevos usuarios que estos deben valorar. En [85,86] proponen mostrar a los nuevos usuarios ítems basándose en la popularidad y entropía del ítem así como la personalización del usuario.
- **Arranque en frío:** El arranque en frío se suele dar en la etapa inicial de la utilización del sistema, cuando la mayoría de los ítems aún no han sido calificados y, por lo tanto, no son utilizados para los cálculos de similitud. El mismo problema puede manifestarse, aún cuando el sistema ha sido extensamente utilizado, cuando se agregan nuevos ítems o con aquellos ítems que representan gustos selectos de usuarios particulares [87,85]. Algunos métodos se han propuesto para paliar el arranque en frío, como el uso de agentes [88] (ver “6.1.2 Agentes” para una explicación más detallada sobre el término de agentes). En el artículo de Schein et al. [89] propone el algoritmo de “Aspect Model” que abarcan el problema de la siguiente manera: Existe una hipótesis llamada “causa oculta”, la cual motiva a que un usuario seleccione un determinado ítem. El algoritmo establece que al determinar esta hipótesis de “causa oculta” es posible establecer las preferencias del usuario en base a los atributos del ítem. Por ejemplo, en un sistema de recomendación de canciones un usuario tiene especial interés por canciones del tipo “Pop” en donde se cuenta con una nueva canción, de la que no se tiene información sobre las preferencias de los usuarios del sistema sobre ella, pero se conoce que entra dentro del tipo “pop”. Según este modelo, es posible de que la nueva canción

sea del agrado del usuario, debido a la preferencia del usuario por las canciones de ese estilo.

- **Problema de Matriz dispersa:** Surge debido a que cada usuario sólo califica a un número relativamente reducido de ítems. Por tanto la matriz de calificaciones usuario-ítems es una matriz dispersa, con numerosos valores nulos. Como resultado de lo anterior, la probabilidad de encontrar perfiles similares es habitualmente baja.
- **Confianza entre usuarios:** En las recomendaciones del tipo colaborativo, se pude dar el caso de que existan usuarios mal intencionados que intentan corromper el sistema de recomendación. Massa y Avesani en [90] proponen la creación de una "red de confianza" para solventar el problema. La idea es que un usuario, además de puntuar los ítems del sistema, también exprese su confianza en las valoraciones que han hecho otros usuarios. El enfoque propuesto por Massa y Avesani permite basar las recomendaciones únicamente con las valoraciones de usuarios en los que se confía. La red de confianza es asimétrica, es decir, si un usuario A confía en las valoraciones de un usuario B, no implica que el usuario B confíe en las valoraciones del usuario A. El sistema permite, mediante los valores bajos de confianza, una rápida detección de usuarios malintencionados.

4.5 Tendencias generales

Al tratar con información de carácter personal, es inevitable entrar en temas relacionados con la seguridad y privacidad de la información. Los usuarios, de forma implícita, confían en que los sistemas de recomendación sólo utilizarán sus preferencias para realizar recomendaciones y que sus datos permanecerán en el anonimato. Según comenta Canny en [91], en el futuro, los usuarios querrán obtener recomendaciones sobre diferentes aspectos de sus actividades diarias, como pueden ser, comidas en restaurantes, sesiones de cine, o visitas a lugares cercanos de su lugar de residencia. Aunque en este sentido la ley tiene presente aspectos de privacidad, se puede dar el caso de que empresas que se dedican al comercio electrónico recojan información sobre sus usuarios y hagan el esfuerzo de mantenerla de forma privada hasta el momento que por ejemplo, entran en banca rota y aprovechan para vender

este tipo de información, de carácter privado, a terceros. Por ello, los usuarios deben poder tener control exclusivo a toda la información relacionada con él.

Otro aspecto en el que se trabaja es la confianza. Diferentes estudios demuestran la importancia de la confianza entre los usuarios en las recomendaciones [92,88]. Sinha et al. en [93] muestra que las personas, si tienen que elegir entre recomendaciones de amigos o de sistemas de recomendación, prefieren a los amigos que a los sistemas de recomendación, incluso sabiendo que los sistemas de recomendación presenten un exitoso factor de recomendación. Ziegler et al. ha realizado estudios empíricos que demuestran la existencia de una correlación entre la confianza y la similitud de usuarios en determinadas comunidades de usuarios [94]. Por estos motivos, existe una tendencia hacia la creación de modelos arquitectónicos de sistemas de recomendación de filtro colaborativo, en donde no sólo el parentesco o similitud entre usuarios es importante, sino que el factor "confianza" va adquiriendo mayor presencia. O'Donovan en [92] ha demostrado que teniendo en cuenta este aspecto, ha reducido el impacto de los errores de predicción en las recomendaciones.

5 Recomendaciones basadas en el conocimiento

A continuación se definen los sistemas de recomendación basados en el conocimiento (*knowledge based systems*) así como un caso particular de éstos que son los sistemas de recomendación basados en la utilidad (*utility based*). La definición y características de estos sistemas se describe en los apartados "5.1 Definición y características" y "5.2 Recomendaciones basadas en la utilidad". En los siguientes apartados se muestra una generalización de su funcionamiento "5.3 Funcionamiento" y por último, en el apartado "5.4 Conclusiones" se realiza un resumen de los aspectos más importantes de las recomendaciones basadas en el conocimiento.

5.1 Definición y características

Los sistemas de recomendaciones tradicionales, como son los de filtrado colaborativo (ver "4 Sistemas de filtrado colaborativo") o basados en el contenido (ver "3 Sistemas basados en el contenido"), se apoyan en perfiles de usuario y ciertas características de los ítems para realizar las recomendaciones. No obstante, estos sistemas no explorar en profundidad el conocimiento sobre el dominio de los ítems. Así pues, los sistemas de recomendaciones tradicionales son perfectamente validos para procesos de recomendación de productos singulares como CDs musicales, libros o películas, pero no son útiles en dominios complejos.

Los sistemas de recomendación basados en el conocimiento pretenden profundizar en el conocimiento sobre los usuarios y los ítems -mayoritariamente de domino complejo- para la elaboración de recomendaciones que encajen de forma adecuada con los requisitos del usuario. Un caso de dominio complejo pueden ser los servicios financieros o cámaras digitales en donde no es suficiente con establecer propiedades del ítem o producto, sino que se tienen que tener en cuenta otros aspectos como pueden ser la experiencia o expectativas del usuario en el dominio a tratar por el sistema de recomendación.

Por ejemplo, si un usuario está interesado en la compra de un coche, el sistema de recomendación basado en el conocimiento debería cuestionarse aspectos como: ¿Por qué el usuario quiere comprar el coche? O ¿Qué es más importante para el usuario, el

confort o el consumo? Basándose en este tipo de información, el sistema realiza un razonamiento sobre qué productos se ajustan en mayor medida a las necesidades del usuario. Así pues, la clave de los sistemas basados en el conocimiento radica en el significado de "necesidad" por parte del usuario.

Los sistemas basados en el conocimiento hacen una representación explicita sobre el producto así como las oportunidades que lo envuelven, permitiendo:

1. La recomendación de ítems que satisfagan ciertos requisitos de calidad.
2. La explicación del porqué de la recomendación realizada. Este hecho tiene una implicación positiva en la confianza de los usuarios hacia las recomendaciones del sistema [95].
3. El soporte a los usuarios cuando no se puede encontrar una solución o recomendación adecuada para el caso del usuario.

La forma de operar de estos sistemas permite que ciertos problemas de ámbito general de los sistemas de recomendación (ver "2.4 Problemas generales") desaparezcan o sean eliminados:

- Carecen del problema del "cold-start" o arranque en frío ya que sus recomendaciones no dependen de las valoraciones de los usuarios.
- No necesitan almacenar información sobre un usuario en particular porque las similitudes entre las preferencias de los usuarios son independientes las unas de las otras. No obstante, sí que necesitan una retroalimentación de las necesidades del usuario, es decir, el usuario debe informar de cuáles son sus necesidades.
- Ya que las recomendaciones están basadas en el conocimiento del dominio del producto, el sistema es inmune a anomalías estáticas del mercado [96], es decir, el problema de "información cambiante" (ver "2.4 Problemas generales"), referente a las modas de productos o ítems, no afecta a este tipo de sistemas.

Por otra parte, los sistemas de recomendación basados en el conocimiento presentan las siguientes desventajas:

- Es necesaria una ingeniería del conocimiento.
- Las recomendaciones son del tipo estático, es decir, para los mismos casos, se realizan las mismas recomendaciones. Para los usuarios que tengan las mismas necesidades, se realizaran las mismas recomendaciones. Esto no ocurre, por ejemplo, en los sistemas de recomendación de filtro colaborativo ya que a medida que transcurre el tiempo el sistema va adquiriendo más información sobre los ítems y usuarios, por lo que va adaptando las recomendación a medida que va adquiriendo más información.

5.2 Recomendaciones basadas en la utilidad

Las recomendaciones basadas en la utilidad son un caso particular de las recomendaciones basadas en el conocimiento. Su objetivo es el de crear un valor de "utilidad" (beneficio, ventaja o interés) para los ítems a recomendar para un usuario en particular. En principio el valor de "utilidad" se basa en la utilización del conocimiento [21]. Por ello, este tipo de sistemas hereda los beneficios de los sistemas de recomendación basados en el contenido como es la eliminación del problema del arranque en frío. No obstante, añaden la necesidad de crear una "función de utilidad", que permite obtener un valor (de "utilidad") para cada ítem a recomendar el cual proporciona un baremo al usuario que refleja el grado de satisfacción del ítem con sus necesidades, gustos o preferencias.

La contribución de la función de utilidad permite incorporar aspectos de gama más amplia a las recomendaciones realizadas. Mientras que los sistemas de recomendación basados en el conocimiento sólo tienen en cuenta las especificaciones de los productos, los sistemas basados en utilidad añaden otras características que no pertenecen a los ítems en sí; Por ejemplo: los plazos de entrega de un ítem o la garantía de éste. Por este motivo, se considera que los sistemas basados en la utilidad permiten expresar al usuario todas las consideraciones, referentes a tipo de ítem, necesarias para encontrar las recomendaciones que mejor encajen en el marco de sus necesidades.

Cuanto más elaborada sea la función de recomendación, mejores serán las recomendaciones que se ajusten a las necesidades del usuario. Esta situación

incrementa la tendencia de elaborar funciones de utilidad más complejas. No obstante, la elaboración de funciones de utilidad más complejas implica que los usuarios realicen la medición de pesos, valoraciones y asignación de función de utilidad a las diferentes propiedades de los ítems de forma muy precisa. Mientras que para ítems con pocas características puede ser razonable este procedimiento, en dominios donde las características de los ítems son cuantiosas, la asignación de pesos en la función de utilidad se vuelve inmanejable para el usuario. Por ejemplo, no sería apropiada la utilización de sistemas basados en la utilidad en dominios como la recomendación de películas o canciones dado que estos ítems presentan gran cantidad de propiedades. Un caso en el que sí es apropiado este sistema puede ser en la venta de cargadores de teléfonos móviles, en donde existen pocas características (modelo, marca, precio y tiempo de envió) de estos dispositivos.

5.3 Funcionamiento

Los algoritmos que se usan en los sistemas basados en el conocimiento están fundados en el razonamiento basado en el caso. Se pueden distinguir tres tipos diferentes de conocimiento [97]:

- **Conocimiento del catálogo (*catalog knowledge*):** Es el conocimiento que el sistema tiene sobre los ítems y sus características. Por ejemplo, el tipo de cocina “Thai” pertenece al grupo de cocina “Asiático”.
- **Conocimiento Funcional (functional knowledge):** Es el conocimiento de cómo los ítems pueden coincidir con las necesidades de los usuarios. Por ejemplo, el sistema puede saber que para la necesidad de una cena “romántica”, un restaurante apropiado seria uno “tranquilo con vistas al mar”.
- **Conocimiento del usuario (user's knowledge):** El sistema necesita reunir información sobre las necesidades del usuario para poder encontrar los ítems que satisfagan sus necesidades.

La adquisición del conocimiento del usuario es el factor más importante para el sistema ya que cuánta más información se tenga sobre las necesidades del usuario, más precisas serán las recomendaciones hacia éste. Como consecuencia, es habitual

obtener este tipo de conocimiento directamente del usuario de forma explícita mediante el uso de formularios u otros sistemas del mismo estilo.

5.4 Conclusiones

Los sistemas de recomendación basados en el conocimiento incorporan soluciones validas para usuarios que requieren de una recomendación que se ajuste a unas necesidades concretas. Además, ya que presentan una ingeniería del conocimiento del dominio en el que trabajan, eliminan problemas generales de los sistemas de recomendación como son el arranque en frío o anomalías estáticas del mercado (ver "2.4 Problemas generales").

Por otro lado, aunque la eliminación de los problemas mencionados anteriormente es positiva, tienen el defecto de no presentar un carácter innovador, es decir, siempre realizan las mismas recomendaciones para las mismas situaciones. Por lo tanto, se puede decir que su funcionamiento se asemeja a la de un autómata. Como consecuencia, se condiciona que estos sistemas se centren en dominios muy concretos, como pueden ser los entornos financieros, cuya peculiaridad es que no presentan muchos cambios a lo largo del tiempo y donde el objetivo de la recomendación es encontrar un producto que se adapte lo máximo posible a las necesidades del usuario. En el caso concreto de los sistemas de recomendación basados en la utilidad, los dominios para poder ser aplicados se reduce aún más, siendo únicamente recomendables aquellos donde los ítems a recomendar presentan pocas características o propiedades.

El hecho de que los posibles dominios a incorporar los sistemas basados en el conocimiento sean acotados por unas características singulares hace que no sean tan utilizados como los sistemas de recomendación basados en el contenido o de filtrado colaborativo. Por ello, en la literatura de los sistemas de recomendación, los sistemas basados en el conocimiento tienen una presencia relativamente marginal donde su uso se limita a dos situaciones muy concretas:

1. En dominios muy precisos como son los entornos financieros. Las recomendaciones realizadas en este tipo de casos son muy ajustadas a unas necesidades y no son de carácter repetitivo. Por ejemplo, un usuario que pide

un préstamo al banco para la compra de una casa, muy posiblemente no vuelva a pedir otro préstamo al banco al día siguiente.

2. Como apoyo en fases tempranas de implementación de otros sistemas de recomendación, es decir, para eliminar el problema del arranque en frío de otros tipos de sistemas de recomendación. Por ejemplo, se quiere implementar un sistema basado en el filtro colaborativo, pero no se tiene suficiente información de los usuarios para eliminar el problema de arranque en frío. En este caso, se implementa el sistema de recomendación basado en el conocimiento hasta que el sistema tiene suficiente información sobre los usuarios para que no aparezca el problema del arranque en frío. En ese instante, es cuando se prescinde del sistema de recomendación basado en el conocimiento para pasar a utilizar el filtro colaborativo.

6 Sistemas de recomendación semánticos

Los sistemas de recomendación semánticas (*semantic recommender systems*) aparecen con la creación de la web semántica. Su principal característica es la utilización de ontologías y la descentralización tanto de los procesos como de la información. En el apartado "6.1 Historia y definición" se presenta la definición y aspectos que se tienen en cuenta de los sistemas de recomendación. En el apartado "6.2 Clasificación" se muestra su categorización. Por último, se establecen las tendencias que se están marcando en este tipo de sistemas de recomendación en el apartado "6.3 Conclusiones y tendencias generales".

6.1 Historia y definición

Los sistemas de recomendación basados en el contenido y de filtrado colaborativo se basan en el uso de una base de datos de información generada por el mismo sistema o por los usuarios que interactúan con el sistema. Su forma de operar reside en el conocimiento único y exclusivo generado dentro del mismo entorno del sistema de forma que aparecen limitaciones con los sistemas de recomendación de arquitecturas descentralizadas [98], es decir, los procesos de comunicación entre las diferentes partes de un sistema descentralizado se ven dificultados por las diversas formas en que la información se ve representada. Por otra parte, la heterogeneidad en la representación de la información conlleva que no se pueda utilizar por otras aplicaciones. Estas deficiencias son abordadas con la aparición de la web semántica [35].

Mediante la aparición de la web semántica, donde ontologías y taxonomías son de vital importancia para la representación del conocimiento, empezaron a aparecer sistemas de recomendación semánticos. Así pues, los sistemas de recomendación semánticos son aquellos que trabajan con información dotada de significado y además presentan un enfoque arquitectónico descentralizado (esto no implica que puedan tener una arquitectura centralizada). Más concretamente Ziegler específica cuatro aspectos que se deben tener en cuenta para la creación de los sistemas de recomendación semánticos [98]:

- **Compromiso de Ontología (*ontological commitment*):** La creación de un sistema web semántico no tendría sentido sin que los agentes software (ver "6.1.1 Web Semántica" y "6.1.2 Agentes") tuvieran la capacidad de operar y entender el contenido de la misma. Por este motivo, la representación de la información debe estar disponible por medio de ontologías o modelos de contenido comunes.
- **Facilidad de interacción (*interaction facilities*):** Al tratarse de un sistema de recomendación descentralizado en donde operan diferentes agentes, es necesario facilitar la comunicación entre ellos. Esta comunicación es llevada a cabo mediante el intercambio de mensajes en formato RDF (*Resource Description Framework*), OWL (*Ontology Web Language*) o formatos similares.
- **Seguridad y credibilidad (*security and credibility*):** Mientras que los sistemas de recomendación centralizados tienen un control de la identidad de los usuarios, los sistemas de recomendación descentralizados tienen que establecer mecanismos de seguridad y credibilidad (o confianza) para que los sistemas de recomendación funcionen de forma verosímil y trasmitan confianza a los usuarios.
- **Complejidad computacional y escalabilidad (*computational complexity and scalability*):** Los sistemas de recomendación centralizados permiten limitar o agrupar (clustering) el tamaño de la comunidad del sistema y de este modo consiguen abordar la complejidad computacional de las operaciones necesarias para realizar las recomendaciones. En los sistemas descentralizados, es inevitable abordar el problema de la escalabilidad mediante mecanismos de filtro inteligentes.

6.1.1Web Semántica

La web semántica extiende la web actual para dotar a la información que reside en ella de significado. Eso hace que la interacción entre humanos y maquinas o entre maquinas (agentes software) mejore considerablemente respecto a la interacción con información que carece de significado: existen dos pilares básicos en que se basa la idea de la web semántica:

- **El marcado semántico de la información o recursos:** Es la idea fundamental de la web semántica, es decir, toda la información está dotada de significado: todas las propiedades de un ítem dejan de ser una lista de palabras sin sentido aparente para convertirse en un conjunto de datos que aportan un significado. Por otra parte, al hablar de tecnologías web, también se considera la separación entre el contenido y la forma de representación del mismo.
- **Desarrollo de agentes software:** Los agentes software deben ser capaces de interactuar entre ellos y operar con la información a nivel semántico mediante la utilización de herramientas tales como las ontologías (esquema conceptual de uno o varios dominios cuya finalidad es la de facilitar la comunicación dentro de los diferentes sistemas que operan en el dominio).

6.1.2 Agentes

El concepto de agente es empleado con frecuencia en el ámbito computacional. Puesto que el significado varía de un autor a otro, a continuación se presentan algunas de las definiciones de agentes:

Nwana, en [99], define a los agentes como un programa de computadora que tiene cierto grado de autonomía, se comunica con otros agentes y trabaja en beneficio de un usuario en particular. Por su parte, Laurel, en [100], se refiere al término agente como el que toma la acción y ejecuta una tarea en beneficio de una persona, ya sea en tiempo real o de manera asíncrona. Riecken indica que los agentes son una oportunidad de integrar resultados significativos de diversas áreas de investigación y mostrarlos a los usuarios. Señala también que la idea básica de la investigación en agentes es desarrollar sistemas de software que ayuden a todo tipo de usuarios y se adecuen a sus necesidades [101].

Se distinguen también diversas características sobre los agentes como autonomía, definición del perfil del usuario y confiabilidad. Sánchez, en [102], menciona que la autonomía se ve representada en dos actividades del agente: el agente trata de lograr sus metas de manera autónoma y, trabaja en beneficio de otro con un cierto grado de independencia. Maes, sugiere que el agente "aprende" su comportamiento, observando las acciones del usuario y de otros agentes y es a través del monitoreo de

las acciones del usuario que el agente define las características que distinguen a su usuario (el perfil del usuario)[103]. Para finalizar, Cypher destaca que las acciones del agente deben poder ser anuladas por el usuario, para que este perciba que en todo momento tiene el control y pueda delegar más tareas a su agente [104].

Los diferentes tipos de agentes que se pueden encontrar en el ámbito computacional son los siguientes[102]:

- **Agentes de programador:** Puesto que la complejidad de los sistemas computacionales ha incrementado, los métodos de representación tradicional de entidades de software y hardware (por ejemplo los diagramas de flujo, diagramas de flujo de datos y modelos de objetos) pueden no ser suficientes. Con el propósito de modelar el proceso ejecutado por la computadora para el beneficio del programador, las abstracciones animadas (agentes de programador) pueden ser una mejor representación para explicar el comportamiento de los sistemas.
- **Agentes de red:** Los agentes de red, también llamados móviles, son entidades que viajan a través de los nodos de redes computacionales de acuerdo a sus tareas o requerimientos para encontrar los recursos que necesiten. Un agente de red puede comenzar su tarea en un nodo de la red y si los recursos que necesita no se encuentran en dicho nodo, el agente puede viajar al nodo que le provea de dichos recursos y continuar con su tarea. Los agentes de red requieren la autorización de cada uno de los nodos (a los que quiera dirigirse) para poder viajar.
- **Agentes de usuario:** La representación explícita de agentes a los usuarios surge de manera natural ya que la mayoría de los usuarios están acostumbrados a atribuir decisiones autónomas o intenciones a los programas de computadora. Los usuarios pueden enfrentarse a la complejidad de los sistemas viendo los programas como entidades animadas. Es útil considerar tres subclases de agentes de usuario:
    - **Agentes de información:** Ayudan al usuario a manejar grandes espacios de información que comúnmente se encuentran desorganizados o son

muy dinámicos. Las grandes bases de datos y el WWW presentan una oportunidad de trabajo para este tipo de agentes.

- **Agentes de tareas:** Ayudan a usuarios en tareas realizadas por computadora mediante la ejecución de manera concurrente con las aplicaciones con las que el usuario trabaje para observar su actividad y ofrecerle algunas acciones automatizadas. Se distinguen dos clases de agentes de tareas: (A) los que ayudan a un usuario en particular (agentes personales); y (B) los que asisten a un grupo de individuos en tareas colaborativas (agentes de grupo).
- **Agentes sintéticos:** Crean ambientes para usuarios introduciendo caracteres vivientes en interfaces de computadora. Es por esta característica que son populares en áreas como el entretenimiento.

6.2 Clasificación

En el campo de las recomendaciones semánticas se pueden distinguir tres tipos de sistemas dependiendo del enfoque que se le pueda dar al sistema:

- **Sistemas basados en ontologías o esquemas de conceptos:** Son sistemas de recomendación que utilizan ontologías para representar la información o modelado de los ítems. Además, también es utilizada para la modelación de usuarios o perfiles de usuario. Actualmente existen diferentes lenguajes de marcado de ontologías tales como RDF (*Resource Description Framework*), RDF-Schema (RDFS) [105], DAML+OIL [106,107] y OWL (*Ontology Web Language*) [108-110]. Estas tecnologías permiten la creación de dominios que soportan la descripción de contenido web. RDF y RDFS son lenguajes para la descripción de recursos. Por su parte, OWL presenta más ventajas para expresar significado y semántica que RDF y RDFS. Además, OWL es una revisión de DAML+OIL. De acuerdo con los estudios realizados actualmente, [109,110], se considera que éste es uno de los mejores lenguajes para el modelado de información. Existen diversos proyectos que se enmarcan dentro de este tipo de sistemas de recomendación semánticos. Por ejemplo, Wang et al. en [111] proponen un sistema que trata de mitigar los problemas de los filtros colaborativos mediante

la utilización de ontologías en las categorías de ítems además de dotar a estas de significado mediante meta información. Por otra parte, el sistema "AVATAR" [112] es otro sistema de recomendación especializado en recomendar programas de televisión que combina diferentes estrategias, como la utilización de razonamiento semántico o el uso de diferentes agentes, para mejorar las recomendaciones.

- **Sistemas basados en redes de confianza:** Sistemas enfocados en garantizar la fiabilidad y precisión de las recomendaciones mediante la creación de redes de confianza entre las diferentes partes que componen el sistema. La confianza de los usuarios, además de aportar credibilidad a los resultados de las recomendaciones, implica un aumento de calidad y usabilidad del sistema. Numerosos estudios se han realizado sobre el impacto y creación de redes de confianza en los sistemas de recomendación [113-115]. El sistema "ConTag" [116] es un ejemplo de este tipo de sistemas el cual realiza recomendaciones sobre las etiquetas más apropiadas para la descripción de recursos de la web. Otro ejemplo a destacar es el llamado "Filmtrust" propuesto por Golbeck en [117,118] que recomienda películas mediante el vocabulario FOAF (Friend of a Friend vocabulary). FOAF es un proyecto para la creación de páginas web legibles por maquinas. Estas webs describen personas así como los links entre ellas además de describir las cosas que crean y hacen. Es decir, es una tecnología descentralizada para la conexión de sitios web sociales y la descripción de sus usuarios [119].
- **Sistemas adaptables (sensibles) al contexto:** Son sistemas que tienen en cuenta otros factores para determinar en qué situación o contexto se encuentra el usuario y así adaptar las recomendaciones a ese contexto o situación. Los factores a tener en cuenta pueden ser: temporales, de lugar, nivel de experiencia del usuario o un dispositivo que se utiliza en el momento de recibir la recomendación. Por ejemplo: Tómese como caso el de un usuario que accede al sistema de recomendación mediante un dispositivo móvil y hace saber al sistema que está interesado en conocer lugares para visitar. El sistema puede determinar la localización del usuario y buscar las previsiones meteorológicas para el lugar donde se encuentra el usuario. Teniendo estos

factores en cuenta, el sistema puede establecer que para días en que la temperatura es muy baja, es mejor recomendar lugares para visitar en donde la climatología no tenga una gran influencia. Es habitual en este tipo de sistemas el uso de ontologías para la definición de contextos de uso. Woerndl et al. en [120] ha desarrollado un sistema de información turística para dispositivos móviles, que gestiona información estática (definida en los perfiles de los usuarios, es decir, preferencias de los usuarios) e información dinámica (contextual, por ejemplo, el factor climatológico). En el campo del e-learning, Yu et al. propone un modelo de sistema de recomendación que permita facilitar los recursos que se necesitan para una tarea a los estudiantes dependiendo del nivel del estudiante y/o del avance del curso en el momento de la recomendación. Este sistema utiliza ontologías para representar a los usuarios, recursos y dominios específicos de cada situación [121].

6.3 Conclusiones y tendencias generales

Los sistemas de recomendación semánticos son aquellos que añaden significado a la información que manejan. Es por ello que utilizan ontologías para la representación del conocimiento. El uso de ontologías presenta ciertas ventajas como una mayor compatibilidad entre sistemas dada la homogenización de la información del sistema y por consiguiente facilitan el trabajo con redes sociales además de la comunicación entre agentes software. Además, también disminuyen el alcance del arranque en frío (ver “2.4 Problemas generales”) ya que la información contiene significado y de este modo es posible inferir en ella cuando ésta se encuentra de forma incompleta.

Por otra parte, existe una tendencia que se enfoca en la creación de sistemas mixtos que incorporan redes de confianza con el objetivo de dar credibilidad a las recomendaciones además del uso de información contextual para adaptar las recomendaciones a cada situación que se encuentra el usuario.

7 Sistemas híbridos

A continuación se presentan los sistemas de recomendación híbridos. Su enfoque es distinto a los sistemas de recomendación vistos hasta el momento puesto que proponen la combinación de diferentes sistemas de recomendación para realizar las recomendaciones. En la sección "7.1 Definición" se presenta esta idea formalmente así como en el apartado "7.2 Clasificación", se realiza un análisis de las diferentes formas de combinación de sistemas de recomendación que se pueden dar. Para finalizar la sección "7.3 Conclusiones" describe las consideraciones más importantes a tener en cuenta.

7.1 Definición

Los sistemas de recomendación híbridos combinan diferentes métodos de recomendación para, o bien eliminar problemas específicos de un sistema en concreto, o para aumentar la precisión de las recomendaciones. Típicamente los sistemas de recomendación de filtro colaborativo son combinados con otros sistemas de recomendación para paliar el problema del arranque en frío. No obstante, los otros sistemas de recomendación también pueden ser combinados.

7.2 Clasificación

Existen siete formas diferentes en que se pueden combinar los sistemas de recomendación [21,122]. A continuación se describen cada una de ellas:

7.2.1 Por pesos (*weighted*)

El valor de la recomendación de un ítem se obtiene ponderando los diferentes resultados obtenidos por los sistemas de recomendación (ver "Ilustración 1Ilustración 1 Recomendación híbrida (por pesos - resultado)"). No obstante hay ocasiones en que el resultado de una recomendación no se puede ponderar ya que el sistema de recomendación utilizado no ofrece un valor que expresa el grado de similitud o nivel de agrado del ítem. En estos casos, en lugar de ponderar los resultados de las recomendaciones, se realiza la unión o intersección de los resultados obtenidos por los diferentes sistemas de recomendación y los elementos resultantes son llamados

candidatos (ver “Ilustración 2 Recomendación híbrida (por pesos - candidatos)”). Por ejemplo, el sistema “P-Tango” [41] al principio da a los sistemas de recomendación colaborativo y basado en el contenido el mismo peso para las recomendaciones, pero éste se va ajustando a medida que los usuarios valoran los ítems.

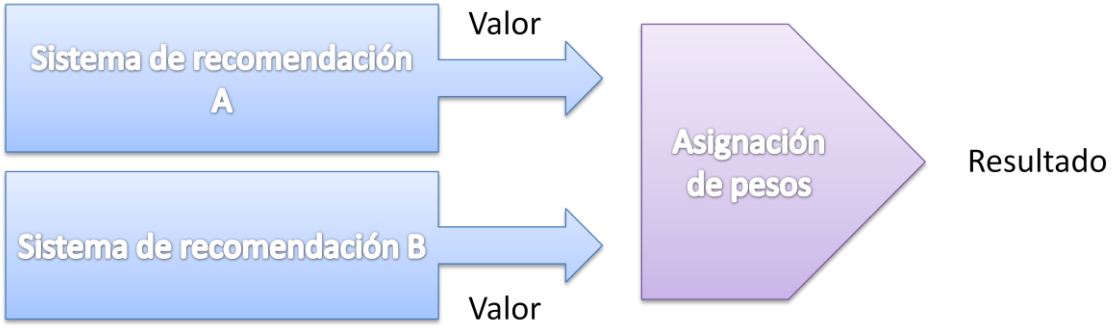

Ilustración 1 Recomendación híbrida (por pesos - resultado)

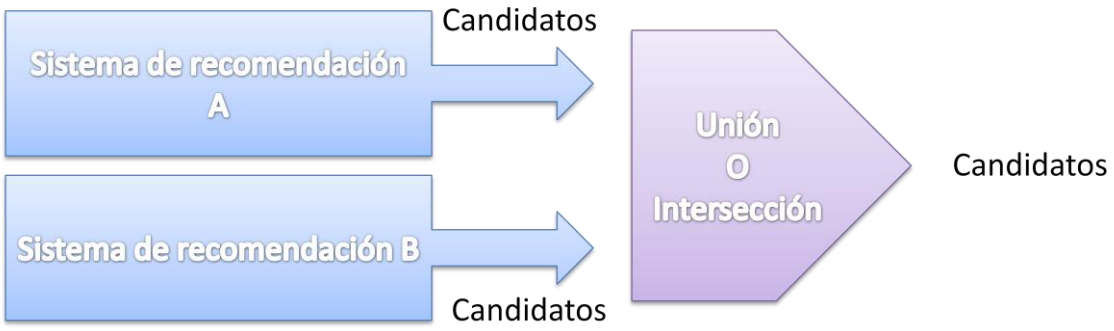

Ilustración 2 Recomendación híbrida (por pesos - candidatos)

7.2.2 Conmutados (*switching*)

El sistema utiliza un criterio para establecer qué sistema de recomendación utilizar en cada momento. Existen dos posibilidades para este caso:

1. A partir de los resultados obtenidos por los sistemas de recomendación involucrados, se determine qué resultados mostrar (ver “Ilustración 3 Recomendación híbrida (Conmutados A)”).
2. Se selecciona el sistema de recomendación a utilizar antes de procesar ninguna información (ver “Ilustración 4 Recomendación híbrida (Conmutados B)”).

Por ejemplo, “DailyLearner” utiliza el método de recomendación basada en el contenido para recomendar noticias, pero cuando éste no tiene la suficiente confianza

para realizar una recomendación, se cambia al sistema de recomendación basado en el filtro colaborativo [21].

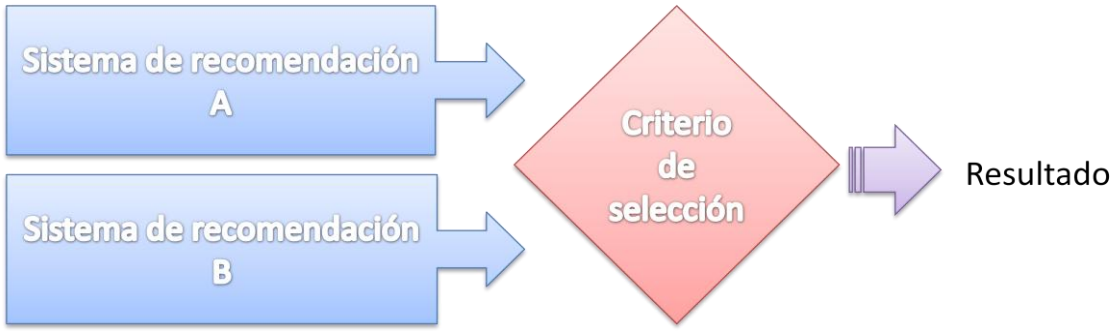

Ilustración 3 Recomendación híbrida (Conmutados A)

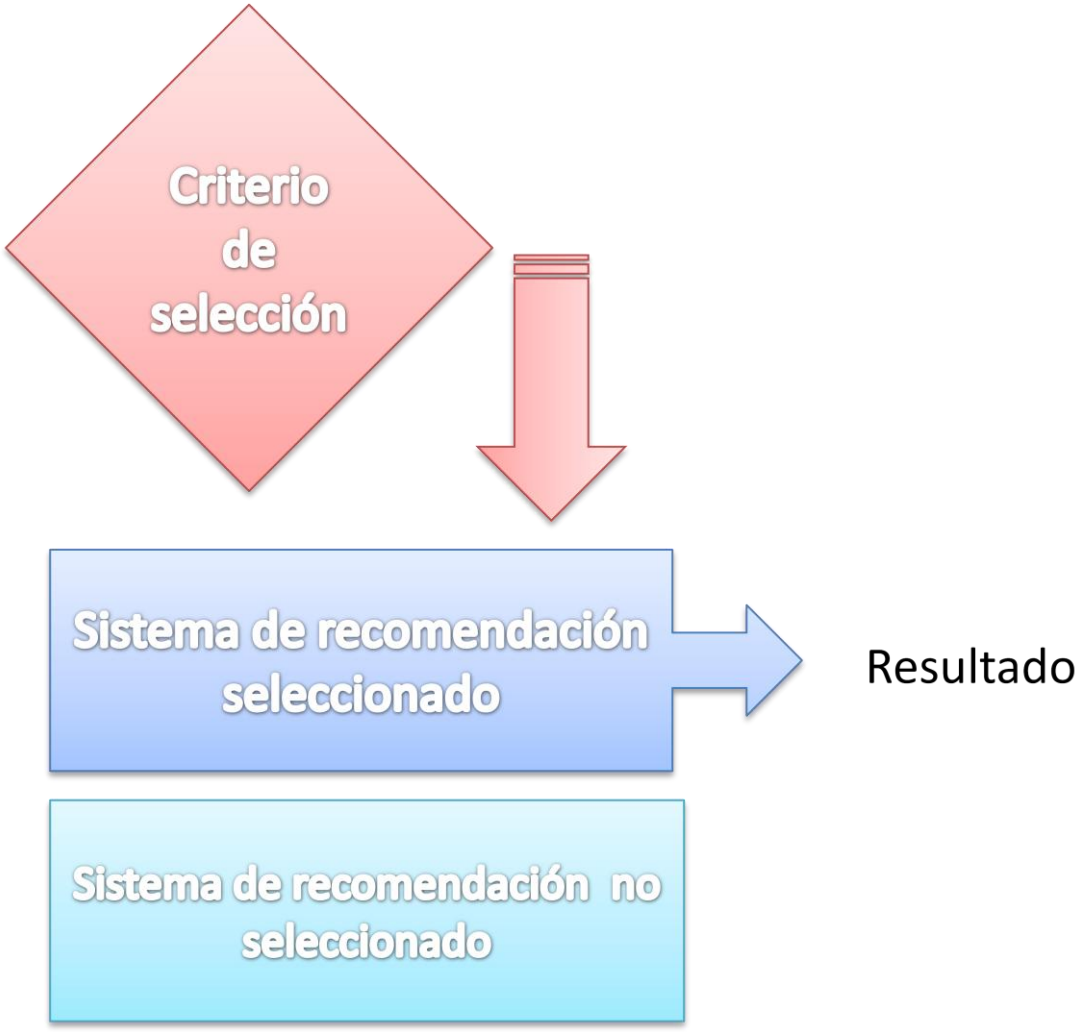

Ilustración 4 Recomendación híbrida (Conmutados B)

7.2.3 Mezclados (*mixed*)

Las recomendación de más de un método se realizan simultáneamente, es decir, diferentes recomendaciones se presentan al mismo tiempo (ver "Ilustración 5 Recomendación híbrida (mezclados)"). El Sistema PTV [123] utiliza esta forma para recopilar recomendaciones sobre programas de televisión. Utiliza técnicas basadas en el contenido para mostrar descriptores textuales de los programas y además información colaborativa sobre las preferencias de los usuarios.

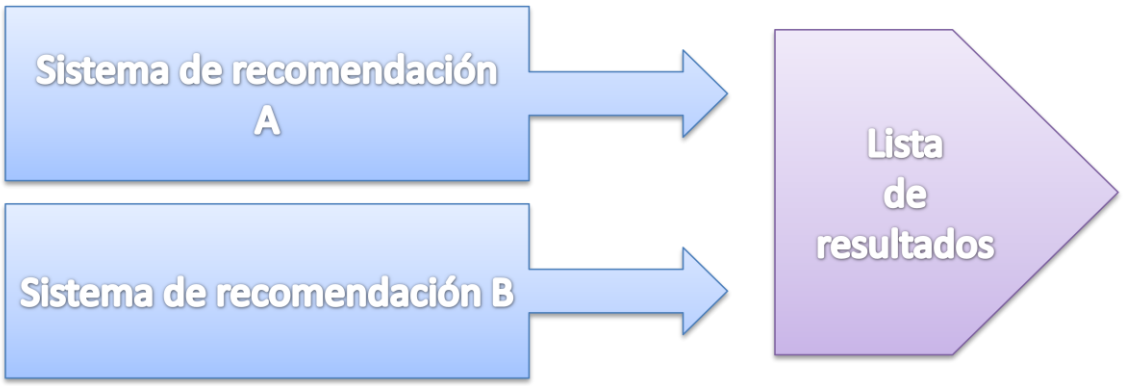

Ilustración 5 Recomendación híbrida (mezclados)

7.2.4 Combinación de propiedades (*feature combination*)

Las propiedades o rasgos de un tipo de sistema de recomendación son usados, mediante una adaptación, a otro tipo diferente de sistema de recomendación. Por ejemplo, Basu et al., [36] utiliza el aprendizaje inductivo "Ripper" [124] de Cohen para aprender reglas basadas en el contenido sobre los gustos del usuario. En el sistema de recomendación basado en el contenido que expone Basu para la recomendación de películas, se añadieron reglas de aprendizaje al sistema de naturaleza colaborativa. Así pues se trataban hechos como "Usuario1 y Usuario2 les gusta la película X" de la misma manera que "Actor1 y Actor2 son protagonistas en la película X". A continuación, en "Ilustración 6 Recomendación híbrida (Combinación de propiedades)" se muestra el esquema de su funcionamiento.

La combinación de propiedades no es un sistema hibrido de recomendación desde el punto de vista de que sólo utiliza un sistema de recomendación. No obstante, se considera que forma parte de los sistemas de recomendación híbridos ya que realiza una combinación de fuentes de conocimiento: La combinación de propiedades toma la

lógica de las recomendaciones de un tipo de técnica en lugar de utilizar un componente o artefacto que la implementa. En el ejemplo expuesto anteriormente de Basu, sólo existe un sistema de recomendación, basado en el contenido, pero este adquiere conocimiento o propiedades de una fuente asociada a los sistemas de recomendación de filtro colaborativo.

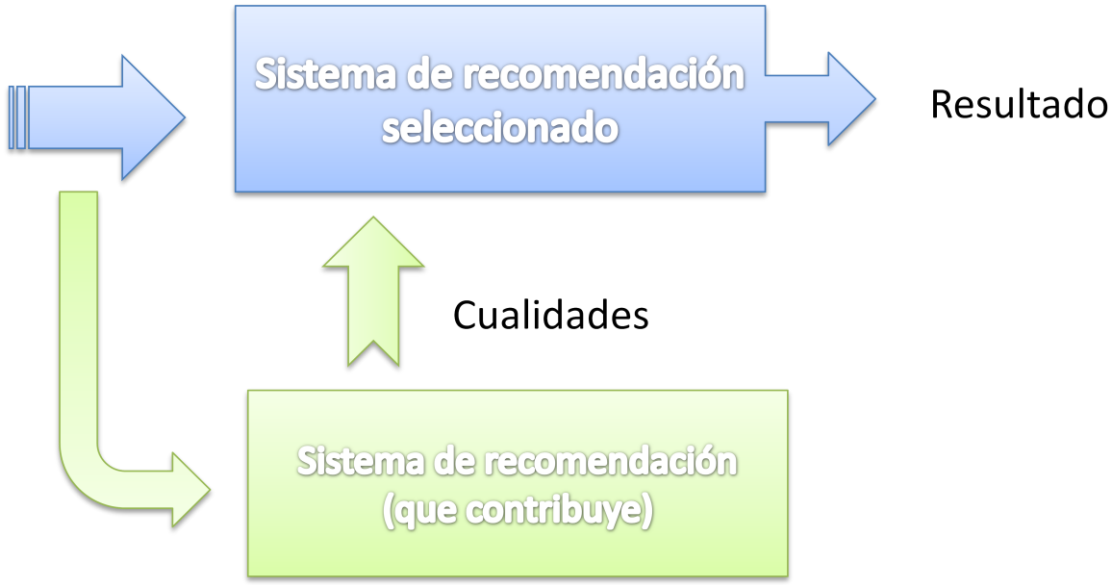

Ilustración 6 Recomendación híbrida (Combinación de propiedades)

7.2.5 En cascada (*cascade*)

Se utiliza de algún método de recomendación para elaborar una lista de posibles recomendaciones y a partir de esta lista de recomendaciones, aplicar un segundo algoritmo de recomendación, es decir, un sistema de recomendación refina las recomendaciones dadas por otro sistema de recomendación. En "Ilustración 7 Recomendación híbrida (cascada)" se muestra de forma esquemática su funcionamiento. Por ejemplo, es posible que un sistema de recomendación basado en el conocimiento, elabore una lista de posibles restaurantes candidatos para recomendar y que partiendo de esta lista, un sistema de recomendación colaborativo seleccione el que tenga mejor valoración popular.

Ilustración 7 Recomendación híbrida (cascada)

7.2.6 Aumento de cualidades (*feature augmentation*)

El aumento de cualidades combina la utilización de dos sistemas de recomendación de la siguiente forma: En primer lugar el sistema de recomendación contribuyente realiza las recomendaciones de un ítem aportando información complementaria del tipo "autores o películas relacionadas con la recomendación". Estas recomendaciones son utilizadas por un segundo sistema de recomendación como elementos de entrada de éste. En otras palabras, los resultados de una técnica o sistema de recomendación son usados como elementos o cualidades de entrada de otro sistema de recomendación (ver "Ilustración 8 Recomendación híbrida (Aumento de cualidades)"). Por ejemplo, en el artículo [125] de Mooney y Roy se describe un sistema de recomendación de libros basado en el contenido. Este sistema de recomendación extrae información sobre libros a partir de la información que apare en "Amazon". La particularidad de extraer información a partir de los resultados que aparecen en "Amazon" es que esta información contiene las recomendaciones que "Amazon" hace al usuario. Estas recomendaciones que "Amazon" muestra al usuario son del tipo "títulos relacionados" o "Autores que también pueden ser de interés". Así pues, el sistema de recomendación de aumento de cualidades, además de obtener como elementos de entrada para el aprendizaje del sistema de recomendación las propiedades del ítem recomendado (como puede ser, autor de la obra, tipo de obra o número de páginas), también incorpora las recomendaciones, en este caso del tipo "títulos relacionas" o "Autores que también pueden ser de interés", a su motor de aprendizaje. Esta característica añade calidad a las recomendaciones [21].

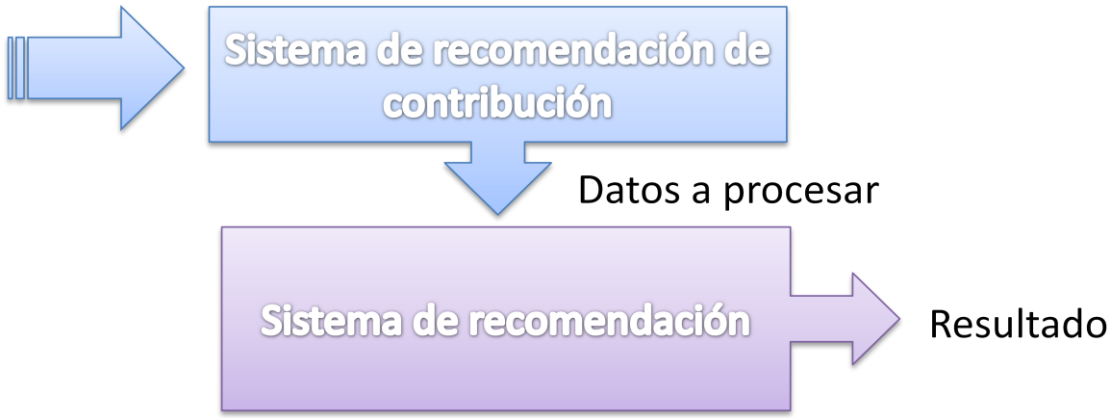

Ilustración 8 Recomendación híbrida (Aumento de cualidades)

7.2.7 Meta niveles (*meta-level*)

En los meta niveles, el modelo aprendido por un sistema de recomendación es usando como fuente de entrada para otro sistema de recomendación. El enfoque se asemeja al de "aumento de cualidades" en el hecho de que el sistema de recomendación primero contribuye en la entrada de información del sistema de recomendación. No obstante, en el caso de los meta niveles, la contribución del primer sistema de recomendación reemplaza completamente la fuente original de conocimiento. Por ejemplo, Pazzani en [36] utiliza técnicas de clasificación bayesiana para crear modelos de preferencias de usuarios que se basan en el contenido. A continuación en "Ilustración 9 Recomendación híbrida (meta niveles)" se muestra de forma gráfica el funcionamiento de este sistema.

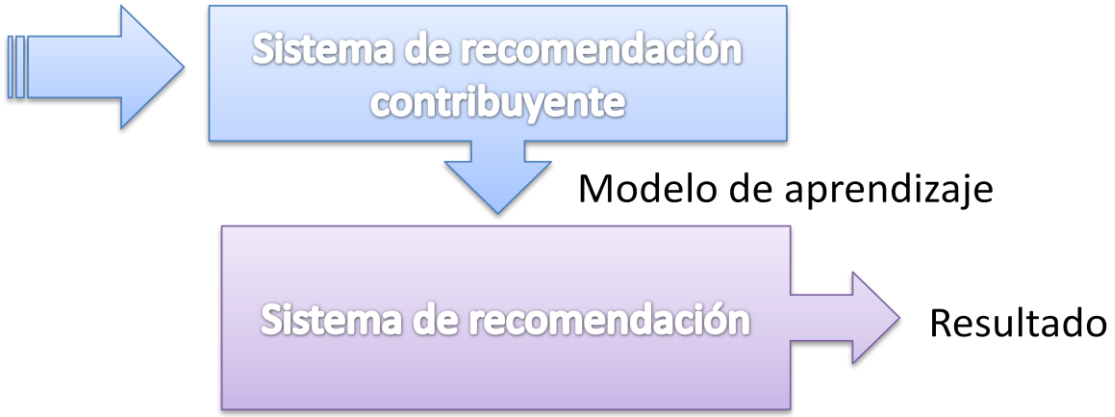

Ilustración 9 Recomendación híbrida (meta niveles)

7.3 Conclusiones

Todos los sistemas de recomendación presentan algún problema o inconveniente. No obstante, al combinarlos, es posible atenuar o incluso eliminar ciertos inconvenientes que los sistemas de recomendación presentan cuando se utilizan de forma aislada. Por ejemplo, los sistemas de filtro colaborativo y basados en el contenido son sistemas de recomendación que funcionan bien con gran cantidad de datos, pero tienen el problema del arranque el frío (ver "2.4 Problemas generales"). Este inconveniente puede ser eliminado mediante la utilización de recomendaciones basadas en el conocimiento. La idea consistiría en realizar las primeras recomendaciones mediante un sistema basado en el conocimiento y a medida que la base de información de sistema vaya aumentando, ir aplicando los sistemas colaborativos o basados en el contenido.

Por otra parte, la implementación de sistemas de recomendación híbridos añade complejidad al sistema y requiere además la construcción de dos o más sistemas de recomendación así como la necesidad de interoperabilidad entre ellos. Además, cabe considerar que las diferentes formas de sistemas híbridos –por pesos, conmutados, mezclados y combinación de propiedades– presentan la propiedad asociativa, es decir, no importa el orden en que se apliquen los diferentes sistemas de recomendación ya que el resultado será el mismo, mientras que las restantes –en cascada, aumento de cualidades y meta niveles– se debe tener en cuenta el orden de los algoritmos a aplicar puesto esto implicaría resultados diferentes.

8 Seguridad en los sistemas de recomendación

La seguridad siempre ha sido un aspecto a tener en cuenta en el hábito de la informática. Por ello, a continuación se aborda el tema desde el punto de vista de los sistemas de recomendación. El enfoque adoptado es desde la perspectiva de sesgo de la información, es decir, no se consideran aspectos como por ejemplo, el robo de la clave para acceder a la base de dados del sistema o las fisuras del sistema debido a errores de programación. En otras palabras, el tema se centra en los casos en que mediante el uso (mal intencionado o no) del sistema, se pueda desinhibir o anular la función principal de los sistemas de recomendación, es decir, realizar recomendaciones para cada usuario lo más precisas posibles y que más bien se ajusten a las necesidades o preferencias de este.

Cabe destacar las diferentes nomenclaturas, de los ataques a los sistemas de recomendación, que han aparecido en la literatura. En un principio, se han denominado "*shilling*" [126,115] pero también existen referencias que prefieren denominarles "*profile injection attacks*" (más concretamente, dentro del campo de los sistemas de recomendación de filtro colaborativo) [115]. A pesar de tener nombres diferentes, ambos se refieren al mismo aspecto, atacar a los sistemas de recomendación.

Dada la naturaleza de los sistemas de recomendación tratados –basados en el contenido, filtro colaborativo, semánticos y basados en la utilidad- el sistema de recomendación más propenso a ser atacado es el de filtro colaborativo dado que el mecanismo de recomendación funciona gracias a las valoraciones (*inputs*) que proporcionan los usuarios. Estas valoraciones, a su vez, tienen repercusiones sobre las recomendaciones de los demás usuarios del sistema. Por otra parte, los otros sistemas de recomendación –basados en el contenido, semánticos y basados en la utilidad- trabajan sobre información cerrada, es decir, los usuarios del sistema no pueden manipular la información de forma tan directa como los sistemas de filtrado colaborativo.

En primer lugar, “8.1 Objetivo del ataque” se realiza una distinción de los posibles ataques que puede sufrir un sistema de recomendación y en el apartado “8.3 Soluciones para los ataques” se discuten posibles soluciones que brinda la literatura así como sus inconvenientes.

8.1 Objetivo del ataque

La seguridad en informática es un aspecto a tener en cuenta en todo momento, especialmente cuando se trata con información de carácter privado. Los sistemas de recomendación, al trabajar con información referente a preferencias y gustos de usuarios no deberían de ignorar esta consideración. Existen tres razones básicas por las que un sistema de recomendación puede ser atacado:

1. **Obtención de información:** Al contener información de carácter privado y además presentar las preferencias y gustos de los usuarios, los sistemas de recomendación son objeto de ataques con el fin de substraer datos de carácter privado de los usuarios. Por ejemplo, una empresa que se dedica al marketing de productos, puede estar interesada en estos aspectos.
2. **Cambiar la información del sistema:** El objetivo de un ataque que cambia la información del sistema de recomendación es el de sesgar la información para que o bien el sistema de recomendación realice recomendaciones de un ítem preferiblemente no recomendable (*push*) o de forma contraria, no recomiende ítems susceptibles a ser recomendados (*nuke*). Por ejemplo, una empresa que se dedica a la venta de un determinado producto, podría tener la intención de cambiar la información de sistema de recomendación para que éste recomendara sus productos en lugar de los de la competencia.
3. **Dañar al sistema:** Existen personas o grupos de usuarios que se dedican a tantear o probar la robustez de los sistemas realizando ataques al sistema con el objetivo de corromper su seguridad, datos o confianza de los usuarios.

Los ataques corrompen el sistema consiguiendo que no formule las recomendaciones adecuadas o simplemente no realice recomendaciones además de tener consecuencias para la reputación o confianza que trasmiten los sistemas de recomendación a los usuarios. Cabe destacar que el aspecto de la reputación de los

sistemas de recomendación ha adquirido gran importancia como riesgo a tener en cuenta, sobre todo en las comunidades online [127]. Si bien cuando un amigo recomienda, por ejemplo, ver una determinada película, y se acepta la recomendación es por dos motivos: La recomendación encaja con los gustos del usuario y, además, existe una confianza en la veracidad de la información aportada por parte de los dos amigos. En los sistemas de recomendación, existe un paralelismo a esta situación: Los perfiles de usuario contienen la información de preferencias del usuario y la confianza del usuario hacia el sistema representa el grado de veracidad de las recomendaciones.

Tomando como referencia el punto de "Cambiar la información del sistema", se pueden distinguir diversos casos. Para ello se parte del hecho de que las recomendaciones se basan en baremos o valoraciones de los usuarios. En este contexto, en la mayoría de casos se encuentran dos grupos de usuarios: Los "consumidores", que son los usuarios del sistema de recomendación a los que el sistema proporciona las recomendaciones, y los "vendedores", que son aquellos usuarios que se encuentran detrás del ítem a recomendar. Mediante la combinación de los dos grupos de usuarios, "consumidores" y "vendedores" y dependiendo de las valoraciones de cada uno de ellos, se distinguen cuatro tipos de situaciones [126,128]:

1. **Manipulación de la valoración de los consumidores:** Hace referencia a las valoraciones aportadas por los consumidores con el objetivo de sesgar la información relativa a los vendedores. Se distinguen dos casos:
    a. **Sobrevalorar al vendedor (*ballot stuffing*):** Los consumidores valoran de forma más positiva a los vendedores para aumentar la reputación del vendedor. En el contexto del comercio electrónico, sobrevalorar al vendedor, permite obtener más solicitudes de ítems del vendedor. Impulsar la recomendación de un determinado ítem, mediante un ataque al sistema, es conocido con el nombre de "push".
    b. **Infravalorar al vendedor (*bad-mouthing*):** Representa la situación contraria al punto anterior, es decir, se reduce la reputación del vendedor. El hecho de restar importancia a la recomendación de un determinado ítem, mediante un ataque al sistema, es conocido con el nombre de "nuke".

2. **Manipulación de la valoración de los vendedores:** Se refiere a las acciones por parte del vendedor para la manipulación del mercado de usuarios. Se pueden distinguir dos casos:
    a. **Discriminación negativa:** El vendedor provee un buen servicio a todos los usuarios excepto a un grupo de ellos. Este grupo específico de usuarios, "*que no gusta*" al vendedor son discriminados por lo que la reputación del vendedor se basa sólo con los usuarios atendidos de forma adecuada.
    b. **Discriminación positiva:** Los vendedores proveen un servicio excepcionalmente bueno a un grupo de usuarios con el objeto de incrementar la media de valoraciones del servicio en general.

8.2 Tipos de ataque

O'mahony et al. introduce los conceptos de robusteza y estabilidad [129]. El término de robustez mide la interpretación del sistema antes y después del ataque con el objetivo de determinar la medida en que el ataque ha afectado al sistema. Se dice que un sistema de recomendación es robusto si puede mantener la calidad de sus recomendaciones a pesar de los ataques que reciba. Por otro lado, el término de estabilidad hace referencia al cambio del sistema de valoraciones que el ataque provoca mediante la introducción de perfiles en el sistema. Un término relacionado con la robusteza es el impacto. Por ejemplo, si un ataque cambia los valores de predicción de un ítem de tal forma que aun con el cambio realizado el ítem no aparece en la lista de recomendaciones de usuarios, el impacto de este ataque se considera nulo. A continuación se distinguen diferentes tipos de ataque que de un modo u otro afectan a la robustez, estabilidad o el impacto de los sistemas de recomendación atacados [115,126]:

- **Ataque perfecto del conocimiento (*Perfect Knowledge Attack*):** La forma de proceder del atacante consiste en una reproducción precisa de la distribución de los datos en el perfil de la base de datos. En otras palabras, se introducen perfiles de usuario en el sistema de forma que sean similares o idénticos al perfil del sistema a atacar. Una vez introducidos los perfiles en el sistema, se

procede a realizar valoraciones especiales para un determinado ítem. Dependiendo del tipo de valoración que se le asigne al ítem, se pueden distinguir dos casos:

  - **"Push":** En el que se intenta que un ítem determinado presente mayor importancia.
  - **"Nuke":** Es el efecto contrario al "*push*", es decir, se pretende quitar importancia a la recomendación de un cierto ítem.

- **Ataque aleatorio (*Random Attack*)** [130]**:** Es un tipo de ataque que está enfocado en la realización de push o nuke a un determinado ítem para que tenga repercusion para todos los tipos de perfiles de usuario del sistema. El procedimiento consiste en (1) realizar la valoración del ítem en concreto, de forma positiva o negativa (depende del objetivo a conseguir) y (2) escoger una lista de ítems aleatorios del sistema para ser valorados de forma genérica. Mediante la valoración de un grupo de ítems de forma genérica, se consigue que el ataque tenga efecto para todos los tipos de perfiles de usuario del sistema.
- **Ataque medio (*Average Attack*)** [130]**:** Procedimiento similar al ataque aleatorio (ataque aleatorio), pero con la peculiaridad de que el atacante conoce las valoraciones promedio de los ítems aleatorios. De este modo, los ítems son apreciados con un valor promedio y el ítem a atacar con los valores deseados. El impacto del ataque es mayor que el anterior (ataque aleatorio) dado que se utiliza información adicional del sistema para realizar el ataque.
- **Ataque consistente (*Consistency Attack*):** También conocido como "ataque al ítem favorito" (*Favorite item attack*) [131]. La "consistencia" o regularidad de las valoraciones de los diferentes ítems son manipuladas a pesar de su valor absoluto. En la "Tabla 3 Ejemplo de ataque consistente" extraída de un ejemplo proporcionado por Mobasher et al,. en el artículo [132] muestra el funcionamiento del ataque consistente. En la tabla aparecen las valoraciones de diferentes usuarios para diferentes ítems. El valor que representa mayor agrado del usuario por el ítem es 10 mientras que el valor cero corresponde a la peor puntuación posible de un ítem. Como se puede observar, el patrón de ataque para los ítems 1, 3 y 4 es el mismo y además son ítems que Alicia tiene

valorados positivamente. Por otro lado, los ítems que no son del agrado de Alicia (2 y 5) son valorados de forma distinta al patrón de ataque. Según muestra el ejemplo, el ítem que más le gusta Alicia es el primero, el cual se mueve desde un nivel muy bajo de similitud al ítem 6 a un nivel muy alto similitud una vez efectuado el ataque. El ataque consigue que un ítem que no representaba ningún valor para determinar las valoraciones de las recomendaciones (ítem 1), pase a tomar importancia para el usuario. Es importante señalar la necesidad de saber exactamente qué elementos son preferentes o agradan al usuario para poder realizar el ataque.

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Alicia	10	2	6	6	Ø	?
Usuario 1	2	Ø	8	8	Ø	2
Usuario 2	2	0	6	Ø	2	2
Usuario 3	8	2	6	6	Ø	0
Usuario 4	6	5	5	Ø	5	2
Usuario 5	Ø	6	Ø	5	5	2
Usuario 6	4	4	Ø	5	5	2
Usuario 7	Ø	10	Ø	2	10	0
Ataque 1	10	10	10	10	0	10
Ataque 2	10	0	10	10	10	10
Ataque 3	10	10	10	10	0	10
Ataque 4	10	0	10	10	10	10
Ataque 5	10	10	10	10	0	10
Cosine vs ítem 6 (antes)	**-7.8**	**0.9**	**114**	**3.8**	**-7**	
Cosine vs ítem 6 (después)	**7.7**	**-0.9**	**9.2**	**1.1**	**-2.7**	

Tabla 3 Ejemplo de ataque consistente

- **Ataque de segmentación (*Segmented Attack*)** [133]**:** El ataque consiste en hacer nuke o push de un determinado ítem teniendo en cuenta los ítems similares o de misma categoría. Supóngase el caso en que el atacante quiere realizar un ataque de segmentación del tipo push sobre un determinado ítem, por ejemplo, un libro de acción. Mediante este tipo de ataque, lo que pretende es que el ítem sea recomendado a los usuarios que consumen ítems de una determinada categoría. En el ejemplo del libro, el atacante quiere que el libro a recomendar sea recomendado por el sistema a los usuarios consumidores o de preferencias de libros de acción. El resto de usuarios del sistema no se tiene en

cuenta. Para llevar a cabo el ataque, se crean perfiles de usuario que contengan altas valoraciones de ítems que estén dentro las preferencias de los usuarios del segmento (en el ejemplo del libro de acción, el segmento se refiere a los usuarios que tienen marcado como preferencias los libros de acción). Al mismo tiempo que se crean estos perfiles de usuario, se realizan valoraciones positivas para los ítems pertenecientes al segmento o grupo de usuarios en concreto y valoraciones negativas para el resto de ítems del sistema. Para llevar a cabo el ataque de segmentación, sólo es necesario saber qué tipo de ítems son similares al ítem atacado (push o nuke).

- **Ataque en pandilla (*Bandwagon Attack*)**[115]**: A**taque cuyo objetivo es asociar un determinado ítem con un grupo determinado de ítems populares. Por ejemplo, el ataque intenta que un libro determinado sea asociado a la categoría de "más vendidos" o "*bestsellers*". Mediante la creación de esta asociación el sistema de recomendación potenciaría la recomendación del libro.
- **Ataque de sondeo (*Probing attack*):** Ataque enfocado a la averiguación del tipo de algoritmo (o parámetros) que rige(n) detrás de las recomendaciones. Por ejemplo, el ataque consistente es más efectivo si el algoritmo de recomendación está basado en los ítems. Mediante el ataque de sondeo se puede encontrar el algoritmo que se utiliza para realizar las recomendaciones y posteriormente aplicar el ataque más conveniente dependiendo del sistema de recomendación. Una variante de este ataque es el ataque de muestra (*sampling attack*) cuyo objetivo es la obtención de las valoraciones de los ítems de la base de datos [129].

8.3 Soluciones para los ataques

En el artículo de Dellarocas se presentan diversas formas de combatir los cambios de información del sistema [128]. A continuación se comentan las soluciones propuestas:

En las situaciones de Bad-mounthing y discriminación negativa, en donde el ataque consiste en escoger una serie de usuarios para sesgar el resultado, se puede evitar este tipo de ataque si se ocultan las identidades de los usuarios, tanto consumidores como

vendedores. Por consiguiente, los usuarios realizan sus votaciones, pero el sistema no muestra el origen o quien ha realizado la votación. El inconveniente de ocultar las identidades de los usuarios es que no es aplicable para todos los dominios ya que algunas veces no es compaginable el anonimato en el sistema. Los casos de ballot stuffing y discriminación positiva se pueden combatir mediante un filtro basado en clusters. El artículo de Dellarocas [128] presenta una argumentación a esta proposición así como diferentes estudios.

Por otra parte, los temas de seguridad son relativamente fáciles de manejar cuando un sistema está distribuido de forma centralizada, pero se convierten en algo más complicado cuando se trata de arquitecturas distribuidas. Canny en [91] ha desarrollado un sistema de protección de privacidad para estos casos basado en la encriptación y compartición de claves. No obstante la mayoría de soluciones pasa por crear, mantener y fomentar la confianza entre de usuarios y usuario-sistema [88,90,94,113,114,117,127].

9 Conclusiones

Actualmente existe una gran variedad de tipos de sistemas de recomendación. La gran mayoría se distinguen por pequeños matices o propiedades, es decir, como se ha mostrado en el apartado "2.3.1 Clasificación exhaustiva" dependiendo del enfoque lógico, de aproximación y la manera de realizar las operaciones del sistema, se pueden dar gran diversidad de tipos de sistemas de recomendación. No obstante, si se realiza un enfoque más generalista y, por consiguiente práctico, esta gran variedad de tipos de sistemas de recomendación se ve reducida a los siguientes:

1. Sistemas de recomendación basados en el contenido (ver "3 Sistemas basados en el contenido").
2. Sistemas de recomendación de filtro colaborativo (ver "4 Sistemas de filtrado colaborativo").
3. Sistemas de recomendación que utilizan meta información para enriquecer o establecer recomendaciones más apuradas. En este sentido, se distinguen dos tipos de sistemas de recomendación:
    a. Sistemas de recomendación basados en el conocimiento (ver "5 Recomendaciones basadas en el conocimiento").
    b. Sistemas de recomendación semánticos (ver "6 Sistemas de recomendación semánticos").

Cada uno de los diferentes tipos de sistemas de recomendación realiza un enfoque diferente para la realización de recomendaciones. Por ello, aparece la necesidad de estudio y comprensión del dominio o medio en que deben ser aplicados o implementados para la maximización de los resultados (en forma de recomendaciones). Además, el estudio y comprensión del dominio de aplicación del sistema de recomendación debe comprender la necesidad de minimización de los problemas que acarrean los diferentes tipos de sistemas de recomendación (ver "2.4 Problemas generales","3.2 Problemas generales","4.4 Problemas","8 Seguridad en los sistemas de recomendación"). Una posible solución para la mitigación de los problemas asociados a cada uno de los diferentes tipos de sistemas de recomendación es la combinación de estos (ver "7 Sistemas híbridos"). En este sentido, existen

diferentes posibilidades y formas de combinación. Una vez más, el estudio y comprensión del dominio a aplicar el sistema de recomendación es necesaria para la correcta implementación de una combinación de distintos tipos de sistemas de recomendación.

En el siguiente apartado, "9.1 Comparación", se realiza un resumen comparativo sobre los aspectos a favor y en contra de los distintos sistemas de recomendación.

9.1 Comparación

A continuación se presenta una tabla a modo de resumen que permite comparar los aspectos positivos ("Tabla 4 Comparativa sistemas de recomendación (a favor)") y negativos ("Tabla 5 Comparativa sistemas de recomendación (en contra)") de cada uno de los sistemas de recomendación presentados en las secciones anteriores.

Sistema	A favor
Basados en el contenido	No es necesario un conocimiento del dominio. Adaptativo: La calidad es incremental con el tiempo. Es suficiente el Feedback implícito. Técnicas de automatización de atributos.
Filtro colaborativo	Recomienda diferentes géneros. No es necesario un conocimiento del dominio. Adaptativo: La calidad es incremental con el tiempo. Estabilidad (no se sobre-especializa). Es suficiente el Feedback implícito. Diferencia ítems homogéneos (mismos atributos).
Basados en el conocimiento	No tienen arranque en frío. Sensible a los cambios de preferencias. Adaptable a las necesidades de los usuarios. Incluye características que no son propias de los productos.
Semánticos	Utilización de ontologías e interoperabilidad entre agentes. Atenúa el arranque en frío.
Basados en la utilidad	No tienen arranque en frío. Sensible a los cambios de preferencias. Adaptable a las necesidades de los usuarios. Incluye características que no son propias de los productos.

Tabla 4 Comparativa sistemas de recomendación (a favor)

Como se puede observar, cada tipo de sistema de recomendación presenta unos pros y contra diferentes. Por consiguiente, dependiendo de la situación o dominio en el que se quiera aplicar un sistema de recomendación, se debe optar por uno u otro sistema a implementar. A todo ello, cabe destacar la posibilidad de utilización de sistemas híbridos, los cuales pueden aportar soluciones a los diferentes problemas de sistemas de recomendación.

Sistema	En contra
Basados en el contenido	Arranque en frío (nuevo usuario) La calidad depende de la cantidad de información. Estabilidad vs manejabilidad. Sobre-especialización. Limitación por el análisis del contenido. No diferencia ítems homogéneos (mismos atributos).
Filtro colaborativo	Arranque en frío (nuevo usuario). Arranque en frío (nuevo ítem). La calidad depende de la cantidad de información. Usuarios con gustos poco comunes. Matriz dispersa. Confianza entre usuarios.
Basados en el conocimiento	Recomendaciones estáticas. Ingeniería del conocimiento.
Semánticos	Confianza de las recomendaciones. Enfocado al contexto web.
Basados en la utilidad	Recomendaciones estáticas. Ingeniería del conocimiento. Creación de la función de utilidad.

Tabla 5 Comparativa sistemas de recomendación (en contra)

Las tablas expuestas anteriormente reflejan un resumen de cada sistema de recomendación y las cualidades o defectos expuestos en ellas se encuentran explicados en cada sección pertinente. No obstante el término "Estabilidad vs manejabilidad" expuesto en el apartado de filtro colaborativo, hace referencia a lo siguiente: Si un usuario ha realizado valoraciones para un grupo de ítems pertenecientes a un conjunto C, el sistema de recomendación será capaz de

recomendarle ítems de otro conjunto diferente al C. Así pues, en el contexto de sistemas de recomendación se dice que la estabilidad es el antónimo de sobre-especialización.

9.2 Líneas futuras

En un principio los sistemas de recomendación se basaban en los datos que el mismo sistema generaba. De este modo, se creaba un círculo cerrado de usuarios del sistema. No obstante, la evolución de los sistemas de recomendación se ha ido encaminando en dirección opuesta. La interconexión y descentralización, tanto de datos como usuarios de otros sistemas es el enfoque que actualmente se le está dando a los diferentes sistemas de recomendación. Es decir, si se tiene en cuenta que la mayoría de sistemas de recomendación operan en internet, se ha pasado de (1) la construcción de sistemas de recomendación cerrados donde las métricas, recomendaciones, noticias, stocks, referencias de usuarios y amigos son propias del mismo sistema a, (2) la activa participación de todos los usuarios de la comunidad de internet. Lo que antes era creado y mantenido por una sola comunidad de usuarios pertenecientes a un dominio en concreto, ahora es creado y mantenido por toda la comunidad de usuarios de la web.

Por otra parte, como consecuencia de la interconexión y descentralización de los diferentes sistemas de recomendación, se ha creado la necesidad (por parte del usuario) de definir aspectos personales (o privados) y públicos de su perfil de preferencias así como poder seleccionar el contenido y presentación de las recomendaciones.

2º Acto: Preludio

Calidad en los sitios del área de la salud

Leer sin reflexionar es como comer sin digerir

10 Calidad en los sitios del área de la salud

El proyecto de calidad en los sitios del área de la salud está enmarcado dentro del campo de investigación de "Tecnologías de la Información y las Comunicaciones" del "Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED)". A continuación, en los apartado "10.1 CYTED", "10.2 CYTED: Tecnologías de la Información y las Comunicaciones" y "10.3 Calidad en los sitios del área de la salud" se describen estos organismos o grupos de trabajo. Posteriormente se evalúa el proyecto desde la perspectiva de los sistemas de recomendación.

10.1 CYTED

El Programa Iberoamericano de Ciencia y Tecnología Para el Desarrollo (CYTED) fue creado en 1984 mediante un acuerdo marco interinstitucional firmado por 19 países de América latina, España y Portugal. Se define como un programa intergubernamental de cooperación multilateral en ciencia y tecnología, que contempla diferentes perspectivas y visiones para fomentar la cooperación en investigación e innovación para el desarrollo de la región Iberoamericana.

Su objetivo principal es contribuir al desarrollo armónico de la región iberoamericana mediante el establecimiento de mecanismos de cooperación entre grupos de investigación de las universidades, centros de investigación y desarrollo (I+D) y empresas innovadoras de los países Iberoamericanos que pretenden la consecución de resultados científicos y tecnológicos transferibles a los sistemas productivos y a las políticas sociales. Desde 1995, el Programa CYTED se encuentra formalmente incluido entre los programas de cooperación de las cumbres iberoamericanas de jefes de estado y de gobierno.

La importancia de CYTED en Iberoamérica radica en el hecho de que es un instrumento común de los sistemas de ciencia y tecnología nacionales de la región Iberoamericana, generando una plataforma que promueve y da soporte a la cooperación multilateral orientada a la transferencia de conocimientos, experiencias, información, resultados y tecnologías. Además, promociona la investigación e innovación como herramientas

esenciales para el desarrollo tecnológico y social, así como para la modernización productiva y el aumento de la competitividad para el desarrollo económico.

En la II cumbre Iberoamericana de jefes de estado y de gobierno [134], celebrada en Madrid, en Julio de 1992, se aprobó una resolución que incluye el siguiente texto: "*En el campo de la investigación científica y de la innovación tecnológica, la conferencia, a la vista de los logros alcanzados desde su creación por el programa Iberoamericano de ciencia y tecnología para el desarrollo. Quinto centenario (CYTED), así como de la opinión de todos los países participantes, aprobó su fortalecimiento y continuidad, como instrumento válido de integración.*"

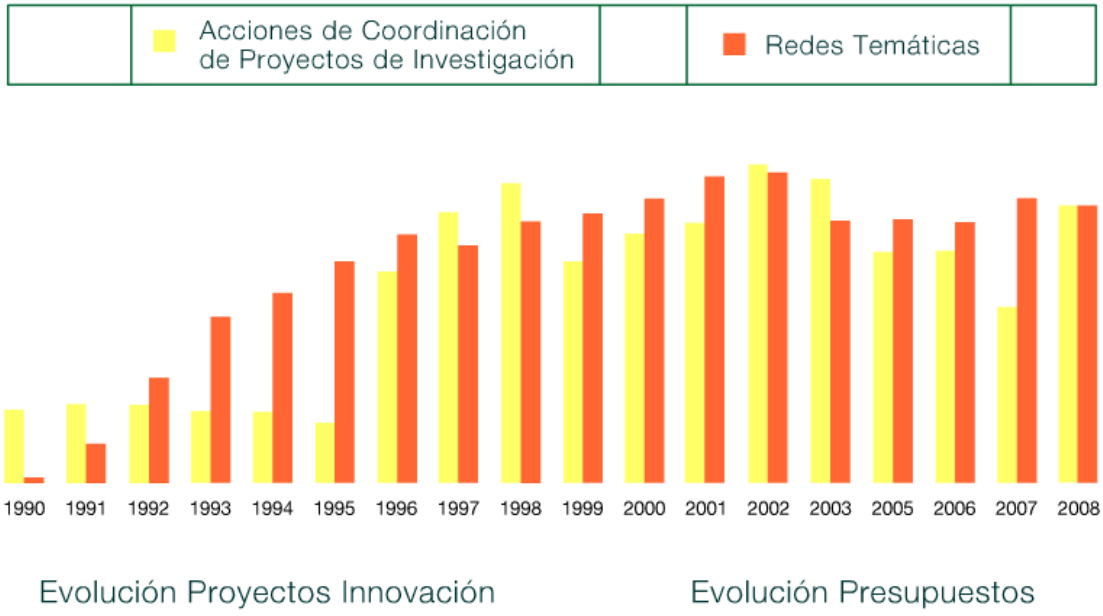

Ilustración 10 Evolución del número de actividades de CYTED

Desde 1993 se han organizando anualmente las conferencias científicas preparatorias de las cumbres iberoamericanas de Jefes de Estado y de Gobierno y se encuentra formalmente incluido entre los programas de cooperación de las cumbres iberoamericanas de jefes de estado y de gobierno. Hasta la fecha (año 2009) se han generado 191 redes temáticas, 193 acciones de coordinación, 3 proyectos de investigación consorciados y 614 proyectos de innovación IBEROEKA con una participación anual de más de 10.000 científicos y tecnólogos iberoamericanos. En la "Ilustración 10 Evolución del número de actividades de CYTED" se muestra la evolución que ha tenido el programa CYTED desde sus comienzos hasta el año 2008.

10.2 CYTED: Tecnologías de la Información y las Comunicaciones

Esta área incluye todas aquellas ciencias, tecnologías y aplicaciones relacionadas con la adquisición, almacenamiento, tratamiento, comunicación, difusión, y uso de la información entendida en su sentido más amplio.

La aplicación generalizada de estas tecnologías en los distintos ámbitos ha generado, especialmente en los países más desarrollados, una transformación social profunda hasta tal punto que se ha llegado a llamar "Sociedad de la Información". Representa un salto cualitativo, una "revolución digital", que sucede a la revolución industrial y que imprime un ritmo evolutivo mucho mayor que los desarrollos tecnológicos del pasado. Por su parte, el día a día está demostrando que, a pesar del sentido de igualación social que podría ingenuamente esperarse de estas nuevas tecnologías, el dominio de las mismas por parte de los países desarrollados y la rapidez del cambio contribuyen a aumentar decisivamente la distancia entre países desarrollados y aquellos en vías de desarrollo.

Desde este punto de partida, los objetivos y actividades que se plantean en esta área en el marco del Programa CYTED se encaminan a intentar disminuir en lo posible esta brecha. Sobre la base del trabajo en común, los objetivos estratégicos que se plantean se focalizan en mejorar el nivel de formación y capacitación, en la identificación de posibles nichos de mercado asociados a sectores económicos estratégicos para la región, en la puesta en marcha de proyectos de investigación y desarrollo orientados en esta dirección que se apoyen en las infraestructuras nacionales de cada uno de los países signatarios y en asegurar en lo posible una transferencia tecnológica básicamente hacia las PYMES de los sectores industriales que constituyen mayoritariamente el tejido industrial de nuestros países.

Así pues, el objetivos del área es reducir la brecha digital existente entre los diferentes países de la región Iberoamericana, para mejorar el nivel de formación y capacitación; identificar nichos de mercado asociados a sectores económicos estratégicos para la región, apoyándose en las infraestructuras nacionales de cada uno de los países signatarios; asegurar una transferencia de tecnología básicamente hacia las PYMES;

realizar esfuerzos hacia aquellas actividades en las que exista o sea de prever una fuerte dependencia tecnológica, exista o pueda crearse una infraestructura empresarial capaz de asumir los desarrollos alcanzados, y se puedan transferir los resultados a otros países.

10.3 Calidad en los sitios del área de la salud

El proyecto, con el código de referencia en CYTED 508RT0361 y nombre original del portugués "*qualidade em sites na área da saúde salus*" (ver "Apéndice A: A qualidade em sites na área da saúde salus" para obtener todos los detalles del documento original, en portugués, del proyecto), consiste en un conjunto de actividades centradas en la investigación para validar una plataforma de consultas y recomendaciones sobre la web semántica con especial énfasis en los sitios de referentes a la salud. Esta plataforma se ha diseñado como un portal web con accesos adaptables a los diferentes usuarios en función de su nivel cultural y de las particularidades regionales.

Los principales temas de investigación del proyecto se centran en la web, los sistemas distribuidos, sistemas de bases de datos, consultas en archivos XML, servicios web, sistemas de información en salud y los sistemas de recomendación. Durante los cuatro años propuestos para el desarrollo del proyecto (2008-2011) se pretende que los grupos involucrados (ver "10.3.2 Grupos representantes de la unidad de investigación") desarrollen una cooperación bi o trilaterales con el objetivo de crear proyectos específicos para la financiación de la investigación y, sobre todo, el desarrollo y finalización del sistema mencionado.

10.3.1 Objetivo

El objetivo principal del proyecto consiste en la creación de un portal web con acceso adaptable de los usuarios en función de su nivel cultural y aspectos regionales. El contenido del portal web consistirá en información sobre "sitios web relacionados con temas de la salud humana". Esta información, de los sitios web relacionados con temas de la salud humana, tendrá especial énfasis en aquellos aspectos relacionados con (A) la calidad y educación del público y (B) credibilidad y opiniones de los usuarios. Además, estos sitios web deberán referirse o contener información sobre: Cuidado de

la salud de los pacientes, tratamiento de enfermedades y prevención para la población en general.

Si bien ya existen proyectos como "*Information Therapy*" [135] que abordan estos aspectos, lo que se pretende con la actual proposición de proyecto es dar una visión más amplia y ofrecer este tipo de información tanto a los pacientes como al público en general. Más concretamente el objetivo del proyecto es crear un "mostrador" al que se le pasa una página web o URL relacionada con temas de salud y mediante la minerización de la URL presentar en un "*dashboard*" los principales indicadores sobre el sitio web así como las opiniones de los grupos de usuarios.

10.3.2 Grupos representantes de la unidad de investigación

A continuación en "Tabla 6 Grupos representantes de la unidad de investigación" se muestra una tabla con los grupos de trabajo que representan la unidad de investigación del proyecto. El coordinador del grupo es José Palazzo Moreira De Oliveira (palazzo@inf.ufrgs.br), profesor de la Universidade Federal Do Rio Grande Do Sul.

País	Universidad y coordinador
Argentina	Universidad Nacional de La Plata (UNLP). Gustavo Rossi (gustavo@sol.info.unlp.edu.ar)
Brasil	Universidade Estadual de Londrina (UEL). Mário Lemes Proença Jr. (proenca@uel.br)
Chile	Pontifícia Universidad Católica de Chile (PUC-CHILE). David Fuller Padilla (dfuller@ing.puc.cl)
Colombia	Universidad del Cauca (UNICAUCA). Álvaro Rendón Gallón (arendon@unicauca.edu.co)
España	Universidad Politécnica de Valencia (UPV). Oscar Pastor (opastor@dsic.upv.es)
España	Universidad Politécnica de Catalunya (UPC). Alberto Abelló Gamazo (aabello@lsi.upc.edu)
Portugal	Faculdade de Engenharia da Universidade do Porto (FEUP). João Falcão E Cunha (jfcunha@fe.up.pt)
Uruguay	Universidad de La república, Instituto de Computación (INCO). Motz Carrano Maria Regina (rmotz@fing.edu.uy)
Uruguay	Evimed Limitada (EVIMED). Alvaro Margolis Hirt (alvaro.margolis@evimed.net)

Tabla 6 Grupos representantes de la unidad de investigación

10.3.3 El papel de los sistemas de recomendación

La parte correspondiente al proyecto de los sistemas de recomendación consiste en la implementación y testeo de un sistema de recomendación automático de contenidos relacionados con temas de salud. A partir de una lista de documentos clasificados por sus características y basándose en el perfil del usuario, el sistema debe realizar una recomendación al usuario.

El objetivo principal del sistema de recomendación es ayudar a las personas a encontrar la información que necesitan, de manera pro-activa, es decir, permitiendo detectar las necesidades de los usuarios sin que éstos precisen formular-las. Además de la recolección de datos del usuario de forma implícita (por ejemplo, analizando las acciones del usuario en el sistema), también se puede realizar de forma explícita (cuando el usuario proporciona datos sobre sí mismos).

Por otro lado, es necesario distinguir entre los diferentes usuarios (nivel de conocimiento o especialización sobre temas de salud) así como el nivel de los diferentes documentos a recomendar, ya sean básicos o avanzados, con el objetivo de recomendar documentos que se ajusten al nivel, formación o especialización del usuario. Por ejemplo, un doctor en un área específica no puede recibir las mismas recomendaciones que un principiante en este ámbito. Del mismo modo, que un documento básico no se recomienda a un especialista o un tema avanzado no se recomienda a un usuario novel en el área. Una forma de abordar estas cuestiones es mediante la identificación del nivel o grado de conocimiento de cada persona o usuarios en las áreas de temas específicos. Inicialmente se proponen cinco tipos distintos de usuarios conforme a su nivel de experiencia (ver “6.3.5 Recomendação” del “Apéndice A: A qualidade em sites na área da saúde salus”):

- **Usuarios expertos:** Aquellos con más tiempo de actividad en el sistema.
- **Usuarios activos:** Actúan con regularidad con el sistema además de realizar distintas acciones.
- **Usuarios contribuyentes (de centros de estudios):** Son aquellos que ayudan activamente a la lectura y almacenamiento de documentos en la biblioteca digital.

- **Especialistas:** Los que confirma su plan de estudios de especialidad y su alto grado de conocimientos.
- **Autoridades:** Aquellos que son nombrados por otros usuarios como referencias en algunos grupos de usuarios.

10.4 Evaluación del sistema de recomendación

Teniendo en cuenta el objetivo del sistema de recomendación para el proyecto de calidad en los sitios del área de la salud focalizado en el apartado anterior "10.3.3 El papel de los sistemas de recomendación", a continuación se presenta una evaluación de los distintos sistemas de recomendación estudiados anteriormente.

10.4.1 Sistemas basados en el contenido

Desde el punto de vista de que el proyecto requiere de un sistema de recomendación automático de contenidos relacionados con temas de salud que a partir de una lista de documentos clasificados por sus características y basándose en el perfil del usuario, se deben realizar recomendaciones al usuario, los sistemas de recomendación basados en el contenido se adaptan perfectamente a este proyecto.

En el proceso de recomendación basado en el contenido, es necesario la adquisición de atributos o propiedades de los ítems (páginas webs, documentos o revistas), para su posterior procesado. Este proceso se realiza por el grupo de proyecto que trabaja en el área de minería de datos web. Así como se ha comentado anteriormente, en el apartado del estado del arte de los sistemas de recomendación basados en el contenido "3.1.1 Representación de los ítems", las siguientes referencias contienen información sobre proceso de extracción de información: [66-69].

Por otra parte, cabe mencionar que los sistemas basados en contenido adolecen de tres tipos de problemas:

- **Limitación por el análisis del contenido:** Hace referencia a que la recomendación sólo se basa en los atributos de los ítems. Si a los ítems no se les asignan una serie de propiedades, es imposible realizar una recomendación del ítem. Es posible que las técnicas de minería no puedan obtener datos de ciertas páginas webs de forma automática por la forma de la arquitectura de la

misma (por ejemplo, páginas web con contenido java o flash). En este caso, es necesaria la evaluación manual realizada por especialistas del sector o bien utilizar técnicas colaborativas para la eliminación de este problema.

- **Sobre-especialización:** La sobre-especialización aparece cuando el sistema sólo muestra al usuario ítems similares a los que ya ha visto antes. En este caso, dado que el usuario puede cambiar las preferencias de su perfil, este problema es evitable. Además, si se complementa el sistema de recomendación, basado en el contenido, con el sistema de recomendación de filtro colaborativo, la sobre-especialización se puede eliminar.
- **Usuario nuevo:** Un usuario tiene que haber valorado algunos ítems antes de que el sistema pueda saber de sus gustos y preferencias para poderle recomendar. No obstante, al igual que en el caso anterior, el usuario puede indicar sus preferencias al editar su perfil. Por consiguiente, de esta forma el problema quedaría solventado. Además, la posibilidad de utilizar técnicas de recomendación basadas en el conocimiento puede ayudar a mitigar este factor.

Teniendo en cuenta los argumentos anteriormente expuestos, se puede concluir que este tipo de sistema de recomendación es perfectamente adaptable al proyecto. No obstante, éste debería de combinarse con otros sistemas de recomendación a modo de reducción de ciertos problemas o inconvenientes inherentes a los sistemas de recomendación basados en el contenido.

10.4.2 Sistemas de filtro colaborativo

Dadas las características del proyecto, es posible la utilización de los sistemas de recomendación de filtro colaborativo. Así como muestra la “Tabla 4 Comparativa sistemas de recomendación (a favor)” del apartado “9.1 Comparación”, este tipo de sistemas tiene dos propiedades bastante útiles para poder ser aplicadas en el proyecto de CYTED: (A) Adaptabilidad (no se sobre-especializa) y (B) diferenciación de ítems homogéneos (con los mismos atributos).

En el sentido de que las recomendaciones de las páginas web siempre serán de un mismo ámbito –el sanitario– seguramente los sistemas de clasificación de documentos (minería de datos) realizarán el etiquetado de diferentes documentos con los mismos

atributos y, por lo tanto, aparecerán ítems o páginas webs con los mismos atributos o propiedades indistinguibles entre sí dado que mostraran las mismas cualidades o propiedades (problema de la homogeneidad). En estas situaciones, es donde los sistemas de recomendación de filtro colaborativo pueden ayudar a distinguir entre unos elementos u otros aportando la perspectiva de calidad del ítem en las recomendaciones y eliminando el problema de la homogeneidad mencionado anteriormente. Por otro lado, los sistemas de filtrado colaborativo también permiten que las recomendaciones se estabilicen, factor a tener en cuenta para las recomendaciones de sitios web con usuarios ocasionales (aquellos que no buscan un tipo de información en concreto).

Por otro lado, este tipo de sistemas presentan una serie de inconvenientes como puede ser el arranque en frío tanto para un nuevo usuario como para el nuevo ítem. No obstante, teniendo en cuenta los requisitos del proyecto, los mayores problemas que muestran este tipo de sistemas de recomendación son dos: (1) Los usuarios con gustos poco comunes y (2) la confianza entre usuarios. El proyecto de CYTED, "Calidad en los sitios del área de la salud", pretende ser utilizado por una gran diversidad de usuarios, en su mayoría con conocimientos generales sobre la salud. También será utilizado por los profesionales del sector cuyas preferencias son de un tipo muy selecto, es decir, los ítems a recomendarles deben ser específicos, precisos, así como especializados. El uso de sistemas de recomendación de filtrado colaborativo puede hacer que los usuarios especializados padezcan de discriminación o falta de recomendaciones por parte del sistema dadas sus preferencias poco comunes (debido su especialización en la temática y la dificultad de encontrar grupos de usuarios vecinos afines a su perfil). El problema es difícil de solventar por lo que no sería aconsejable la utilización, como motor de sistema de recomendación principal, un sistema de recomendación de filtrado colaborativo.

Por lo que se refiere al otro inconveniente de los sistemas de filtrado colaborativo mencionado anteriormente –la confianza entre usuarios–, el proyecto cuenta con distintos tipos de usuarios, por lo que se podría crear una red de confianza con pesos o ponderaciones dependiendo del tipo de usuario que se trate.

10.4.3 Recomendaciones basadas en el conocimiento

Los sistemas de recomendación basados en el conocimiento son aparentemente bastante útiles en el proyecto dado que solventan el problema del arranque en frío además de ser sensibles a las preferencias de los usuarios. No obstante, sus recomendaciones son de carácter estático, aspecto no deseable en el proyecto (los entornos web varían constantemente y rápidamente, siendo éste pues, un entorno no apto para los sistemas de recomendación basados en el conocimiento). Además, el inconveniente de tener que implementar una ingeniería del conocimiento hace que se replantee su uso en el proyecto.

Teniendo en cuenta los aspectos mencionados en los apartados anteriores "10.4.1 Sistemas basados en el contenido" y "10.4.2 Sistemas de filtro colaborativo", el papel de los sistemas de recomendación basados en el conocimiento sería fundamental para la eliminación del arranque en frío tanto para un nuevo usuario como para un nuevo ítem. No obstante, una vez el sistema haya recopilado la suficiente información para poder prescindir del de sistema de recomendación basado en el conocimiento, su uso se vería relevado a un segundo plano, llegando incluso a su des utilización. Es por este motivo, y teniendo en cuenta la necesidad de la creación de una ingeniería del conocimiento, que el uso de este tipo de sistemas es desaconsejable en el proyecto. Además, así como se ha comentado en el estado del arte "5.4 Conclusiones", su uso se restringe a dominios muy precisos con pocas variaciones a lo largo del tiempo. Consecuente mente, el uso de sistemas basados en el conocimiento, no encaja dentro del entorno (dominio) web, donde la información es muy diversa así como sus variaciones son rápidas a lo largo del tiempo.

10.4.4 Sistemas de recomendación semánticos

El proyecto de CYTED está enfocado al uso de la web como principal medio de comunicación. Los sistemas de recomendación semánticos son el tipo de sistema de recomendación más adaptado al contexto planteado por el proyecto de CYTED ya que basan su funcionamiento sobre una base de conocimiento, normalmente definida a través de un esquema de conceptos (como una taxonomía o un tesauro) o una ontología, y que utilizan tecnologías de Web Semántica. Son estos dos últimos

aspectos, el uso de ontologías y tecnologías de web semántica, nociones presentes en el proyecto. Motivo por el cual, este tipo de sistemas pueden aportar gran conocimiento útil para la aplicación del sistema de recomendación.

De los tres tipos de sistemas de recomendación semánticos, son los basados en ontologías o esquemas de conceptos y los sistemas basados en redes de confianza los más importantes para el proyecto. Este tipo de sistemas aportan el conocimiento necesario para el trabajo con ontologías (utilizadas para la definición de los usuarios) y las redes de confianza (así como se comenta en "8 Seguridad en los sistemas de recomendación", son necesarias en las redes colaborativas como medida de protección contra los ataques a los sistemas de filtro colaborativo).

10.4.5 Conclusiones

De los diferentes sistemas de recomendación valorados en los apartados anteriores, se puede extraer que existen tres tipos de sistemas de recomendación que aportan una parte de la solución para el sistema de recomendación del proyecto de CYTED. En primer lugar, los sistemas de recomendación basados en el contenido se adaptan perfectamente a las condiciones del proyecto, así mismo, los sistemas de recomendación colaborativo, permitirían eliminar problemas como la sobre especialización.

Por otra parte, los sistemas de recomendación semánticos son un tipo de sistema de recomendación basado en la web semántica de los cuales se encuentra mucha información en la literatura sobre su implementación así como trabajo con ontologías y redes de confianza. Es por ello y en este sentido que también pueden realizar importantes contribuciones para determinar el "road-map" definitivo para la creación del sistema de recomendación del proyecto de CYTED.

3º Acto: Cierre

Proposición del Sistema de recomendación

La constancia de pequeños detalles construye las grande cosas

11“Road-map”

A continuación se describen las diferentes partes de un sistema de recomendación desde el punto de vista del proyecto de CYTED. El objetivo de este capítulo es la proposición de un “*road-map*” para la adopción o implementación de un sistema de recomendación particular para el proyecto mencionado anteriormente en el apartado “10 Calidad en los sitios del área de la salud”.

11.1 Elementos de entrada

Los elementos de entrada de un sistema de recomendación hacen referencia a la información que se le proporciona al sistema por parte de los usuarios y de los ítems a recomendar. Así pues, se tienen dos tipos diferentes de elementos de entrada en un sistema de recomendación: usuarios e ítems. Para cada tipo de elemento de entrada es necesario crear una estructura de datos que permita almacenar los datos relativos a éstos así como las relaciones entre ambos (usuarios e ítems). Otro elemento a tener en cuenta es la forma en que se obtiene la información o propiedades de cada elemento de entrada al sistema. La forma en que se utiliza esta información corresponde a la manera en que el sistema de recomendación maneja la información para procesar las recomendaciones. A continuación se presenta la mejor aproximación posible para determinar la estructura y forma de obtención de propiedades de los dos tipos de elementos de entrada del sistema de recomendación. La utilización la información del sistema será analizada en el apartado “11.3 Método de generación de las recomendaciones”.

11.1.1 Usuarios

Teniendo en cuenta la forma en que se ha planteado el proyecto y el estudio realizado en el estado del arte, se recomienda utilizar una estructura del tipo “*Friend of a friend*” (FOAF) la cual provee de un “*framework*” para la representación de información sobre las personas, sus intereses y relaciones entre estas y las conexiones sociales.

Anteriormente, en el apartado “10.3.3 El papel de los sistemas de recomendación” se ha visto que existen, a priori, cinco tipos distintos de usuarios del sistema que se encuentran registrados. No obstante, si es posible que un usuario acceda a las

facilidades del sistema, sería conveniente considerar la creación de perfiles temporales los cuales corresponderían a aquellos usuarios que acceden a la página web y no se identifiquen como miembros de ésta. La principal característica de estos usuarios, con respecto al perfil de usuario, es que la información obtenida a partir de ellos carece de total credibilidad o confianza dado que no están identificados.

El proyecto contempla dos formas de obtener información sobre el perfil del usuario: De forma implícita y explícita. Para la adquisición de la información de forma explícita, se puede crear un formulario donde el mismo usuario puede cambiar o modificar sus preferencias, además, es posible obtener información de este tipo mediante el uso de técnicas de *feedback*. Por ejemplo, sería de gran utilidad para el sistema preguntar le al usuario, una vez accedido o seleccionado la recomendación, el grado de utilidad de ésta así como el nivel de complejidad de la información seleccionada/recomendada. Por lo que respeta a la adquisición de información de forma implícita, el sistema deberá monitorizar las acciones del usuario en todo momento. Los aspectos más comunes a ser monitorizados de forma implícita son tres: Determinar la zona geográfica donde se encuentra el usuario (mediante técnicas de rastreo de IP), tener en cuenta los recursos que selecciona el usuario y medir el tiempo que le dedica a cada uno de ellos.

11.1.2 Ítems

En el contexto del proyecto “Calidad en los sitios del área de la salud”, los ítems corresponden a un tipo información que se proporciona al usuario en forma de sugerencia, la cual abarca desde un simple link a una determinada web, hasta páginas de un libro, papers, documentos, revistas o cualquier recurso web que pueda ser accesible desde internet.

La parte correspondiente a la búsqueda tanto de ítems como de sus atributos, así como se ha comentado en los apartados “10.3 Calidad en los sitios del área de la salud” y “10.4.1 Sistemas basados en el contenido”, corresponde al grupo de minería de datos. No obstante, debido a la presencia del sistema de recomendación de filtro colaborativo, se recomienda la creación de una estructura de datos que además de permitir la clasificación de los ítems y la asignación de atributos (derivados del proceso

de minería), se tenga en cuenta que los usuarios puedan realizar valoraciones o dar a conocer su opinión personal de cada ítem. En el apartado "3.1.1 Representación de los ítems" se han nombrado diferentes referencias que pueden ayudar al campo de minería de datos en la abstracción de información de forma automática de diferentes ítems.

11.2 Elementos de salida

La salida del sistema de recomendación corresponde a las recomendaciones generadas por el sistema, que variarán dependiendo del tipo, cantidad y formato de la información proporcionada al usuario. Existen diferentes formas de presentarle esta información al usuario. La forma más común de presentar las recomendaciones al usuario es mediante una lista de ítems que el sistema de recomendación determina que son recomendables para el usuario en cuestión dependiendo de su perfil. No obstante, también se le pueden mostrar al usuario predicciones del grado de satisfacción que se asignará al ítem concreto. Estas estimaciones pueden ser presentadas como personalizadas al usuario o como estimaciones generales del conjunto de colaboradores. Por otra parte, para los usuarios que no estén identificados en el sistema, se pueden mostrar listas de ítems como por ejemplo: "los 10 ítems/documentos más vistos" o "los documentos más valorados durante la última semana".

11.3 Método de generación de las recomendaciones

Así como se ha comentado anteriormente en los apartados "10.4 Evaluación del sistema de recomendación" y "10.4.5 Conclusiones", el tipo de sistema de recomendación a implementar podría ser una combinación de los sistemas basados en el contenido y los filtros colaborativos. La combinación de ambos se describe a continuación:

En el momento de la iniciación del sistema, cuando aún no se tienen los suficientes datos de los usuarios para presentarle recomendaciones, se podrían presentar las recomendaciones única y exclusivamente basándose en el contenido de los ítems. Posteriormente, a medida que vayan aumentando las opiniones de los usuarios sobre

los ítems, se podría utilizar una combinación de resultados de ambos sistemas de recomendación (ver "Ilustración 11 Motor del sistema de recomendación (1ª fase)").

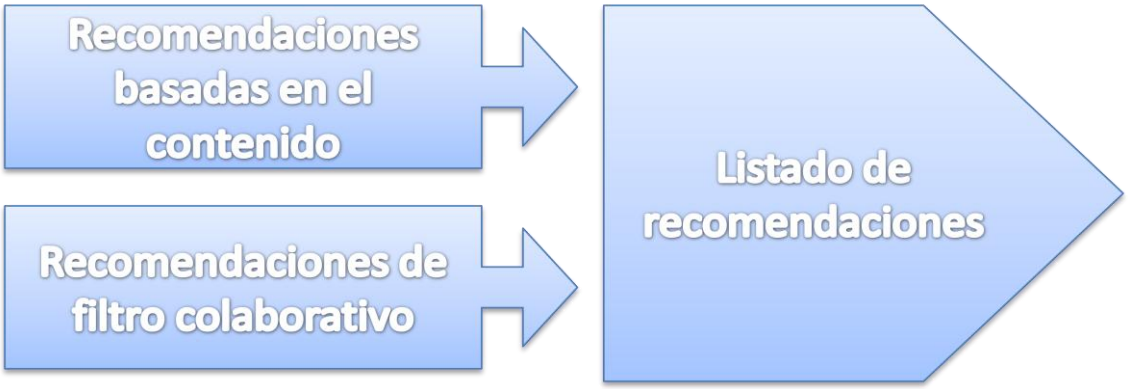

Ilustración 11 Motor del sistema de recomendación (1ª fase)

Por último, en cuanto se tengan las suficientes valoraciones de usuarios y una gran cantidad de ítems en el sistema para recomendar, se podría seguir utilizando el mismo sistema, pero filtrando las recomendaciones mediante la combinación en cascada (ver "7.2.5 En cascada (*cascade*)") de ambos sistemas de recomendación.

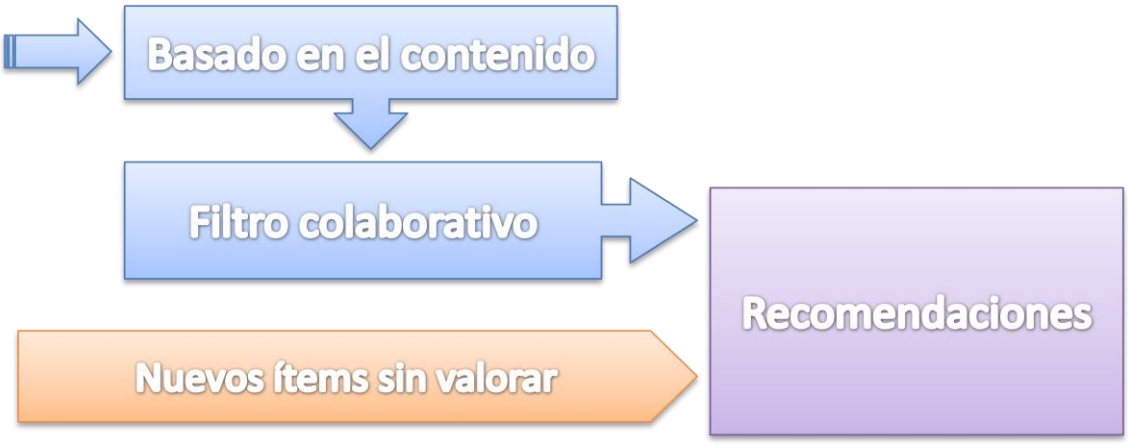

Ilustración 12 Motor del sistema de recomendación (2ª fase)

El primer sistema de recomendación sería el basado en el contenido y el segundo sistema de recomendación, el cual permitiría discriminar los resultados menos relevantes obtenidos con el primer sistema de recomendación, el sistema de recomendación de filtro colaborativo. Una vez implementado el sistema en cascada, se le podría presentar al usuario las recomendaciones derivadas de este proceso además de algunos ítems nuevos añadidos al sistema (siempre y cuando tengan relación con la información que está buscando el usuario) que no tengan valoraciones por parte de los

usuarios (ver “Ilustración 12 Motor del sistema de recomendación (2ª fase)”). De este modo, al añadir un nuevo ítem o documento al sistema, se conseguiría obtener valoraciones de los usuarios y por consiguiente de aceleraría el proceso para determinar si el nuevo ítem del sistema es apropiado para recomendar (obtiene valoraciones positivas) o si por el contrario no se debe recomendar (obtiene valoraciones negativas).

11.4 Estructura basada en agentes

El sistema de recomendación que se propone está basado en la utilización de diferentes agentes (ver “6.1.2 Agentes”) para la realización del proceso de recomendación. Esta es una práctica que aparece con los sistemas de recomendación semánticos y que proporciona todas las técnicas y aspectos necesarios para tratar adecuadamente el carácter dinámico de los sistemas de recomendación. Por otra parte, la implementación del sistema de recomendación basado en agentes permite: La realización de sistemas distribuidos capaces de realizar tareas complejas a través de cooperación e interacción entre agentes además de un análisis teórico y experimental de mecanismos de auto-organización y adaptación que tienen lugar cuando las entidades autónomas interaccionan. En la “Ilustración 13 Estructura sistema de recomendación” se muestra la arquitectura del sistema de recomendación con el papel de los agentes.

En el diagrama se pueden distinguir tres partes que corresponden a (1) el motor del sistema de recomendación (explicado anteriormente en el apartado “11.3 Método de generación de las recomendaciones”); (2) la base de datos que contiene tanto la información de los ítems del sistema, los perfiles de usuarios y las relaciones entre ítems y usuarios (que son las valoraciones que realizan los usuarios a los diferentes ítems) y (3) los agentes que actúan de intermediarios entre el sistema de recomendación, los usuarios y la base de datos.

En el esquema mostrado en la “Ilustración 13 Estructura sistema de recomendación” se puede observar que la base de datos está dividida en tres categorías (Ontologías, Relaciones y perfiles de usuario) para facilitar una mayor comprensión de la

arquitectura del sistema de recomendación, no obstante, no es necesario que en la práctica la base de datos sea estrictamente definida en estas tres partes.

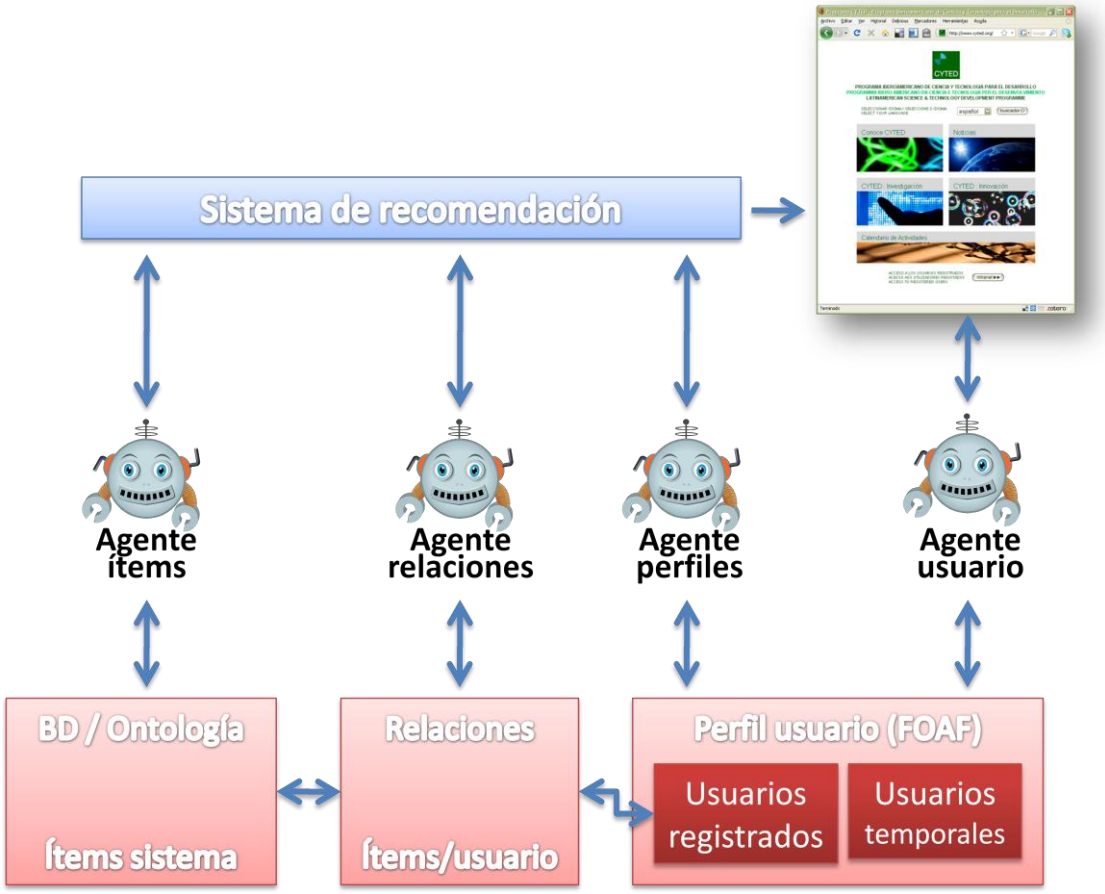

Ilustración 13 Estructura sistema de recomendación

Por último, destacar la figura del agente del usuario, que además de tener la responsabilidad de interactuar con los perfiles de usuario, debería poder observar las acciones que de forma implícita el usuarios realiza sobre el sistema para así, de este modo, poder obtener información relativa a las preferencias o gustos del usuarios sin la necesidad de que el usuario la proporcione explícitamente.

12 Bibliografía

[1] U. Hanani, B. Shapira, y P. Shoval, "Information filtering: Overview of issues, research and systems," *User Modeling and User-Adapted Interaction*, vol. 11, 2001, págs. 203-259.

[2] N.J. Belkin, "Helping people find what they don't know," 2000.

[3] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, y J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews."

[4] D.S. Stodolsky, *Consensus journals: invitational journals based upon peer consensus*, Roskilde University Centre, Computer Science, 1990.

[5] D. Goldberg, D. Nichols, B.M. Oki, y D. Terry, "Using collaborative filtering to weave an information tapestry," 1992.

[6] P. Resnick y H.R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, 1997, pág. 57.

[7] E.M. Housman y E.D. Kaskela, "State of the art in selective dissemination of information," *IEEE Transactions on Engineering Writing and Speech*, vol. 13, 1970, págs. 78-83.

[8] K.H. Packer y D. Soergel, "The importance of SDI for current awareness in fields with severe scatter of information," *Journal of the American Society for Information Science*, vol. 30, 1979.

[9] R.B. Allen, "User models: theory, method, and practice," *International Journal of Man-Machine Studies*, vol. 32, 1990, págs. 511-543.

[10] P.W. Foltz y S.T. Dumais, "Personalized information delivery: An analysis of information filtering methods," *Communications of the ACM*, vol. 35, 1992, págs. 51-60.

[11] T.W. Malone, K.R. Grant, K.Y. Lai, R. Rao, y D. Rosenblitt, "Semistructured messages are surprisingly useful for computer-supported coordination," *ACM Transactions on Office Information Systems*, vol. 5, 1987, págs. 115-131.

[12] W.E. Mackay, T.W. Malone, K. Crowston, R. Rao, D. Rosenblitt, y S.K. Card, "How do experienced Information Lens users use rules?," *Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, ACM New York, NY, USA, 1989, págs. 211-216.

[13] N. Manouselis y C. Costopoulou, "Analysis and Classification of Multi-Criteria Recommender Systems," *World Wide Web*, vol. 10, 2007, págs. 415-441.

[14] J.B. Schafer, J.A. Konstan, y J. Riedl, "E-commerce recommendation applications," *Data mining and knowledge discovery*, vol. 5, 2001, págs. 115-153.

[15] K. Goldberg, T. Roeder, D. Gupta, y C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *Information Retrieval*, vol. 4, 2001, págs. 133-151.

[16] J.L. Herlocker, J.A. Konstan, L.G. Terveen, y J.T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, 2004, págs. 5-53.

[17] M. Deshpande y G. Karypis, "Item-based top-n recommendation algorithms," *ACM Transactions on Information Systems (TOIS)*, vol. 22, 2004, págs. 143-177.

[18] P. Han, B. Xie, F. Yang, y R. Shen, "A scalable p2p recommender system based on distributed collaborative filtering," *Expert systems with applications*, vol. 27, 2004, págs. 203-210.

[19] Y.S. Kim, B.J. Yum, J. Song, y S.M. Kim, "Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites," *Expert Systems with Applications*, vol. 28, 2005, págs. 381-393.

[20] S.H. Min y I. Han, "Detection of the customer time-variant pattern for improving recommender systems," *Expert Systems with Applications*, vol. 28, 2005, págs. 189-199.

[21] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, 2002, págs. 331-370.

[22] N.J. Belkin y W.B. Croft, "Information filtering and information retrieval: two sides of the same coin?," 1992.

[23] M. Balabanovic y Y. Shoham, "Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, 1997, pág. 67.

[24] G. Adomavicius y A. 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*, vol. 17, 2005, págs. 734-749.

[25] M. Van Setten, "Supporting people in finding information: hybrid recommender systems and goal-based structuring," 2005.

[26] B. Krulwich y C. Burkey, "Learning user information interests through extraction of semantically significant phrases," *Proceedings of the AAAI spring symposium on machine learning in information access*, 1996, págs. 100–112.

[27] K. Lang, "Newsweeder: Learning to filter netnews," *In Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[28] W. Hill, L. Stead, M. Rosenstein, y G. Furnas, "Recommending and evaluating choices in a virtual community of use," *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1995, págs. 194-201.

[29] U. Shardanand y P. Maes, “Social information filtering: algorithms for automating "word of mouth",” *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1995, págs. 210-217.

[30] T.W. Malone, K.R. Grant, F.A. Turbak, S.A. Brobst, y M.D. Cohen, “Intelligent information-sharing systems,” 1987.

[31] E. Rich, “Users are individuals: individualizing user models,” *International journal of man-machine studies*, vol. 18, 1983, págs. 199-214.

[32] H.J. Ahn, “A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem,” *Information Sciences*, vol. 178, 2008, págs. 37-51.

[33] R. Burke, “Knowledge-based recommender systems,” *Encyclopedia of Library and Information Systems*, vol. 69, 2000, págs. 175-186.

[34] N. Manouselis y D. Sampson, “Recommendation of Quality Approaches for the European Quality Observatory,” *Proc. of ICALT*, 2004.

[35] T. Bemers-Lee, J. Hendler, y O. Lassila, “The semantic web,” *Scientific American*, vol. 284, 2001, págs. 34-43.

[36] C. Basu, H. Hirsh, y W. Cohen, “Recommendation as classification: Using social and content-based information in recommendation,” *Proceedings of the National Conference on Artificial Intelligence*, JOHN WILEY & SONS LTD, 1998, págs. 714-720.

[37] J.B. Schafer, “The application of data-mining to recommender systems,” *Encyclopedia of data warehousing and mining*, 2002, págs. 44–48.

[38] S. Perugini, M.A. Gonçalves, y E.A. Fox, “Recommender systems research: A connection-centric survey,” *Journal of Intelligent Information Systems*, vol. 23, 2004, págs. 107-143.

[39] C.P. Wei, M.J. Shaw, y R.F. Easley, “Recommendation Systems in Electronic Commerce,” *E-Service: new directions in theory and practice*, 2002, pág. 168.

[40] M. Montaner, B. López, y J.L. De La Rosa, “A taxonomy of recommender agents on the internet,” *Artificial intelligence review*, vol. 19, 2003, págs. 285-330.

[41] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, y M. Sartin, “Combining content-based and collaborative filters in an online newspaper,” *Proceedings of ACM SIGIR Workshop on Recommender Systems*, 1999.

[42] N. Manouselis y D. Sampson, “A multi-criteria model to support automatic recommendation of e-learning quality approaches,” *Proc. of EDMEDIA*, 2004.

[43] P. Karampiperis y D. Sampson, “Adaptive learning resources sequencing in educational hypermedia systems,” *Educational Technology & Society*, vol. 8, 2005, págs. 128-147.

[44] B. Roy, *Multicriteria methodology for decision aiding*, Kluwer Academic Pub, 1996.

[45] Y.H. Cho, J.K. Kim, y S.H. Kim, "A personalized recommender system based on web usage mining and decision tree induction," *Expert Systems with Applications*, vol. 23, 2002, págs. 329-342.

[46] J.S. Lee, C.H. Jun, J. Lee, y S. Kim, "Classification-based collaborative filtering using market basket data," *Expert Systems with Applications*, vol. 29, 2005, págs. 700-704.

[47] W.P. Lee, C.H. Liu, y C.C. Lu, "Intelligent agent-based systems for personalized recommendations in Internet commerce," *Expert systems with applications*, vol. 22, 2002, págs. 275-284.

[48] W. Lihua, L. Lu, L. Jing, y L. Zongyong, "Modeling user multiple interests by an improved GCS approach," *Expert Systems With Applications*, vol. 29, 2005, págs. 757-767.

[49] W.P. Lee y T.H. Yang, "Personalizing information appliances: a multi-agent framework for TV programme recommendations," *Expert Systems with Applications*, vol. 25, 2003, págs. 331-341.

[50] B.J. Mirza, B.J. Keller, y N. Ramakrishnan, "Studying recommendation algorithms by graph analysis," *Journal of Intelligent Information Systems*, vol. 20, 2003, págs. 131-160.

[51] S.E. Middleton, N.R. Shadbolt, y D.C. De Roure, "Ontological user profiling in recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, 2004, págs. 54-88.

[52] S.W. Changchien y T.C. Lu, "Mining association rules procedure to support on-line recommendation by customers and products fragmentation," *Expert Systems with Applications*, vol. 20, 2001, págs. 325-335.

[53] Y.H. Cho y J.K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, vol. 26, 2004, págs. 233-246.

[54] L. Terveen, W. Hill, B. Amento, D. McDonald, y J. Creter, "PHOAKS: A system for sharing recommendations," *Communications of the ACM*, vol. 40, 1997, págs. 59-62.

[55] J. Zimmerman, L. Parameswaran, y K. Kurapati, "Celebrity recommender," *Proceedings of the 2nd Workshop on Personalization in Future TV, Malaga, Spain*, 2002, págs. 33-41.

[56] J.S. Breese, D. Heckerman, y C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Learning*, vol. 9, 1992, págs. 309-347.

[57] J. Herlocker, J.A. Konstan, y J. Riedl, “An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms,” *Information Retrieval*, vol. 5, 2002, págs. 287-310.

[58] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, y J. Riedl, “GroupLens: applying collaborative filtering to Usenet news,” *Communications of the ACM*, 1997.

[59] B. Sarwar, G. Karypis, J. Konstan, y J. Riedl, “Analysis of recommendation algorithms for e-commerce,” *Proceedings of the 2nd ACM conference on Electronic commerce*, ACM New York, NY, USA, 2000, págs. 158-167.

[60] B.N. Miller, J.A. Konstan, y J. Riedl, “Pocketlens: Toward a personal recommender system,” *ACM Transactions on Information Systems (TOIS)*, vol. 22, 2004, págs. 437-476.

[61] R. Rafter y B. Smyth, “Passive profiling from server logs in an online recruitment environment,” *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001)*, 2001, págs. 35-41.

[62] C. Boutilier, R.S. Zemel, y B. Marlin, “Active collaborative filtering,” *In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2003.

[63] D. Maltz y K. Ehrlich, “Pointing the way: Active collaborative filtering,” *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1995, págs. 202-209.

[64] M. Gnasa, S. Alda, N. Gul, y A.B. Cremers, “Personalized Peer Filtering for a Dynamic Information Push,” *Anonymous Foundations of Intelligent Systems", Berlin, Heidelberg: Springer Verlag*, 2005, págs. 650-659.

[65] M. Richard, “5 Problems of Recommender Systems, http://www.readwriteweb.com/archives/5_problems_of_recommender_systems.php.”

[66] M.J. Pazzani y D. Billsus, “Content-based recommendation systems,” *Lecture Notes in Computer Science*, vol. 4321, 2007, pág. 325.

[67] M.F. Porter, “An algorithm for suffix stripping,” *Program: electronic library and information systems*, vol. 40, 2006, págs. 211-218.

[68] G. Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Addison-Wesley, 1989.

[69] M. Pazzani y D. Billsus, “Learning and revising user profiles: The identification of interesting web sites,” *Machine learning*, vol. 27, 1997, págs. 313-331.

[70] G. Linden, B. Smith, y J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” *IEEE Internet computing*, vol. 7, 2003, págs. 76-80.

[71] J.R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, 1986, págs. 81-106.

[72] J.W. Kim, B.H. Lee, M.J. Shaw, H.L. Chang, y M. Nelson, “Application of decision-tree induction techniques to personalized advertisements on Internet storefronts,” *International Journal of Electronic Commerce*, vol. 5, 2001, págs. 45-62.

[73] P.E. Danielsson, “Euclidean distance mapping,” *Computer Graphics and Image Processing*, vol. 14, 1980, págs. 227-248.

[74] G. Salton, A. Wong, y C.S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, 1975, págs. 613-620.

[75] D. Billsus, M.J. Pazzani, y J. Chen, “A learning agent for wireless news access,” *Proceedings of the 5th international conference on Intelligent user interfaces*, ACM New York, NY, USA, 2000, págs. 33-36.

[76] J.J. Rocchio, “Relevance feedback in information retrieval,” *The SMART retrieval system: experiments in automatic document processing*, 1971, págs. 313-323.

[77] S.B. Cousins, A. Paepcke, T. Winograd, E.A. Bier, y K. Pier, “The digital library integrated task environment (DLITE),” *Proceedings of the second ACM international conference on Digital libraries*, ACM New York, NY, USA, 1997, págs. 142-151.

[78] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, y A. Waibel, “Machine learning,” *Annual Review of Computer Science*, vol. 4, 1990, págs. 417-433.

[79] E. Pampalk, A. Rauber, y D. Merkl, “Content-based organization and visualization of music archives,” *Proceedings of the tenth ACM international conference on Multimedia*, ACM New York, NY, USA, 2002, págs. 570-579.

[80] J.T. Foote, “Content-based retrieval of music and audio,” *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, 1997.

[81] E. Rich, “User modeling via stereotypes,” *Cognitive Science*, vol. 3, 1979, págs. 329-354.

[82] D. Billsus y M.J. Pazzani, “Learning collaborative information filters,” *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

[83] C.C. Aggarwal, J.L. Wolf, K.L. Wu, y P.S. Yu, “Horting hatches an egg: A new graph-theoretic approach to collaborative filtering,” *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM New York, NY, USA, 1999, págs. 201-212.

[84] J.B. Schafer, D. Frankowski, J. Herlocker, y S. Sen, “Collaborative filtering recommender systems,” *Lecture Notes in Computer Science*, vol. 4321, 2007, pág. 291.

[85] I.A. Al Mamunur Rashid, D. Cosley, S.K. Lam, S.M. McNee, J.A. Konstan, y J. Riedl, "Getting to know you: Learning new user preferences in recommender systems," *Proceedings of the 7th international conference on Intelligent user interfaces, January*, 2002, págs. 13-16.

[86] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, y H.P. Kriegel, "Probabilistic memory-based collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, 2004, págs. 56-69.

[87] A. Popescul, L.H. Ungar, D.M. Pennock, y S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," *17th Conference on Uncertainty in Artificial Intelligence*, 2001, pág. 444.

[88] P. Bedi, H. Kaur, y S. Marwaha, "Trust based recommender system for the semantic web," *Proc. of the IJCAI07*, 2007, págs. 2677-2682.

[89] A.I. Schein, A. Popescul, L.H. Ungar, y D.M. Pennock, "Methods and metrics for cold-start recommendations," *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, 2002, págs. 253-260.

[90] P. Massa y P. Avesani, "Trust-aware collaborative filtering for recommender systems," *Lecture Notes in Computer Science*, 2004, págs. 492-508.

[91] J. Canny, "Collaborative filtering with privacy," *2002 IEEE Symposium on Security and Privacy, 2002. Proceedings*, 2002, págs. 45-57.

[92] J. O'Donovan y B. Smyth, "Trust in recommender systems," *Proceedings of the 10th international conference on Intelligent user interfaces*, ACM New York, NY, USA, 2005, págs. 167-174.

[93] R. Sinha y K. Swearingen, "Comparing recommendations made by online systems and friends," *Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, 2001.

[94] C.N. Ziegler y G. Lausen, "Analyzing correlation between trust and user similarity in online communities," *Lecture notes in computer science*, 2004, págs. 251-265.

[95] A. Felfernig y B. Gula, "An empirical study on consumer behavior in the interaction with knowledge-based recommender applications," *Proceedings of the The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services*, IEEE Computer Society Washington, DC, USA, 2006, pág. 37.

[96] R. Burke, "Integrating knowledge-based and collaborative-filtering recommender systems," *Proceedings of the Workshop on AI and Electronic Commerce*, 1999.

[97] D. Ruan, F. Hardeman, y K. Van der Meer, *Intelligent Decision and Policy Making Support Systems*, Springer, Berlin, 2008.

[98] C.N. Ziegler, “Semantic web recommender systems,” *Proceedings of the Joint ICDE/EDBT Ph. D. Workshop*, Springer, 2004.

[99] H.S. Nwana, “Software agents: An overview,” *Knowledge Engineering Review*, vol. 11, 1996, págs. 205-244.

[100] B. Laurel, “Computers as theatre,” *Reading, MA*, 1991.

[101] D. Riecken, “Intelligent agents,” 1994.

[102] J.A. Sánchez, “A taxonomy of agents,” *Rapport technique, ICT-Universidad de las Américas-Puebla, México*.

[103] P. Maes, “Agents that reduce work and information overload,” 1994.

[104] A. Cypher, “Eager: Programming repetitive tasks by example,” *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, ACM New York, NY, USA, 1991, págs. 33-39.

[105] D. Brickley y R.V. Guha, “Rdf vocabulary description language 1.0: Rdf schema, w3c recommendation 10 february 2004,” *World Wide Web Consortium*, 2004.

[106] I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, y M. Dean, “SWRL: A semantic web rule language combining OWL and RuleML,” *W3C Member Submission*, vol. 21, 2004.

[107] D.L. McGuinness, R. Fikes, J. Hendler, y L.A. Stein, “DAML+ OIL: an ontology language for the Semantic Web,” *IEEE Intelligent Systems*, vol. 17, 2002, págs. 72-80.

[108] S. Bechhofer, F. Van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, y L.A. Stein, “OWL web ontology language reference,” *W3C recommendation*, vol. 10, 2004, págs. 2006-01.

[109] M.K. Smith, C. Welty, y D.L. McGuinness, “OWL web ontology language guide. W3C Recommendation, 10 February 2004,” *World Wide Web Consortium*, 2004.

[110] G. Antoniou y F. Van Harmelen, “Web ontology language: Owl,” *Handbook on ontologies*, vol. 2, 2004, págs. 45-60.

[111] R.Q. Wang y F.S. Kong, “Semantic-enhanced personalized recommender system,” *Machine Learning and Cybernetics, 2007 International Conference on*, 2007.

[112] Y. Blanco-Fernández, J.J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, B. Barragáns-Martinéz, M. López-Nores, J. García-Duque, A. Fernández-Vilas, y R.P. Díaz-Redondo, “AVATAR: An advanced multi-agent recommender system of personalized TV contents by semantic reasoning,” *Lecture notes in computer science*, 2004, págs. 415-421.

[113] M. Richardson, R. Agrawal, y P. Domingos, “Trust management for the semantic web,” *Lecture Notes in Computer Science*, 2003, págs. 351-368.

[114] J. Golbeck, B. Parsia, y J. Hendler, “Trust networks on the semantic web,” *Lecture Notes in Computer Science*, 2003, págs. 238-249.

[115] B. Mobasher, R. Burke, R. Bhaumik, y C. Williams, “Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness,” 2007.

[116] B. Adrian, L. Sauermann, y T. Roth-Berghofer, “ConTag: A Semantic Tag Recommendation System.”

[117] J. Golbeck, “Combining provenance with trust in social networks for semantic web content filtering,” *Lecture Notes in Computer Science*, vol. 4145, 2006, pág. 101.

[118] J. Golbeck y J. Hendler, “Filmtrust: Movie recommendations using trust in web-based social networks,” *Proceedings of the IEEE Consumer Communications and Networking Conference*, 2006, págs. 43-44.

[119] D. Brickley y L. Miller, “FOAF vocabulary specification,” *Namespace Document*, vol. 3, 2005.

[120] W. Woerndl, C. Schueller, y R. Wojtech, “A Hybrid Recommender System for Context-aware Recommendations of Mobile Applications,” *2007 IEEE 23rd International Conference on Data Engineering Workshop*, 2007, págs. 871-878.

[121] Z. Yu, Y. Nakamura, S. Jang, S. Kajita, y K. Mase, “Ontology-based semantic recommendation for context-aware e-learning,” *Lecture Notes in Computer Science*, vol. 4611, 2007, pág. 898.

[122] R. Burke, “Hybrid web recommender systems,” *Lecture Notes in Computer Science*, vol. 4321, 2007, pág. 377.

[123] P. Cotter y B. Smyth, “PTV: Intelligent personalised TV guides,” *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2000, págs. 957-964.

[124] W.W. Cohen, “Fast effective rule induction,” *Proceedings of the Twelfth International Conference on Machine Learning*, Citeseer, 1995, págs. 115–123.

[125] R.J. Mooney y L. Roy, “Content-based book recommending using learning for text categorization,” *Proceedings of the fifth ACM conference on Digital libraries*, ACM New York, NY, USA, 2000, págs. 195-204.

[126] R. Burke, B. Mobasher, y R. Bhaumik, “Identifying attack models for secure recommendation,” *Beyond Personalization*, vol. 2005, 2005.

[127] P. Kollock, “The production of trust in online markets,” *Advances in group processes*, vol. 16, 1999, págs. 99-123.

[128] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” *Proceedings of the 2nd ACM Conference on Electronic Commerce*, ACM New York, NY, USA, 2000, págs. 150-157.

[129] M. O'Mahony, N. Hurley, N. Kushmerick, y G. Silvestre, “Collaborative recommendation: A robustness analysis,” *ACM Transactions on Internet Technology (TOIT)*, vol. 4, 2004, págs. 344-377.

[130] S.K. Lam y J. Riedl, “Shilling recommender systems for fun and profit,” *Proceedings of the 13th international conference on World Wide Web*, ACM New York, NY, USA, 2004, págs. 393-402.

[131] R. Burke, B. Mobasher, y R. Bhaumik, “Limited knowledge shilling attacks in collaborative filtering systems,” *Proceedings of the 3rd IJCAI Workshop in Intelligent Techniques for Personalization*, 2005.

[132] B. Mobasher, R. Burke, R. Bhaumik, y C. Williams, “Effective attack models for shilling item-based collaborative filtering systems,” *Proceedings of the 2005 WebKDD Workshop, held in conjuction with ACM SIGKDD’2005*, Citeseer, 2005.

[133] R. Burke, B. Mobasher, R. Bhaumik, y C. Williams, “Segment-based injection attacks against collaborative filtering recommender systems,” *Fifth IEEE International Conference on Data Mining*, 2005, pág. 4.

[134] “Cumbres Iberoamericanas de Jefes de Estado y de Gobierno, http://www.cumbresiberoamericanas.com.”

[135] “Center for Information Therapy, http://www.ixcenter.org/.”

Apéndice A: A qualidade em sites na área da saúde salus

O presente projeto trata de um conjunto de atividades centradas na pesquisa para a validação de uma Plataforma de Recomendação e Consulta na Web Semântica com especial ênfase em sites na área da saúde. Esta plataforma será planejada como um Portal Web com acesso adaptável aos diferentes usuários em função de seu nível cultural e peculiaridades regionais. O tema central deste projeto surgiu muito naturalmente pela evolução e convergência das diversas pesquisas desenvolvidas pelos grupos envolvidos. Estas linhas de pesquisa estão centradas em sistemas distribuídos e Web, sistemas para consultas em SGBD e arquivos XML, Web services, Sistemas de Informação na área de Saúde, e Sistemas de Recomendação. Ao longo dos quatro anos propostos os grupos envolvidos desenvolverão a cooperação aqui proposta, mas, também, desenvolverão cooperações bi ou trilaterais com o objetivo de criarem projetos específicos para o financiamento das pesquisas e, principalmente, da implantação dos sistemas especificados. Contamos com a Rede Temática CYTED para suportar a mobilidade dos pesquisadores e para apoiar alguns estágios de jovens pesquisadores em parceiros da Rede.

Este Projeto de Rede tem como principal motivação gerar um Portal Web cujo conteúdo refira-se às informações sobre Sites Web que abordem temas relacionados com a saúde, especialmente nos aspectos relacionados à sua qualidade e adequação ao público, a sua credibilidade e opiniões de usuários dos mesmos. Estes sites podem ser caracterizados como aqueles que provem informações sobre o cuidado com a saúde de pacientes, tratamento de enfermidades e prevenção para a população em geral. Esta é uma área nova que na parte dedicada ao oferecimento de informação específica a pacientes é denominada em Inglês “Information Therapy” (Ix 2006). Nossa proposta é um pouco mais ampla, pois procuramos oferecer informações não somente para pacientes, mas também, para o público geral. Não ofereceremos uma avaliação absoluta de qualidade, mas uma série de indicadores que permitam aos usuários ter uma boa percepção da adaptação dos sites às suas necessidades. Entre estes indicadores podemos citar (em Inglês): “Decision-focused and actionable,

Evidencebased, Reviewed by experts, Referenced, Up to date, Unbiased, User-fiendly" (Kemper 2006). A base conceitual para o desenvolvimento desse repositório está centrada nas áreas de Web Semântica e de Mineração de Dados na Web (Web Mining). O objetivo final extrapola a avaliação de Sites Web e tem por meta a especificação e prototipação de mecanismos automáticos e semi-automáticos de avaliação de conteúdos Web. Esperamos, ao final do projeto, contar com um demonstrador que, recebida uma URL de um Site Web em tema ligado à saúde, minere na Web dados e apresente um dashboard com uma apresentação visual clara dos diversos indicadores sobre o site ai incluídas as opiniões e recomendações dos grupos de usuários.

Entre os fatores que devem ser considerados nesta avaliação estão as diferenças de comportamento dos usuários e de percepção destes indicadores pelos mesmos, neste item está incluído o nível de conhecimento necessário para a leitura e compreensão do site: um ótimo site científico sobre um determinado assunto pode não ser adequado, e até incompreensível, para um leitor de menor nível cultural e tudo depende de variações regionais. Este fato existe, atualmente, nas bulas de medicamentos onde as precauções e interações medicamentosas são ininteligíveis para uma grande parte das populações de nossos países. O mencionado problema é de central importância para o desenvolvimento de ações de certificação de sites da área da saúde e de proteção à saúde pública e mesmo de alertas às autoridades. Esta proposta vincula-se a um tema de grande importância social nos diferentes países associados: a geração de parâmetros que auxiliem usuários na avaliação de qualidade e adequação de sites tratando de assuntos sobre a saúde a seu nível de conhecimento e necessidades. O tratamento deste problema necessita da cooperação de diferentes competências. Esta proposta de Rede Temática está incluída na linha 6.1 do Edital CYTED e visa fomentar e facilitar a comunicação entre investigadores nos domínios das Ciências da Vida, Informática e Ciências do Conhecimento com vistas ao desenvolvimento de projeto conjunto. Para tanto há grupos com experiência em Ciências do Conhecimento nos aspectos Ontológicos e de modelagem de conhecimento, Ciência da Computação com adaptabilidade e aplicações multiculturais e em Ciências da Vida com um grupo médico e grupos com experiência em aplicações na área da saúde. A seguir

descrevemos algumas das cooperações já existentes e que asseguram a viabilidade de nossa proposta.

1 Justificativa da qualidade científica da proposta

Os dois requisitos deste campo serão tratados conjuntamente. A presente proposta é inovadora dentro do princípio que está baseada em atividades que já passaram pela análise de comitês científicos para a avaliação de projetos de pesquisa tal como descritos no campo "Indicadores de producción científico-tecnológica del coordinador y de los grupos participantes". Como os projetos de pesquisa locais estão em andamento bem como as dissertações e teses associadas, podemos certificar a atualidade dos componentes da proposta. Complementarmente a presente proposta foi formulada para agregar os conhecimentos e tecnologias em desenvolvimento em torno de um tema aglutinador e de importância para os Países aqui representados e para toda a região, o oferecimento de indicadores de qualidade e de adaptação de sites ligados à saúde destinado a usuários leigos.

Em cada um dos tópicos os temas são de atualidade. Na área de Informação na Saúde estamos tratando com uma área muito nova parte dedicada ao oferecimento de informação específica a pacientes (denominada em Inglês "Information Therapy") e ampliando este conceito para sugestões preventivas para a população em geral. Na Web Semântica estamos trabalhando com Ontologias de domínio, com alinhamento de ontologias e com inferência para recomendação semântica. Na área de Mineração na Web (Web mining) torna-se um desafio na medida em que páginas Web são por natureza semi-estruturados ou totalmente desestruturadas e os sites Web apresentam uma diversidade muito grande de design (i.e., projeto de conteúdo, navegação e layout). Associado a essas questões, ainda é necessário gerenciar adequadamente o extenso volume e a cobertura de conteúdos e usuários. A modelagem de Perfis de Usuários e sua clusterização em grupos significativos de comportamentos e de seqüências de acesso a conteúdos é, também, um problema não totalmente resolvido. Finalmente há o problema de especificação do ambiente de software baseado em serviços para permitir a implementação dos serviços de recomendação e qualificação de conteúdos digitais de forma distribuída.

2 Justificativa da inserção social da proposta

Há uma enorme proliferação de sites Web tratando de assuntos ligados à saúde. Esses sites são uma fonte inestimável de informação para a população. Entretanto, entre sites altamente confiáveis de grandes Universidades de pesquisa encontram-se sites com informações extremamente perigosas e sem o menor fundamento científico. Esse problema é de central importância para o desenvolvimento de ações de certificação de sites da área da saúde e de proteção à saúde pública e mesmo de alertas às autoridades. Esta proposta vincula-se a um tema de grande importância social nos diferentes países associados: a geração de parâmetros que auxiliem usuários na avaliação de qualidade de sites tratando de assuntos sobre a saúde.

Há uma enorme influência destes sites na vida cotidiana do público. Para tratar desse desafio estamos propondo a criação de uma rede cooperativa unindo as competências de vários centros de pesquisa na elaboração de modelos e protótipos de mecanismos de qualificação de sites que permitam a geração de indicadores de fácil compreensão pelo público geral sobre o grau de confiabilidade desses sites.

Entre os fatores que devem ser considerados estão as diferenças de comportamento dos usuários e de percepção desses indicadores pelos mesmos, nesse item está incluído o nível de conhecimento necessário para a leitura e compreensão do site: um ótimo site científico sobre um determinado assunto pode não ser adequado, e até incompreensível, para um leitor de menor nível cultural. Esse fato existe, atualmente, nas bulas de medicamentos onde as precauções e interações medicamentosas são ininteligíveis para uma grande parte das populações de nossos países.

3 Proposta

É preciso, para avaliar a qualidade de sites Web, desenvolver mecanismos que garantam uma forma pública e de qualidade garantida para a avaliação aberta de conteúdos. Um sistema de indexação e avaliação de conteúdos digitais tem por objetivo auxiliar no processo social de criar conhecimento, aperfeiçoar este conhecimento através da revisão pelos pares e indicar ou receber indicação de conhecimento relevante.

Sistemas de avaliação e recomendação de produtos são largamente usados em comércio e marketing, tanto para sugerir produtos ou fornecer informações como para ajudar o cliente a decidir o que comprar. Esses sistemas se baseiam, em grande parte, na avaliação realizada pelos usuários para selecionar os produtos ou serviços oferecidos; essa mesma metodologia pode ser aplicada para a avaliação de sites Web. Neste caso o problema é mais complexo do que na recomendação de produtos comerciais, onde todos os consumidores são considerados iguais, pois é preciso associar à avaliação um indicador de qualidade do revisor. Nota-se claramente uma métrica bidimensional: a qualidade do conteúdo publicado e a qualidade dos avaliadores.

De forma complementar à avaliação humana, e talvez mais importante por sua independência de interesses específicos, é relevante considerar mecanismos e estratégias para a obtenção de meta-dados que permitam uma inferência gerando indicadores de qualidade associados a um determinado site. Entre estes meta-dados podem ser citados: a competência científica do autor obtida através da avaliação de suas publicações associada à qualidade dos meios na qual foi realizada; a qualidade da instituição que suporta o site, como uma Universidade de Pesquisa ou um alerta sobre conflitos de interesse como um laboratório publicando a avaliação de qualidade de um medicamento produzido pelo mesmo; ou o número de referências e a qualidade das referências a esse site.

Em um modelo de avaliação de qualidade de sites é tratado o problema de validação de conteúdos baseado no perfil e nas competências do autor, das instituições que o publicam e das referências ao mesmo. A modelagem do perfil do usuário serve para aprimorar o processo de qualificação e permitir a avaliação indireta de conteúdos. Para que um sistema possa avaliar de forma automática a qualidade dos itens considerando o autor, a qualificação dos revisores e o nível de conhecimento necessário para a leitura do item é necessário o acesso a fontes de informação adequadas.

O objetivo desta proposta de cooperação é especificar uma arquitetura, um modelo conceitual e de realizar protótipos que permitam uma avaliação experimental para

permitir que um sistema parcialmente automatizado avalie a qualidade de um site (ou de partes de um site). Esta avaliação é baseada na qualificação do autor, e na qualidade dos pareceres de usuários de uma forma similar ao modelo de page ranking, utilizado pelo Google, onde referências de maior qualidade são ponderadas mais fortemente do que referências de menor peso. Dentro do espírito deste edital, nossa proposta está centrada em consolidar o grupo internacional de pesquisa onde o núcleo central de parceiros já mantém ativa interação com a agregação de novos grupos. Os pesquisadores membros desta proposta têm desenvolvido atividades especificamente na área deste projeto, sendo que, assim, a reunião das competências existentes permitirá, certamente, a obtenção de resultados adequados. Há uma grande possibilidade de realizarmos mais do que o especificado no projeto, com a agregação de alunos realizando trabalhos de diplomação, de mestrado ou incluindo parcelas de seus trabalhos de doutorado na linha de pesquisa aqui proposta. Essa agregação permitirá o desenvolvimento de protótipos para a validação dos conceitos e modelos desenvolvidos no projeto. Dentro das reuniões de coordenação e de trabalho sempre serão reservados espaços para a apresentação de mini-cursos ou seminários sobre o tema para divulgar e atrair alunos e pesquisadores dos grupos participantes a desenvolverem trabalhos na área.

4 Áreas de pesquisa relacionadas

Este Projeto inclui o estudo e a aplicação em áreas de pesquisa atualmente muito ativas, quais sejam: Web Semântica; Mineração na Web *(Web mining)* e Sistemas de Recomendação, sempre voltados para o tema de divulgação de tópicos sobre a saúde na Web.

4.1 Web semântica

Ultimamente, a Web Semântica tem se tornado uma área de pesquisa comum em aplicações Web. Conforme analisam Decker et al. (2000), a primeira geração Web iniciou com o desenvolvimento manual de páginas HTML. A segunda geração concentrou-se no desenvolvimento automático de páginas HTML, muitas das quais ativas e dinâmicas. Tanto a primeira quanta a segunda geração Web caracterizaram-se por serem dirigidas à compreensão e ao processamento por seres humanos (i.e.,

leitura, navegação e preenchimento de formulários). Entretanto, a terceira geração Web tem sido referenciada como *Web Semântica,* o que coincide com a visão de Tim Berners-Lee em *"Weaving the Web"* (Berners-Lee, 1999):

> *"A Web Semântica corresponde à geração Web na qual os seus recursos são acessíveis não somente por humanos, mas também para automatizar processos, i.e., agentes automáticos navegando na Web e executando tarefas úteis tais como buscar informação de forma eficiente e precisa, descobrir recursos, filtrar informações etc."*

Para realizar a visão de Berners-Lee, Bechhofer et al. (2001) apresentam e discutem três requisites técnicos essenciais: (i) a descrição do conteúdo e das funcionalidades dos recursos Web através de *metadados;* (ii) uma *linguagem padrão* para a descrição destes conteúdos e funcionalidades *(Resource Description Framework* - RDF); e (iii) uma *ontologia,* com o objetivo de prover a representação semântica dos conceitos compartilhados de um domínio particular, e com o objetivo de permitir que estes conceitos sejam comunicados entre aplicações e pessoas. Os parceiros do centra Médico serão essenciais para o fornecimento de informações e validação dos resultados.

4.2 Mineração na Web

Considerando que o nosso interesse neste Projeto é o de desenvolver um serviço Web para disponibilizar informações sobre Sites da área da Saúde, um importante desafio consiste em extrair da Web, de forma automática ou semi-automática, as informações pertinentes. Conforme discutem Lim e Sun (2005), procurar, organizar e manter uma biblioteca digital com conteúdos capturados na Web torna-se um desafio na medida em que páginas Web são por natureza semi ou totalmente desestruturadas e os *sites* Web apresentam uma diversidade muito grande de *design* (i.e., projeto de conteúdo, navegação e *layout).* Associado a essas questões, ainda é necessário gerenciar adequadamente o extenso volume e a cobertura de conteúdos e usuários.

Com o objetivo de encontrar soluções que facilitem a busca e a indexação de conteúdos Web, pesquisadores em bibliotecas digitais estão investindo nas técnicas de mineração de dados na Web (KOSALA; BLOCKEEL, 2000). As técnicas mais usuais de

mineração de dados na Web incluem a *classificação* e a *extração* de dados. Especificamente neste Projeto, enfocamos as técnicas de *extração* de dados na Web, o que pode envolver a busca por elementos HTML e frases ou *tuplas* que representem a instância de algum conceito requerido, i.e., nomes de pessoas, nomes de locais etc. (LIM; SUN, 2005).

Conforme apresentam esses autores, *ontologia* e *mineração de dados na Web* são áreas que podem ser aplicadas conjuntamente de três formas:

- Quando a *ontologia* e suas *instâncias* são conhecidas. Nesse caso, as instâncias são identificadas entre os dados Web de entrada, possibilitando uma descoberta de conhecimento semanticamente mais significativa para o usuário.
- Quando somente a *ontologia* está disponível como estrutura semântica de entrada de dados. Nesse caso, a estrutura semântica da ontologia é usada como base para a mineração de dados e os dados extraídos da Web funcionarão como instâncias de *conceitos,* e essas instâncias funcionarão como instâncias de *relacionamentos,* o que também gera um conhecimento semanticamente mais significativo para o usuário.
- Quando a *ontologia* está disponível, mas não atende diretamente aos requisites do usuário. Esta situação decorre do caráter extremamente volátil da informação, especialmente em temas relacionados à saúde. Novos tratamentos e novas tecnologias surgem muito rapidamente, de modo que uma ontologia em pouco tempo se torna obsoleta, devido à ausência da definição dos novos conceitos. Por exemplo, há 10 anos atrás o uso de células tronco para regeneração de tecidos ainda não era discutido. Uma forma de evitar que a ontologia se torne desatualizada envolve a sua manutenção gradual, incorporando novos conceitos de forma manual à medida que estes se tornem conhecidos.

A combinação entre ontologia (semântica) e mineração de dados na Web gera vantagens importantes para as aplicações Web, tais como: (i) a indexação de dados Web como *conceitos* e seus *relacionamentos* provê suporte para consultas semanticamente expressivas, com resultados mais precisos e com uma provável

redução de informações Web irrelevantes como resultado; (ii) ao invés de simplesmente navegar na Web à procura de informação através do uso de *links,* páginas Web podem ser visitadas com base nos conceitos e relacionamentos de sua ontologia; (iii) da vantagem (ii), decorre uma significativa facilitação para os sistemas de recomendação.

4.3 Sistemas de recomendação

A web semântica permite a descrição mais detalhada e aprofundada do conteúdo de cada elemento disponibilizado na web (i.e. páginas, web-services etc.), e a mineração na web oferece uma panacéia de ferramentas e métodos de processamento e análise das informações contidas na web (semântica ou não) e também da própria interação do usuário com esses elementos, permitindo a identificação de padrões e de informações ou conhecimento ocultos (implícitos) nos mesmos.

De posse das informações e conhecimentos identificados, é extremamente importante poder recomendar o que foi descoberto aos usuários que utilizam a web. Isso porque a quantidade de informações disponíveis na web é gigantesca e cresce a todo o momenta (NELSON 1994). Os usuários não conseguem identificar a informação e o conhecimento relevante dentre o que lhe é fornecido, mesmo com as ferramentas existentes, ficando perdidos - *lost in cyberspace* (BRAKE 1997) ou sobrecarregados - *information overload* (FARHOOMAND and DRURY 2002; CARLSON 2003).

Os sistemas de recomendação (RESNICK and VARIAN 1997) são desenvolvidos justamente para minimizar esses problemas, filtrando as informações inúteis ou irrelevantes e apresentando ao usuário aquilo que lhe é mais útil e relevante em determinado momento, tomando como base o seu perfil de interesse.

Existem basicamente dois tipos de sistemas de recomendação. Os primeiros são baseados em filtragem colaborativa (GOLDBERG et al. 1992), ou seja, as recomendações são geradas tendo como base as avaliações feitas por usuários que possuem perfil similar, onde aqueles itens que não são muito bem avaliados são ignorados pelo sistema (constituindo o filtro). A identificação da relevância de cada item é estimada com base nas opiniões dos outros usuários e é construída de forma colaborativa, dando origem ao nome da técnica. Seu problema consiste em necessitar

de opiniões dos usuários para realizar as recomendações, algo extremamente difícil de ocorrer durante as primeiras interações com o sistema, onde não há dados (opiniões) anteriores. Fato semelhante ocorre com itens recentes, ainda não opinados. O segundo tipo é aquele que faz a recomendação baseada no conteúdo do itens, sem levar em conta as recomendações dos usuários em si, sendo denominada, justamente, de recomendação baseada em conteúdo (MOONEY et al. 1998).

Abordagens híbridas (BALABANOVIC and SHOHAM 1997) podem ser utilizadas para combinar os métodos existentes nas duas abordagens, minimizando os seus problemas. As técnicas de recomendação podem ser beneficiadas pelas tecnologias de descrição de conteúdo provenientes da web semântica, recomendando informações mais pertinentes ao usuário, e também pelas técnicas de mineração na web, que permitem a identificação mais correta e adequada do tipo de usuário que consome e que produz informação na web.

5 Objetivos do Projeto

Para a avaliação do Projeto SALUS dividimos os objetivos em: objetivos dos grupos, objetivo geral e objetivos específicos. Os objetivos específicos são tratados pelas Atividades descritas mais adiante.

5.1 Objetivos dos grupos

- Especificar uma arquitetura e um modelo conceitual de um ambiente de avaliação aberta de sites ligados à saúde na Web;
- Criar, manter e consolidar o trabalho cooperativo entre as equipes envolvidas;
- Definir trabalhos a serem desenvolvidos por alunos dos grupos;
- Disponibilizar amplamente o acesso aos conhecimentos, artigos e eventuais protótipos desenvolvidos para suportar novas atividades de pesquisa;
- Divulgar os resultados obtidos por workshops, seminários, artigos e relatórios de pesquisa estimulando a pesquisa na área de avaliação de qualidade;
- Gerar start-ups de alunos e pesquisadores que possam transformar os conhecimentos desenvolvidos em atividades auto-sustentáveis.
- Estimular a submissão de subprojetos por grupos de parceiros para assegurar e consolidar as atividades de pesquisa locais ou regionais.

5.2 Objetivo geral do projeto

O objetivo deste Projeto consiste em desenvolver um modelo conceitual baseado em Ontologia, projetar e desenvolver um Portal Web a partir da mineração de dados na Web realizada sobre sites que tratem de temas relacionados com a saúde. Esse portal destina-se a avaliar a qualidade relativa e a autoridade (instituições suportando, pesquisadores reconhecidos etc.) destes conteúdos digitais. Este objetivo será atingido agregando as diferentes competências dos grupos através dos intercâmbios e eventos realizados.

5.3 Objetivos específicos

- Especificação do modelo de qualidade de sites sobre temas de saúde. Isso inclui a definição dos conceitos relacionados a esse domínio, da estrutura hierárquica de classes e subclasses, bem como definição das regras para classificar um site sobre temas de saúde.
- Especificação dos mecanismos de coleta de dados na Web, com vistas à automatização do processo de geração das instâncias da ontologia.
- Implementação* e avaliação dos mecanismos de coleta de dados especificados.
- Desenvolvimento* e avaliação da ontologia com base no modelo de qualidade especificado e nos mecanismos de coleta de dados implementados.
- Projeto, implementação* e testes do algoritmo de inferência para classificar alguns sites sobre temas de saúde previamente analisados e avaliados.
- Projeto, implementação e testes do algoritmo de recomendação automática de conteúdo relacionado aos sites relevantes identificados pelo algoritmo de inferência.
- Projeto, especificação e implantação, por bolsista contratado, de um Portal Web apresente os resultados do projeto para ampla difusão na sociedade.
- Divulgação dos resultados obtidos por workshops, palestras, publicação de artigos e de duas monografias gerais sobre o tema.
- Formação de jovens pesquisadores por meio de estágios em parceiros do projeto.

*em função da opção de alunos

6 Metodologia

A forma de trabalho dos grupos Universitários está intimamente ligada à formação de recursos humanos e ao desenvolvimento de projetos que suportem total ou parcialmente os estudantes. Os diversos grupos envolvidos possuem projetos locais ou de cooperação parcial que suportam estas atividades. Dentro desta realidade nossa proposta é de atribuir a cada uma das grandes atividades dois grupos responsáveis sendo que os demais participam como colaboradores. Esta metodologia permitirá, ao longo do projeto, a integração natural dos grupos e a busca de objetivos de pesquisa locais complementares e cooperativos. Para atingir os objetivos propostos, o Projeto será desenvolvido em diversas etapas:

1. Estudo de trabalhos na literatura relacionados com os temas de Web semântica e ontologias.
2. Estudo de trabalhos na literatura relacionados com o tema de Mineração de Dados na Web.
3. Estudo de trabalhos na literatura relacionados com o tema de Sistemas de Recomendação.
4. Estudo de trabalhos na literatura relacionados com o tema qualidade de sites.
5. Especificação do modelo de qualidade de sites sobre temas de saúde.
6. Especificação dos mecanismos de coleta de dados na Web.
7. Implementação da ontologia, com base no modelo de qualidade especificado.
8. Testes (manuais) da ontologia a partir da geração de instâncias utilizando, para isto, sites sobre temas de saúde conhecidos.
9. Implementação e testes dos mecanismos de coleta de dados na Web.
10. Projeto, implementação* e testes do algoritmo de inferência para classificar automaticamente ou semi-automaticamente sites sobre temas de saúde.
11. Projeto, implementação* e teste do algoritmo de recomendação automática de conteúdo relacionado aos temas da Saúde.
12. Projeto, implementação* e testes de um Portal Web.
13. Divulgação dos resultados obtidos.

* em função da opção de alunos

6.1 Organização e gerenciamento

Para suporte e divulgação será criado um site público do projeto sendo os relatórios e demais publicações liberadas sob licença *Creative Commons.* Este site servirá como aglutinador dos trabalhos e como referencial para o alinhamento e potencializarão contínua da cooperação. Este site será mantido por um software de gestão de sites como, por exemplo, o Limbo ou equivalente. Esta estrutura de coordenação foi utilizada na implantação do Projeto ProTeM-CC do CNPq e avaliada pelo coordenador geral deste projeto (Oliveira, J. Palazzo M. de and Hoppen 94) tendo sido a primeira aplicação em grande escala de emprego de meios eletrônicos na coordenação de projetos científicos na América Latina. Em paralelo serão utilizadas ferramentas de coordenação e cooperação eletrônicas para uma contínua interação e cooperação, atualmente estamos considerando a utilização do Groove. Periodicamente serão realizadas reuniões de projeto durante as quais ocorrerão *workshops* para divulgar os resultados entre os membros locais e alunos dos grupos envolvidos. Os coordenadores farão o possível para incluir em eventos científicos, como o Simpósios Brasileiros promovidos pela SBC e outros eventos regionais ou Ibero-latinoamericanos, *workshops* sobre os temas de pesquisa da Rede Temática.

Como estrutura de gerenciamento e coordenação serão realizadas quatro reuniões gerais, duas na Península Ibérica e duas na América Latina (os locais estão aqui indicados como sugestão inicial). A primeira, no Brasil, servirá para a organização inicial das atividades. A segunda, na Espanha, será a reunião de acompanhamento intermediário destinada a avaliar o andamento e planejar as atividades de redação da primeira monografia bem como para revisão e ajustes no projeto. A terceira, a ser realizada no Uruguai, servirá para a avaliação final do andamento do projeto para a esquematização das atividades de avaliação experimental. Finalmente, a quarta reunião será realizada em Portugal e destina-se a avaliação e conclusão do projeto e planejamento das próximas atividades do consórcio bem como para planejar a monografia final. Em alternância serão realizadas reuniões regionais para a coordenação e revisão do andamento das atividades.

Como resultado, teremos, ao final, realizado a integração dos pesquisadores, alinhado as linhas de pesquisa, divulgado os resultados e organizado a submissão de projetos de

pesquisa conjuntos. Este modelo foi aplicado com sucesso em vários projetos de pesquisa como o ProTem-CC e os projetos ProSul (CNPq) e a Red Educa (FRIDA).

6.1.1 O coordenador

O Coordenador do projeto, Prof. José Palazzo M. de Oliveira, é pesquisador Ic do CNPq e tem desenvolvido inúmeras atividades de coordenação de projetos e participação em comitês de avaliação de Editais de Pesquisa. Além de suas atividades em Computação participou da implantação do Programa de Doutorado em administração na UFRGS e cooperou com aquele programa por vários anos. Entre algumas das coordenações mais recentes de projetos podem ser citadas a Coordenação Sul do ProTeM-CC - fase I (implantação), pesquisador principal e, posteriormente, Coordenação do Projeto Sidi (ProTeM-CC fase III). Coordenação dos projetos Tapejara, ProTeM National fase II, ProSul 2003 - AdaptWeb Multicultural, DIGITEX CTInfo 2005; ProSul 2005 - Editoração, Indexação e Busca de Documentos Científicos em um Processo de Avaliação Aberta; PRONEX -Recomendação Semântica; PERXML - Representação e Consultas sobre a Evolução de Perfís de Usuários Codificados em XML . Atualmente é membro do Comitê assessor em Ciência da Computação do CNPq e do Comitê de Matemática, Estatística e Computação da Fundação de Amparo à Pesquisa do estado do Rio Grande do Sul, Brasil.

6.2 Atividades

Nesta seção são descritas as atividades a serem desenvolvidas para a obtenção dos resultados esperados. A lista de objetivos apresentada sugere uma seqüência de passos para a satisfação das atividades propostas. As atividades estão organizadas em sete grupos principais, que formam os macro-objetivos do projeto. Assim, a solução do problema será construída por meio de soluções individuais e independentes para cada uma destas etapas, conforme descrito nas subseções apresentadas a seguir. Antes disto, os seguintes itens descrevem uma seqüência de passos a ser utilizada no desenvolvimento do projeto, em cada uma das suas atividades:

- *levantamento bibliográfico* sobre o assunto investigado, como forma de verificar os trabalhos desenvolvidos ou em desenvolvimento por outros grupos de pesquisa. O levantamento bibliográfico permitirá também estabelecer

parâmetros de comparação entre as pesquisas desenvolvidas neste projeto pelos diferentes grupos envolvidos. Tais parâmetros de comparação são de grande valia no detalhamento dos trabalhos a serem desenvolvidos no contexto do projeto;

- *proposta de soluções* através da especificação de soluções baseadas no levantamento bibliográfico realizado e no problema a ser resolvido;
- *divulgação e validação das propostas através de artigos submetidos a congressos (workshops)* que tenham uma característica investigativa nos trabalhos aceitos. O objetivo neste item é formalizar a proposta através de um texto e validar inicialmente, na medida do possível, a mesma através de sua apresentação à comunidade de pesquisadores;
- *implementação das soluções propostas* através da criação de processos de software capazes de fornecer subsídios para a análise de resultados. Apesar de não ser esperado pelo edital a obtenção de um resultado tangível, pela estrutura operacional das Universidades, poderemos oferecer este resultado ao obtermos êxito na motivação de nossos alunos para o desenvolvimento de trabalhos de diplomação, de mestrado e de doutorado de nossos alunos;
- *divulgação dos resultados obtidos* na pesquisa realizada por *workshops* realizados quando da realização dos encontros de pesquisa e, também, através da confecção de artigos. Nesta etapa os artigos serão submetidos a congressos mais exigentes em relação aos resultados e avaliações experimentais nos trabalhos aceitos, Finalmente serão submetidos artigos a periódicos.

6.3 Descrição das atividades

O projeto está centrado em sete *grandes atividades:* Definição da ontologia do domínio; Mineração de dados na Web, Indexação dos *sites;* Desenvolvimento do modelo de qualidade; Modelagem do perfil do usuário; Recomendação dos *sites;* e Validação e avaliação experimental. As atividades contarão com a participação dos grupos de pesquisa tal como descrito na seção 7 abaixo. Cada grupo tem desenvolvido pesquisas nas áreas de sua responsabilidade e constituem grupos de pesquisa com maturidade acadêmica, desta forma a atribuição interna de responsabilidades é uma decisão do coordenador local e não está detalhada nesta proposta. Cada Grande

Atividade será desenvolvida sob a coordenação de dois grupos com a participação dos demais e servirão de aglutinadoras das pesquisas estimulando a cooperação e a troca de experiências.

6.3.1 Definição da Ontologia do domínio

Esta atividade centra-se no elemento conceitual do projeto: o desenvolvimento de uma ontologia abrangente para a representação dos diversos elementos associados à qualidade relativa e a autoridade (instituições suportando, pesquisadores reconhecidos etc.) de documentos digitais da área da saúde. Aqui são considerados não apenas a qualidade científica intrínseca a um determinado site, como também, a sua adequação a um determinado público de leitores. O objetivo principal consiste em prover uma representação semântica dos conceitos compartilhados de um domínio particular e também permitir que esses conceitos sejam comunicados entre aplicações e pessoas. Nesse contexto, os parceiros do centra Médico serão essenciais para o fornecimento de informações e validação dos resultados.

Um segundo problema a ser tratado envolve a situação em que a ontologia está disponível, mas não atende diretamente aos requisitos do usuário. Esta situação decorre do caráter extremamente volátil da informação, especialmente em temas relacionados à saúde. Novos tratamentos e novas tecnologias surgem muito rapidamente, de modo que uma ontologia em pouco tempo se torna obsolete devido à ausência da definição dos novos conceitos. Por exemplo, há 10 anos atrás o uso de células tronco para regeneração de tecidos ainda não era discutido. Uma forma de evitar que a ontologia se torne desatualizada envolve a sua manutenção gradual, incorporando novos conceitos de forma manual à medida que estes se tornem conhecidos. No entanto, com o intuito de tomar esta atualização mais automática, com base em técnicas de mineração (ver a seguir), propõe-se uma solução mais dinâmica, conforme listado abaixo:

- Em um primeira momento, permitir que o usuário realize consultas semânticas mesmo que os conceitos requisitados não existam na ontologia. Isto envolveria uma etapa de mineração de dados *on-the-fly,* para suprir as necessidades do usuário.

- Em um segundo momento, propõe-se que a ontologia seja retro-alimentada através das consultas dos usuários sobre conceitos não existentes. Esta proposta caracteriza uma solução do tipo *late-binding,* onde os conceitos se tornam disponíveis apenas após um usuário requisitar por eles. Isto agilizaria a atualização da ontologia, sendo apenas necessária a intervenção de um especialista para decidir a melhor forma de encaixar o novo conceito na ontologia. Ademais, após a primeira requisição, o mapeamento entre os novos conceitos e suas instancias já se torna disponível, de modo que as consultas futuras sobre estes termos se tomem mais eficientes.

6.3.2 Mineração na web

Uma vez definida a ontologia é possível utilizá-la como base para a identificação e extração de informações em sites e repositórios específicos de dados na web. Um dos primeiros passos consiste em identificar sites relacionados com o tema da saúde que contenham informações úteis e relevantes. Tais sites podem ser indicados pelos usuários ou podem ser localizados por uma ferramenta de análise automática de sitios (crawler ou spider), encarregado de vasculhar a web ou regiões específicas dela a fim de identificar possíveis páginas relacionadas com o assunto desejado, tendo como base a ontologia. A tecnologia que permite realizar tal identificação é a classificação ou categorização, que realiza um casamento (matching) do conteúdo das informações das páginas analisadas com os conceitos presentes na ontologia. As páginas identificadas como relevantes podem ser resguardadas para posterior indexação e recomendação, ou ser utilizadas como fonte de extração de dados ou informações específicas, tais como nomes de especialistas de uma área, enfermidades, sintomas, tratamentos etc.

Para a extração de informações é necessário estudar e compreender a estrutura ou o padrão com que as informações são apresentadas. Atualmente são conhecidas algumas heurísticas que permitem a identificação desses elementos, tendo como base a estrutura sintática dos elementos (disposição no texto), o formato (i.e. datas cujos elementos são separados por caracteres '/', como em 10/10/06), a identificação de palavras-chave específicas (i.e. identificação de medicamentos, tal como em '10 mg de Xilocaína') ou até mesmo o uso de dicionários (i.e. de nomes próprios e entidades). Uma vez definidas as regras de extração e as páginas-alvo (fonte dos dados), é possível

desenvolver parsers para analisá-las, extrair seus dados e popular bases de dados ou a própria ontologia.

6.3.3 Indexação dos sites

O foco desta atividade é tratar a obtenção automática, semi-automática ou manual de meta-dados dos documentos digitais, de seu armazenamento em formato XML compatível com *Dublin Core* bem como de sua disponibilização como um Portal Web. Além disso, para um domínio de aplicação (no caso presente para a área de saúde) será desenvolvida uma Ontologia de Domínio.

6.3.4 Modelagem do perfil do usuário

O foco desta atividade é a análise de dados associados a um usuário tais como CV Lattes, documentos previamente recuperados, características de navegação entre outros, é definido um modelo do usuário. Este modelo de usuário é utilizado no processo de recuperação e recomendação de documento, em conjunto com a ontologia de domínio.

O objetivo desta tarefa é a modelagem computacional e validação do perfil do usuário em um sistema de recomendação para *sites* Web ligados à saúde, baseada em padrões de comportamento navegacional e em um padrão de meta-dados para a descrição do usuário.

Especificamente, esta atividade envolve as seguintes tarefas:

- estudos de trabalhos na literatura relacionados ao uso das técnicas de mineração de uso na Web e ao uso de padrões para a descrição de documentos na Web, assim como estudos relacionados a padronizações referentes à descrição de dados na Web sobre o usuário;
- definição de uma técnica para minerar comportamento navegacional do usuário;
- definição do modelo de perfil do usuário;
- definição de uma arquitetura computacional do modelo de perfil do usuário;
- desenvolvimento e validação de um protótipo que implemente a arquitetura computacional do modelo de perfil do usuário proposta;

- produção de documentação sobre todo o trabalho produzido e posterior publicação das mesmas em eventos e periódicos a nível nacional e internacional, com o objetivo de obter uma avaliação da comunidade científica e reconhecimento da pesquisa realizada.

6.3.5 Recomendação

A partir de uma lista de documentos classificados por suas características de similaridade com o tópico utilizado pela busca de similaridade e baseado no perfil do usuário, o foco desta atividade é realizar um processo de recomendação que classifica os documentos em função da correlação entre as características dos documentos e aquelas dos usuários.

O objetivo principal de um sistema de recomendação é ajudar pessoas a encontrarem informações de que necessitam de maneira pró-ativa, ou seja, sem que as pessoas precisem realizar consultas ou mesmo explicitar do que precisam. Através de processos colaborativos, que permitem a pessoas indicarem ou receberem indicais, tais sistemas auxiliam nos processos de tomada de decisão. Por serem pró-ativos, detectam as necessidades dos usuários sem que estes precisem formalizá-las, auxiliando as pessoas em ambientes com diferentes e volumosas opções. Para realizar recomendações, deve-se coletar dados sobre o usuário, de forma implícita (analisando por exemplo as ações do usuário no sistema) ou explícita (quando o usuário informa o que deseja receber ou suas preferências ou fornece dados sobre si).

A proposta desta atividade é pesquisar as técnicas de recomendação que façam distinção dos tipos de usuários. O objetivo é utilizar somente informações vindas de certos tipos de usuários (por exemplo, os mais ativos ou mais conhecedores do assunto em questão). Também se faz necessário distinguir o tipo do documento, se básico ou avançado, para evitar recomendar algo fora do nível de conhecimento do usuário.

O fato de as técnicas tradicionais considerarem as pessoas sem distinguir experiências passadas pode ser eficiente para filmes ou programas de televisão, mas não quando se trata de documentos eletrônicos ou artigos científicos presentes em uma biblioteca digital. Um doutor em uma área especifica não pode receber as mesmas

recomendações que um aluno iniciante nesta área. Da mesma forma, um item básico não pode recomendado para um especialista, nem um item avançado para um novato.

Uma forma de responder essas questões é identificando o nível ou grau de conhecimento de cada pessoa em áreas ou temas específicos, ou seja, provendo a identificação da expertise de cada usuário. Inicialmente propõe-se a distinção de cinco tipos de usuários, conforme seus níveis de expertise: Usuários Experientes (aqueles com mais tempo de atividade no sistema); Usuários Ativos (que estão regularmente interagindo com o sistema, desempenhando diversas ações); Usuários hubs ou surveys (aqueles que contribuem ativamente com leitura e armazenamento de documentos para a biblioteca digital e que, portanto, podem ser referência para encontrar outros usuários ou documentos); Especialistas (cujo currículo confirma sua especialidade e seu alto grau de conhecimento numa área) e Autoridades (que são apontados por outros usuários como referências em determinadas áreas).

Esta atividade também deverá investigar os diferentes tipos existentes, tanto de usuários quanto de documentos, procurando avaliar se o tipo interfere e como interfere no processo de recomendação. Pretende-se, portanto, gerar recomendações mais adequadas às necessidades dos usuários.

6.3.6 Validação e avaliação experimental

Para cada modelo, ferramenta e método proposto serão desenvolvidos modelos específicos de validação, relacionados com a técnica própria da área, utilizada e recomendada pela comunidade específica, identificada na literatura relacionada.

Apesar de não ser esperado pelo edital a obtenção de um resultado tangível, é importante avaliar e validar os objetos propostos nas atividades do projeto, através de experimentos específicos. Neste momento, para alguns casos, determinados experimentos incluindo suas etapas e métricas já podem ser predefinidos e vislumbrados. Os demais, devido a complexidade do projeto e ao fato de algumas definições e encontros iniciais serem necessários, podem ser propostos no future em momentos mais oportunos. Cabe salientar que muitos experimentos serão igualmente dependentes dos trabalhos a serem desenvolvidos pelos alunos (dissertações) e bolsistas a serem alocados no projeto.

Para a avaliação e validação da ontologia, por exemplo, podem ser realizados testes (manuais) relacionados com a geração de instâncias, utilizando para isso sites sobre temas de saúde conhecidos (a serem definidos).

Para os mecanismos de coleta, é necessária a elaboração de um conjunto de dados (sites) de teste, onde os elementos a serem extraídos são definidos *a priori* e os métodos desenvolvidos devem gerar como saída os mesmos dados. Nesse caso, métricas tradicionais de *recall* e *precision* (BAEZA-YATES and RIBEIRO-NETO 1999), e suas variações (LEWIS 1991), podem ser utilizadas.

Nos mecanismos de inferência, classificação e recomendação um processo similar de avaliação pode ser utilizado, onde um conjunto de elementos predefinidos é encaminhado aos mecanismos de análise e o resultado comparado com o esperado. As métricas são as mesmas ou similares.

Como validação e avaliação geral, o próprio uso do portal a ser implantado, pela comunidade acadêmica e pelo público em geral, onde serão disponibilizados os documentos e as ferramentas desenvolvidos, serve como mecanismo, tendo o respaudo de sua audiência. Finalmente, os artigos a serem publicados, demonstrados e discutidos em conferências também constituem um mecanismo de validação e avaliação dos produtos desenvolvidos e modelados.

7 Resultados esperados

*Nesta seção os resultados, descritos em outras partes deste documento, são agrupados para uma melhor avaliação global.

Há uma enorme proliferação de sites Web tratando de assuntos ligados à saúde. Estes sites são uma fonte inestimável de informação para a população. Entretanto entre sites altamente confiáveis de grandes Universidades de pesquisa encontram-se sites com informações extremamente perigosas e sem o menor fundamento científico. Há uma enorme influência destes sites na vida cotidiana do público. A rede cooperativa unindo as competências de vários centros de pesquisa desenvolverá modelos e protótipos de mecanismos de qualificação de sites que permitam a geração de

indicadores de fácil compreensão pelo público geral sobre o grau de confiabilidade e adequação destes sites para um determinado público.

O resultado esperado é a integração destas competências agregando facetas complementares e gerando um portal de apoio aos pesquisadores da área e testando protótipos que permitam aos usuários obterem informações relevantes para a área de saúde. Teremos, ao final, realizado a integração dos pesquisadores, alinhado as linhas de pesquisa, divulgado os resultados e organizado a submissão de projetos de pesquisa conjuntos.

Entre os fatores de análise dos sites que devem ser considerados estão as diferenças de comportamento dos usuários e de percepção destes indicadores pelos mesmos, neste item está incluído o nível de conhecimento necessário para a leitura e compreensão do site: um ótimo site científico sobre um determinado assunto pode não ser adequado, e até incompreensível, para um leitor de menor nível cultural. Este fato existe, atualmente, nas bulas de medicamentos onde as precauções e interações medicamentosas são ininteligíveis para uma grande parte das populações de nossos países. Esta é uma área nova que na parte dedicada ao oferecimento de informação específica a pacientes é denominada em Inglês “Information Therapy”. Nossa proposta é mais ampla pois procuramos oferecer informações não somente para pacientes mas, também, para o público geral. Não ofereceremos uma avaliação absoluta de qualidade, mas uma série de indicadores que permitam aos usuários ter uma boa percepção da adaptação dos sites às suas necessidades. Entre estes indicadores podemos citar (em Inglês): “Decisionfocused and actionable, Evidencebased, Reviewed by experts, Referenced, Up to date, Unbiased, User-fiendly”.

Para suporte e divulgação será criado um site público do projeto sendo os relatórios e demais publicações liberadas sob licença Creative Commons. Este site servirá como aglutinador dos trabalhos e como referencial para o alinhamento e potencializarão contínua da cooperação. Este site será mantido por um software de gestão de sites. Em paralelo serão utilizadas ferramentas de coordenação e cooperação eletrônicas para uma contínua interação e cooperação.

Periodicamente serão realizadas reuniões de projeto durante as quais ocorrerão workshops para divulgar os resultados entre os membros locais e alunos dos grupos envolvidos. Os coordenadores farão o possível para incluir em eventos científicos, como o Simpósios Brasileiros promovidos pela SBC e outros eventos regionais ou Ibero-latinoamericanos, workshops sobre os temas de pesquisa da Rede Temática.

Como estrutura de gerenciamento e coordenação serão realizadas quatro reuniões gerais, duas na Península Ibérica e duas na América Latina (os locais estão aqui indicados como sugestão inicial). A primeira, no Brasil, servirá para a organização inicial das atividades. A segunda, na Espanha, será a reunião de acompanhamento intermediário destinada a avaliar o andamento e planejar as atividades de redação da primeira monografia bem para revisão e ajustes no projeto. A terceira, a ser realizada no Uruguai, servirá para a avaliação final do andamento do projeto para a esquematização das atividades de avaliação experimental. Finalmente, a quarta reunião será realizada em Por tugal e destina-se a avaliação e conclusão do projeto e planejamento das próximas atividades do consórcio bem como para planejar a monografia final. Em alternância serão realizadas reuniões regionais para a coordenação e revisão do andamento das atividades.

Como os grupos participantes possuem programas de pós-graduação será gerada, em decorrência das dissertações de mestrado e teses de doutorado geradas por esta rede, uma quantidade significativa de artigos científicos. Além disto serão editados pela rede duas monografias, em forma de livros, uma ao final de cada biênio, condensando os resultados e apresentado todo um referencial conceitual para a disseminação dos conhecimentos produzidos.

8 Riscos e dificuldades

O presente projeto trata de um conjunto de atividades centradas na pesquisa conceitual e no desenvolvimento de protótipos para a validação de ambiente de Avaliação da qualidade de Sites na área da saúde. Esta abordagem será baseada em ontologia e mineração de dados na Web. O tema central deste projeto surgiu naturalmente pela evolução e convergência das diversas pesquisas desenvolvidas pelos pesquisadores envolvidos. Estas linhas de pesquisa estão centradas em sistemas

para consultas em servidores Web, modelagem conceitual da área e em sistemas de recomendação. Nos últimos anos, os parceiros do grupo vêm atuando intensamente na área de sistemas na Web, tendo todos seus participantes trabalhado e cooperado em facetas específicas no tema. Alguns em avaliação de qualidade (UdelaR, UFRGS), Ontologias (UFRGS, UPC, UdelaR), Sistemas de Recomendação (UFRGS), Sistemas de saúde (UNICAUCA, UdelaR), entre outras. Os detalhes são encontrados nos sites dos grupos de pesquisa. Desta forma fica evidente que o projeto é um todo consistente e desenvolvido de forma natural por uma complementação de competências e consolidado por projetos ao longo de projetos cooperativos. O risco institucional e de coordenação do projeto são extremamente reduzidos tendo em vista a experiência dos pesquisadores principais e considerando a maturidade científica atingida pelo grupo.

9 Resolução de conflitos

Quaisquer conflitos que eventualmente surjam devem ser tratados da seguinte forma:

- O interessado deve comunicar o problema ao Coordenador de seu grupo para uma tentativa de resolução;
- Caso uma solução não seja encontrada localmente será realizada uma discussão por e-mail ou outra forma de coordenação eletrônica a ser definida entre as partes interessadas, o Coordenador Local e o Coordenador Geral do projeto;
- Se, ainda assim, não for encontrada uma solução negociada o Coordenador Geral do consórcio documentara a situação e, por meio de e-mail ou outra forma de coordenação eletrônica a ser definida, consultará todos os Coordenadores Locais. A decisão será tomada por maioria simples dos votos dos Coordenadores e será definitiva.
- A não aceitação desta decisão pelo interessado implicara no seu afastamento do projeto.

10 Bibliografia

BAEZA-YATES, R. and RIBEIRO-NETO, B. (1999). Modern Information Retrieval. Adisson-Wesley.

BALABANOVIC, M.; and SHOHAM, Y (1997). Fab: content-based, collaborative recommendation. Communications of the ACM. ACM, 40(3), p.66–72.

BECHHOFER, S.; HORROCKS, I.; GOBLE, C.; STEVENS, R. (2001). OilEd: a Reason-able Ontology Editor for the Semantic Web. IN: Lecture Notes in Computer Science, Vol. 2174, 396-?.

BERNERS-LEE, T. (1999). Weaving the web. Harper, San Francisco.

BONHARD, P., HARRIES, C., MCCARTHY, J., and SASSE, M. A. (2006). Accounting for taste: using profile similarity to improve recommender systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). CHI '06. ACM Press, New York, NY, 1057-1066.

BRAKE, D. (1997). Lost in cyberspace. New Scientist Magazine, 2088 (28 jun. 1997). http://www.newscientist.com/article/mg15420882.200.html.

CAPES. (2005). http://qualis.capes.gov.br/pesquisa/ServletPesquisa.

CAPES. (2006). http://www.capes.gov.br.

CARLSON, Christopher N. (2003) Information overload, retrieval strategies and Internet user empowerment. In Haddon, Leslie, Eds. Proceedings The Good, the Bad and the Irrelevant (COST 269) 1(1), pp. 169-173, Helsinki (Finland).

CHAKRABARTI, S, et al. (1999). Mining the link structure of the world wide web. IEEE Computer, 32(8):60-67.

COOLEY, R.; MOBASHER, B. and SRIVASTAVA, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

DECKER, S. et al. (2000). The semantic Web: the roles of XML and RDF. IEEE Internet Computing, 4(5), 63-74.

DESPEYROUX, T. (2004). Practical semantic analysis of web sites and documents. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 685-693.

DONG, J. S., LEE, C. H., LEE, H. B., LI, Y. F., and WANG, H. (2004). A combined approach to checking web ontologies. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 714-722.

DORNELES, Carina F., Heuser, CARLOS A., LIMA, Andrei E. N., SILVA, Altigran Soares da, MOURA, Edleno Silva de. (2004). Measuring similarity between collection of values. WIDM 2004: 56-63.

DÖRRE, J., GERSTL, P., and SEIFFERT, R. (1999). Text mining: finding nuggets in mountains of textual data. In Proceedings of the Fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Diego, California, United States, August 15 - 18, 1999). KDD '99. ACM Press, New York, NY, 398-401

ETZIONI, O. (1996). The World Wide Web: Quagmire or gold mine. COMMUNICATIONS OF THE ACM, 39(11):65-68.

FANG, X. and SHENG, O. R. (2004). LinkSelector: A Web mining approach to hyperlink selection for Web portals. ACM Trans. Inter. Tech. 4, 2 (May. 2004), 209-237.

FARHOOMAND, Ali F. and DRURY, Don H. (2002). Managerial information overload. Communications of the ACM. ACM 45(10) (Oct. 2002), 127-131.

GOLDBERG, D. Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry (1992). Communications of the ACM. ACM 35(12). (Dec.1992), 61—70.

GRAU, B. C. (2004). A possible simplification of the semantic web architecture. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 704-713.

GROBELNIK, M., MLADENIC, D., and MILIC-FRAYLING, N. (2000). Text mining as integration of several related research areas: report on KDD's workshop on text mining 2000. SIGKDD Explor. Newsl. 2, 2 (Dec. 2000), 99-102.

HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., and RIEDL, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 1 (Jan. 2004), 5-53.

HULL, R. and SU, J. (2005). Tools for composite web services: a short overview. SIGMOD Rec. 34, 2 (Jun. 2005), 86-95.

Ix (2006), http://www.ixcenter.org/. Acessada em 29-06-2006.

KEMPER, D. (2006). Information therapy: The strategic role of prescribed information in disease self-management. In: L. Bos, L. Roa, K. Yogesan, B. O'Connell, A. Marsh and B. Blobel (Edrs.): Medical and Care Compunetics 3, Series of Studies in Health Technology and Informatics. V.121, 2006.

KOSALA, R.; BLOCKEEL, H. (2000). Web Mining Research: A Survey. SIGKDD Explorations, 2(1), 1–15.

KROEZE, J. H., MATTHEE, M. C., and BOTHMA, T. J. (2003). Differentiating data- and text-mining terminology. In Proceedings of the 2003 Annual Research Conference of the South African institute of Computer Scientists and information Technologists on Enablement Through Technology (September 17 - 19, 2003). J. Eloff, A. Engelbrecht, P. Kotzé, and M. Eloff, Eds. ACM International Conference Proceeding Series, vol. 47. South African Institute for Computer Scientists and Information Technologists, 93-101.

LEWIS, David Dolan. (1991). Representantion and Learning in Information Retrieval. Amherst: University of Massachusetts, Department of Computer and Information Science. Phd Thesis.

LIM, Ee-Peng; SUN, Aixin. (2005). Web Mining - The Ontology Approach. International Advanced Digital Library Conference (IADLC), Nagoya, Japan, Aug. 2005. (invited paper). (Disponí-vel em: http://iadlc.nul.nagoya-u.ac.jp/archives/IADLC2005/Ee-Peng.pdf).

LIU, B., CHIN, C. W., and NG, H. T. (2003). Mining topic-specific concepts and definitions on the web. In Proceedings of the 12th international Conference on World

Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM Press, New York, NY, 251-260.

MEDJAHED, B., BOUGUETTAYA, A., and Elmagarmid, A. (2003). Composing Web services on the Semantic Web. The VLDB Journal 12, 4 (Nov. 2003), 333-351.

MEI, Q. and ZHAI, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceeding of the Eleventh ACM SIGKDD international Conference on Knowledge Discovery in Data Mining (Chicago, Illinois, USA, August 21 - 24, 2005). KDD '05. ACM Press, New York, NY, 198-207.

MELLO, Ronaldo dos Santos, HEUSER, Carlos A. (2005). BInXS: A Process for Integration of XML Schemata. CAiSE 2005: 151-166.

MIDDLETON, S. E., SHADBOLT, N. R., and DE ROURE, D. C. (2004). Ontological user profiling in recommender systems. ACM Trans. Inf. Syst. 22, 1 (Jan. 2004), 54-88.

MIKA, P., Oberle, D., GANGEMI, A., and SABOU, M. (2004). Foundations for service ontologies: aligning OWL-S to dolce. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 563-572.

MOONEY, R.; BENNETT, P. and ROY, L. (1998). Book recommending using text categorization with extracted information. Proceedings of AAAI-98/ICML-98 Workshop on Learning for Text Categorization and the AAAI-98 Workshop on Recommender Systems,pp.49-54 and pp.70-74, Madison, WI, July 1998.

NANDIGAM, J., GUDIVADA, V. N., and Kalavala, M. (2005). Semantic Web services. J. Com-put. Small Coll. 21, 1 (Oct. 2005), 50-63.

NELSON, M. R. (1994). We have the information you want, but getting it will cost you!: held hostage by information overload. Crossroads 1, 1 (Sep. 1994), 11-15.

OLIVEIRA, J. PALAZZO M. DE AND HOPPEN N. (1994). E-mail as an enabling technology in Brazil: the CNPq/ProTeM-CC experience, in Proccedings of the 13th IFIP Congress '94, v. 2, p. 366-371. E. Raubold and K. Brunnstein ed., Hamburg, Aug. 28 - Sep. 2, 1994, Elsevier Science

PROTÉGÉ. (2005). http://protege.stanford.edu/.

RASHID, A. M., ALBERT, I., COSLEY, D., LAM, S. K., MCNEE, S. M., KONSTAN, J. A., and RIEDL, J. (2002). Getting to know you: learning new user preferences in recommender systems. In Proceedings of the 7th international Conference on intelligent User interfaces (San Francisco, California, USA, January 13 - 16, 2002). IUI '02. ACM Press, New York, NY, 127-134.

RESNICK, P. AND VARIAN, H. R. (1997). Recommender systems. Communications of the ACM. ACM 40(3), 56–58.

SCHRAEFEL, m. c, SHADBOLT, N. R, GIBBINS, N., HARRIS, S., and GLASER, H. (2004). CS AKTive space: representing computer science in the semantic web. In *Proceedings of the 13th international Conference on World Wide Web* (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 384-392.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., and TAN, P. (2000). Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl.* 1, 2 (Jan. 2000), 12-23.

TAI, S., KHALAF, R, and MIKALSEN, T. (2004). Composition of coordinated web services. In *Proceedings of the 5th ACM/IFIP/USENLX international Conference on Middleware* (Toronto, Canada, October 18 - 22, 2004). Middleware Conference, vol. 78. Springer-Verlag New York, New York, NY, 294-310.