ARTíCULO

Sistema de Búsqueda Personalizada y Recomendación de Documentación Científica

Erika J. Salazar G., Oscar Ortega L.

Universidad de Antioquia
Medellin, Colombia
esalazar@psl.com.co, oortega@udea.edu.co

Resumen

La sobrecarga de información ha sido un problema ampliamente tratado entre la comunidad científica de las áreas de recuperación y filtrado de información. Un investigador que se encuentre buscando a través de la Web se enfrenta a dicho problema cuando se encuentra reuniendo información y artículos para la generación de un estado del arte en sus temas específicos de investigación. Las fuentes de información electrónica especializada a consultar son diversas y los documentos obtenidos a partir de ellas son tan numerosos que deben ser examinados uno a uno por los investigadores con el fin de filtrar aquellos que representan la información más relevante y actualizada. Como solución al problema, han surgido los llamados sistemas de recomendación y filtrado de información, los cuales, aunque aplicados con mayor frecuencia en sitios comerciales de ventas en línea, se han planteado como una posibilidad de apoyo a los usuarios en sus búsquedas de información, ayudando en la localización y filtrado automático de documentos interesantes. Sin embargo, sistemas como estos son poco comunes, más allá de aplicaciones experimentales o de comercio electrónico, y son poco conocidos por la comunidad de usuarios en general. En el presente artículo se presenta el desarrollo de un sistema de búsqueda y recomendación automática de documentos, dirigido hacia los usuarios investigadores de una comunidad académica con intereses de información documental especializada. El sistema cuenta con varios módulos. Un módulo de generación de consultas, encargado de extraer y transformar en consultas los términos más importantes contenidos en los perfiles de cada usuario; un módulo de búsqueda y descarga de documentos, encargado de enviar las consultas a un conjunto de buscadores de documentos científicos en la Web y luego descargarlos; un módulo de agrupamiento, encargado de procesar y almacenar los documentos obtenidos a partir de las búsquedas; y un módulo de filtrado, recomendación y retroalimentación, encargado de filtrar los subconjuntos de documentos relevantes para ser recomendados a los usuarios y de ajustar los perfiles de dichos usuarios a partir de los valores de calificación que ellos suministran, ya sea implícita o explícitamente, a los documentos que les son recomendados. Las recomendaciones producidas por el sistema desarrollado fueron evaluadas según el cambio en la calidad de las mismas a lo largo del tiempo para un conjunto de usuarios. Dicha calidad se midió usando el área bajo la curva ROC, la cual debía aumentar a lo largo del tiempo en que es usado el sistema, indicando un aprendizaje y mejora en los resultados de recomendación presentados a los usuarios. Aunque durante la evaluación se obtuvo un buen desempeño y el área bajo la curva ROC demostró un aumento en la calidad de los resultados de recomendación a lo largo del tiempo, dicho aumento fue mucho mayor al comienzo de los experimentos que al final de los mismos. Por lo tanto, para establecer las causas de tales variaciones, se plantearon nuevas hipótesis estableciendo la importancia que tiene la frecuencia de generación de recomendaciones en el desempeño del sistema y la necesidad de realizar experimentación mucho más extensa y detallada.

Palabras clave: Sistemas de recomendación, Filtrado de información, Personalización, Generación de consultas, Agrupamiento

1. Introducción

El auge de la Internet y de la producción científica ha causado un aumento sorprendente en el número de publicaciones electrónicas. Un artículo que está disponible a través de un medio de publicación electrónico tiene grandes posibilidades de ser conocido en todo el mundo de manera rápida y sencilla, convirtiéndose la Internet en la primera opción que un usuario considera para la búsqueda de artículos; reemplazando así, en muchas ocasiones, otros medios más tradicionales como las bibliotecas o las revistas científicas publicadas en papel [10]. Sin embargo, frente a tanta abundancia de información, es difícil encontrar los artículos que se buscan. Un investigador, por ejemplo, invierte bastante tiempo y esfuerzo en la búsqueda de documentos científicos relacionados con el tema de sus investigaciones para estar informado acerca del estado del arte en dichos temas. Debe consultar en diversos buscadores, tanto generales como específicos, bibliotecas digitales y otras fuentes de información, revisando y seleccionando - en la gran cantidad de documentos obtenidos como resultado de las búsquedas -, los artículos más relevantes. Considerando el problema que representa tal sobrecarga de información, dos subcampos de las ciencias de la información se han enfocado en estudiar y proponer posibles soluciones. Estos subcampos son: la Recuperación de Información (IR por sus siglas en inglés) y el Filtrado de Información (IF), siendo este último de origen más reciente. Los sistemas de recuperación de información tratan de satisfacer necesidades de información de corto plazo, indicando la disponibilidad o no disponibilidad de datos relevantes a consultas particulares [17]. Los motores de búsqueda existentes hoy día, como Google [1], son un ejemplo de sistemas de Recuperación de Información. Los sistemas de filtrado de información tratan de satisfacer necesidades de información de largo plazo, representándolas como perfiles y filtrando flujos de datos para retornar sólo aquellos datos que coinciden con uno o más de los perfiles [17]. Los sistemas de recomendación y personalización, usados comúnmente en sitios comerciales como Amazon[2] o en algunas bibliotecas digitales, son ejemplos de sistemas que usan técnicas de filtrado de información. Los sistemas de filtrado de información son de especial interés en el presente artículo; por ello se presenta en lo que resta de esta introducción.

[1] http://www.google.com
[2] http://www.amazon.com

Un sistema de filtrado de información está conformado principalmente de un conjunto de perfiles de usuario, un conjunto de ítems o datos a evaluar, un algoritmo de filtrado y un algoritmo de aprendizaje o retroalimentación. Los perfiles de usuario representan cada uno las necesidades de información del usuario al que pertenecen. El conjunto de ítems a evaluar constituye la fuente de datos a filtrar y cada ítem de dicho conjunto también puede ser representado por algún tipo de perfil, cuando se posee información específica acerca de su contenido. El algoritmo o función de filtrado se encarga de filtrar y seleccionar, del conjunto de ítems, los datos relevantes que van a ser recomendados a los usuarios, generando, para cada ítem y usuario, una predicción de relevancia. El algoritmo de aprendizaje o retroalimentación se encarga de modificar los perfiles de los usuarios con base en valores de calificación que dichos usuarios proveen acerca de la relevancia de los ítems que reciben como recomendados.

Un algoritmo o función de filtrado de información evalúa si un ítem es relevante o no para un usuario, ya sea examinando el contenido de un ítem para compararlo con el perfil del usuario o tomando consejos de otros usuarios que tienen perfiles similares al usuario. De acuerdo con estas diferentes estrategias, existen varios tipos de filtrado de información: Filtrado basado en contenido, Filtrado colaborativo y Filtrado híbrido. El filtrado basado en contenido analiza el contenido de un ítem o documento y predice su relevancia con base en el perfil del usuario [19]; por ello, este filtro es usado cuando se tiene acceso a la descripción o contenido de los ítems a recomendar, como es el caso de páginas Web, documentos electrónicos y noticias. Cuando el usuario recibe las recomendaciones de ítems y las califica, ya sea de forma explícita o implícita, el perfil del usuario es adaptado para que en las siguientes recomendaciones sean favorecidos los ítems que son parecidos a aquellos que recibieron calificaciones positivas por parte del usuario. Una gran ventaja de los filtros basados en contenido es que pueden recomendar ítems a un usuario aún cuando este no tenga preferencias en común con los demás usuarios del sistema y por lo tanto no encaje en ningún grupo particular de usuarios. Sin embargo, su gran desventaja es la sobre-especialización, pues el filtro sólo recomienda al usuario ítems parecidos a otros ítems que le fueron recomendados anteriormente [18].

El filtrado colaborativo, acumula una base de datos de calificaciones de ítems dadas por un conjunto de usuarios y luego, para predecir las preferencias que puede tener un usuario sobre los ítems que no ha visto aún, usa las calificaciones que han dado otros usuarios con preferencias parecidas a dicho usuario [19]. Los filtros colaborativos, a diferencia de los sistemas basados en contenido, son independientes del contenido de los ítems que recomiendan y por tanto pueden recomendar ítems que son diferentes entre si. Sin embargo, los filtros colaborativos también tienen desventajas. Por ejemplo, al tener que considerar los usuarios con gustos similares a otros para generar las recomendaciones de estos últimos, aquellos usuarios que no tengan mucho en común con otros usuarios recibirán recomendaciones de poca calidad [18]. Otra de sus desventajas es que, si un usuario es nuevo en el sistema, no se posee suficiente información de sus preferencias y por lo tanto no puede compararse con ningún otro para generar recomendaciones [18]. El filtrado híbrido, combina técnicas de filtrado colaborativo y filtrado basado en contenido, compensando así las desventajas de cada uno y aprovechando sus fortalezas [19]. La combinación de ambas técnicas de filtrado es realizada ya sea calculando el promedio ponderado de las predicciones generadas por un filtro colaborativo y un filtro basado en contenido, insertando características colaborativas en un filtro basado en contenido o características de contenido en un filtro colaborativo.

Aunque en Colombia no son conocidos ampliamente sistemas de recomendación de documentos científicos como el presentado en este artículo, a nivel internacional se conocen sistemas como el de CiteSeer [4], una biblioteca digital de documentos de carácter científico. Esta biblioteca posee un sistema de recomendación, el cual usa técnicas de filtrado basado en contenido y en las cuales el perfil del usuario puede ser actualizado explícitamente por el mismo usuario o automáticamente de acuerdo con las acciones realizadas por el usuario sobre el documento como: ver detalles o descargar el documento. Sin embargo, aunque existe un artículo donde se hace referencia al desarrollo del sistema de recomendación de CiteSeer [4], este sistema no está disponible en su sitio Web[3]. El sistema de recomendación ACE [13], es otro ejemplo que, aunque no es aplicado para recomendación de documentos científicos sino para artículos de un periódico electrónico de la Agencia Cultural Electrónica (ACE) de Barcelona, combina las predicciones basadas en contenido y colaborativas generadas para cada documento en un solo valor híbrido de predicción. Al recibir las recomendaciones, el usuario puede llevar a cabo acciones de retroalimentación explícitas e implícitas.

En este artículo se presenta el desarrollo de un sistema de recomendación de documentos orientado hacia usuarios interesados en temas de investigación especializada. Dicho sistema obtiene los documentos automáticamente desde buscadores especializados en artículos científicos y reportes técnicos, generando para ello consultas a partir de los perfiles de los usuarios. Para la generación de las recomendaciones, el sistema usa una técnica de filtrado híbrido, la cual combina las predicciones de relevancia producidas por un filtro basado en contenido y un filtro colaborativo con el fin de obtener un valor de predicción global para cada documento a recomendar al usuario. Para actualizar los perfiles de los usuarios, el sistema usa una técnica de retroalimentación implícita y explícita. A través de la retroalimentación implícita, el sistema interpreta acciones del usuario, como la descarga o eliminación de un documento, como calificaciones favorables y desfavorables respectivamente hacia el documento, por parte del usuario. Y, a través de la retroalimentación explícita, el usuario puede dar un valor de calificación, entre cero y cinco inclusive, a cada documento que le fue recomendado indicando su relevancia. Así, con base en los valores de calificación obtenidos, los perfiles de los usuarios son modificados y el sistema aprende constantemente acerca de los intereses de los usuarios.

El resto de este artículo está organizado como sigue: En la segunda sección se presenta una idea general del funcionamiento del sistema, abordado por subsecciones los detalles de cada una de sus partes funcionales y en la última subsección se presenta la medida usada para determinar la exactitud del sistema de recomendación desarrollado. En la tercera sección se presenta la evaluación del sistema de recomendación, las hipótesis a probar y la discusión de los resultados obtenidos. Finalmente, en la cuarta sección se extraen las conclusiones más importantes del desarrollo del proyecto y se formulan posibles trabajos futuros.

[3] http://citeseer.ist.psu.edu

2. Enfoque Abordado

Con el fin de apoyar las labores investigativas de grupos de usuarios con intereses de largo plazo en temas especializados, se ha desarrollado un sistema de recomendación de documentos científicos, a través del cual los usuarios registrados pueden disminuir en gran medida el tiempo que invierten en la recolección de artículos y documentos relacionados con sus temas de interés científico. El sistema desarrollado apoya sus recomendaciones en un filtro híbrido de información, el cual combina, para el cálculo de las predicciones de relevancia, los valores de predicción generados por un filtro basado en contenido y un filtro colaborativo aprovechando las ventajas de ambos tipos de filtrado. Cuando un usuario se registra en el sistema, ingresa un conjunto de palabras clave, que describen sus intereses de información o temas de interés. El usuario puede ingresar hasta cuatro temas de interés descritos por palabras clave en idioma inglés, único lenguaje soportado hasta ahora. Luego de que el usuario ingresa sus temas de interés, el sistema crea un perfil de usuario por cada uno de tales temas de interés, con base en la descripción ingresada por dicho usuario en el momento de su registro. Cada perfil consiste en un vector de términos, ponderados por un peso asociado que indica su importancia en el perfil o tema de interés. Al comienzo, después de su registro, todos los términos en el perfil del usuario poseen el mismo peso; luego, con cada acción de retroalimentación que dicho usuario genere sobre los documentos que le son recomendados, los pesos de los términos en el perfil son ajustados y más términos pueden ser adicionados. Usando los términos con mayor peso de los perfiles creados y de los perfiles que han sido ajustados por retroalimentación, el sistema genera, de manera periódica, consultas para ser enviadas a un conjunto de buscadores, obteniendo así nuevos documentos para ser analizados y filtrados. Los documentos encontrados por medio de las búsquedas realizadas, son descargados, procesados y agrupados para extraer la información que conformará los perfiles de documentos, con la cual el filtro basado en contenido generará predicciones. Cada perfil de documento está constituido, al igual que los perfiles de los usuarios, por un vector de términos donde cada término está acompañado, a su vez, por un valor que representa la frecuencia con que aparece dicho término en el documento. Con los documentos ya procesados, las predicciones de relevancia, tanto del filtro basado en contenido como del filtro colaborativo, son generadas de manera periódica para cada usuario y para cada documento que aún no le ha sido recomendado. La predicción final de relevancia es generada por el filtro híbrido a partir de las predicciones basadas en contenido y colaborativas, asignando para cada usuario y filtro un peso que es ajustado con cada acción de retroalimentación del usuario. Finalmente, con base en las predicciones híbridas generadas, el sistema elige los documentos a recomendar para cada usuario, tomando aquellos documentos que tienen una mayor predicción de relevancia. Las recomendaciones generadas con base en dichos documentos son creadas fuera de línea y quedan listas para cuando el usuario ingrese de nuevo en el sistema para examinarlas.

Una vez presentado en forma global el funcionamiento del sistema, a continuación se detalla en subsecciones las técnicas usadas en cada parte funcional que lo conforma: la generación de consultas, búsqueda y descarga de documentos, el filtrado de información basado en contenido, colaborativo e híbrido (Ver Figura 1) y finalmente, el procesamiento de las acciones de retroalimentación del usuario (Ver Figura 2). Adicionalmente, en la última subsección, se introduce la métrica utilizada para medir la calidad de las recomendaciones producidas por el sistema desarrollado.

2.1. Generación de consultas, búsqueda y descarga de documentos

El flujo dinámico de documentos, desde el cual el sistema desarrollado filtra aquellos documentos que son de interés para cada usuario, proviene de buscadores especializados en documentos científicos y reportes técnicos. Es decir, el sistema desarrollado obtiene los documentos a procesar y analizar desde un conjunto de motores de búsqueda especializados. Actualmente el sistema hace uso de tres buscadores: el servidor de reportes técnicos de la NASA [4], el portal de investigación de IBM [5] y el portal de e-prints de Arxiv[6].

[4] http://ntrs.nasa.gov
[5] http://www.research.ibm.com/journal
[6] http://arxiv.org/corr/home

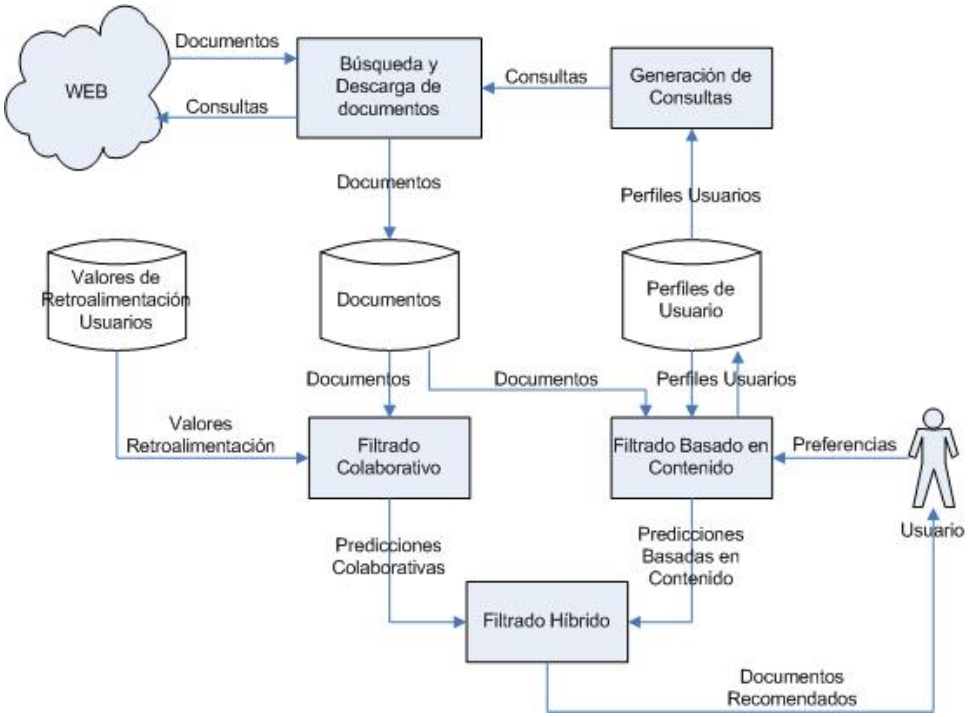

Figura 1. Obtención y recomendación de documentos

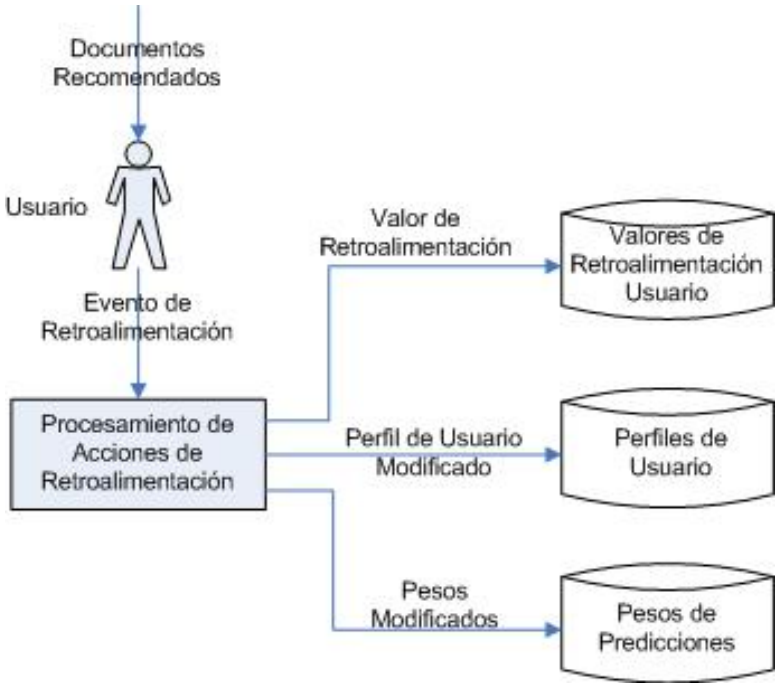

Figura 2. Ajuste de perfiles y datos a partir de la retroalimentación del usuario.

Tales buscadores son accedidos de forma paralela por el sistema, actuando como un metabuscador. Un metabuscador es una herramienta que automática y simultáneamente consulta diversos motores de búsqueda, interpreta los resultados y los presenta en un formato uniforme [6]. Sin embargo, en el sistema de recomendación desarrollado, los resultados de las búsquedas no son presentados al usuario sino que, aquellos resultados que corresponden con documentos en formato PDF, son descargados para su posterior procesamiento y filtrado. Para la realización de las búsquedas y descargas de documentos, el sistema revisa periódicamente los perfiles de usuario que han sido modificados recientemente y, para cada uno de ellos, genera una consulta constituida por un conjunto de términos, la cual es luego ejecutada en los motores de búsqueda anteriormente citados. Para seleccionar el método de generación de consultas a partir de perfiles de usuario, se examinó un estudio de varios métodos realizado por Somlo y otros [16]. En los resultados de dicho estudio se destacan principalmente tres de los ocho métodos explorados: Boley, TFIDF y TF. El método de Boley selecciona la intersección de los k primeros términos en dos listas: una ordenada en modo decreciente de acuerdo con las frecuencias de los términos en los documentos relevantes y la otra, ordenada en modo creciente de acuerdo con el número de documentos en los que se encuentran los términos. De esta manera, los k primeros términos son aquellos que aparecen muchas veces en los documentos relevantes pero muy pocas veces en toda la colección. El método TFIDF selecciona los n términos con el peso TF-IDF más alto en el perfil del usuario. Donde TF-IDF [12] (Term Frequency - Inverse Document Frequency) es una medida de frecuencia ampliamente usada en el campo de la Recuperación de Información para representar los documentos y los perfiles de usuario como vectores de frecuencias o pesos de términos. El método TF selecciona los n primeros términos del perfil de usuario ordenados de acuerdo con la frecuencia TF de los términos. Dónde TF (Term Frequency) es el factor de frecuencia de términos que hace parte de la medida de frecuencia TF-IDF.

La medida TF-IDF [12] asigna un peso $w_{i,k}$ al término T_k en el documento D_i. Para el cálculo de tal medida se tienen en cuenta dos factores TF e IDF así:

$$w_{i,k} = TF * IDF = f_{i,k} * log\left(\frac{N}{n_k}\right)$$

donde $TF = fi, k$ es la frecuencia con la que ocurre el término T_k en el documento D_i, N es el número total de documentos en la colección y n_k es el número de documentos en los cuales el término T_k ocurre al menos una vez.

En el sistema de recomendación desarrollado, los perfiles de usuario y los documentos son representados como vectores de frecuencias de términos TF, los cuales son reducidos a sus raíces usando procesos de stemming y eliminación de stopwords. No se calcula el factor IDF porque se trata de un sistema de flujo dinámico de documentos y el cálculo del factor IDF implicaría un reprocesamiento de las frecuencias de todos los términos en el sistema para todos los documentos de la colección cada vez que ingresan nuevos documentos. Por lo tanto, por facilidades de implementación y por el buen desempeño que presenta según lo demostrado en [16], se seleccionó el método TF para la generación de consultas en el sistema de recomendación desarrollado con base en los perfiles de usuario.

2.2. Filtrado Basado en Contenido

En la elección de una técnica de filtrado basado en contenido para el sistema de recomendación desarrollado, se estudiaron cuatro alternativas. La primera alternativa es la presentada en [7], donde se aplica la técnica de recuperación de información Latent Semantic Indexing (LSI) al campo de filtrado de información. LSI primero crea una estructura semántica multidimensional de la información contenida en los documentos. Para ello, realiza la descomposición en sus valores singulares (SVD) de la matriz de asociación de términos por artículos. Así, se produce una matriz de dimensionalidad reducida en la que están los K mejores factores ortogonales que aproximan la matriz original como un modelo de espacio semántico para la colección de documentos. Luego, para determinar si un documento debe ser recomendado o no a un usuario, dicho documento es usado como si fuera una consulta, llevándolo al mismo espacio semántico en el que están representados los demás documentos, y si éste aparece cerca de otros documentos que han sido interesantes con anterioridad para el usuario, el documento es considerado como posiblemente relevante para el usuario. Si por el contrario, el documento aparece cerca de otros artículos que no son relevantes para el usuario, dicho artículo será considerado como

no relevante para el usuario. La segunda alternativa es el sistema Syskill & Webert, presentado en [9]. Syskill & Webert consiste en un agente de software que aprende perfiles de usuario y los usa para identificar páginas web interesantes. A partir de una página índice construida manualmente, página con miles de enlaces a otros sitios del mismo tema, el sistema aprende un perfil con base en las calificaciones del usuario y usa el perfil para sugerir al usuario enlaces a seguir desde otras páginas. Las predicciones que indican la probabilidad de que a un usuario le interese una página, corresponden con números entre 0 y 1 que son determinados con un clasificador Bayesiano simple. Las páginas son representadas con vectores booleanos de términos. La tercera alternativa es el sistema WAIR (Web Agents for Information Retrieval), el cual es presentado en [21]. Dicho sistema consiste en tres agentes: agente de interfaz, agente de recuperación y agente de filtrado. El agente de interfaz interactúa con el usuario y observa su comportamiento para obtener información de retroalimentación implícita. El agente de recuperación crea consultas a partir del perfil del usuario y recupera hasta N documentos desde diversos motores de búsqueda. El agente de filtrado evalúa la relevancia de los documentos obtenidos comparándolos con los perfiles del usuario, los ordena según dicha relevancia y presenta al usuario los primeros M documentos de la lista original de N documentos ($M <= N$). Para realizar el filtrado de información, los documentos y los perfiles de usuario son representados como vectores de términos con sus frecuencias. Los documentos son comparados con los perfiles de los usuarios usando la métrica del coseno y la información de retroalimentación de los usuarios es obtenida de manera tanto explícita como implícita para actualizar los perfiles de dichos usuarios. Entre los documentos recomendados se incluyen además, con cierta probabilidad, algunos de baja relevancia para el usuario con el fin de explorar regiones desconocidas y de disminuir en algo la sobre-especialización que es característica de los sistemas de filtrado basados en contenido. Finalmente, la cuarta alternativa es la presentada en [2], en la cual se hace uso de la técnica de agrupamiento Star Clusters [1] para conformar grupos con los documentos que los usuarios ya han examinado y, con base en dichos grupos, además del perfil del usuario, realizar el filtrado de nuevos documentos. Las predicciones de relevancia para los documentos se calculan comparando primero dichos documentos con el perfil del usuario y luego comparándolos con otros documentos que pertenecen a su mismo grupo. Para las comparaciones se usa la medida de similaridad del coseno y es posible usar tanto retroalimentación explícita como implícita, pues el funcionamiento de esta técnica no impone ninguna condición al respecto.

Para el desarrollo del filtro basado en contenido del sistema de recomendación presentado en este artículo, fue seleccionada la última de las anteriores cuatro alternativas estudiadas. Entre las razones por las cuales fue seleccionada esta técnica de filtrado basado en contenido se encuentran las siguientes:

1. A diferencia de las dos alternativas presentadas antes que ella, la técnica seleccionada aprovecha las relaciones que existen entre los documentos que pertenecen a un mismo grupo temático, lo cual mejora la efectividad de las recomendaciones, como es demostrado en [2].

2. A diferencia de la primera alternativa estudiada, la técnica seleccionada opera sobre colecciones dinámicas de documentos en las cuales un nuevo documento puede ser adicionado de manera sencilla a la colección y al agrupamiento sin necesidad de grandes recursos de procesamiento.

3. La representación de los perfiles de usuario en un espacio vectorial de frecuencias de términos facilita el uso de técnicas conocidas de retroalimentación y generación de consultas (ver sección 2.1), a diferencia de la matriz de representación semántica multidimensional que usa la primera alternativa presentada.

En la aplicación de la técnica de filtrado basada en contenido seleccionada, fue necesaria también la utilización de la técnica de agrupamiento de documentos Star Clusters [1]. Dicha técnica de agrupamiento consiste primero en organizar los documentos del sistema en un grafo de similaridad $G_\sigma = (V, E_\sigma)$, donde V, el conjunto de vértices en el grafo, corresponde al conjunto de documentos; y $E_\sigma = \{e : w(e) \geq \sigma\}$, el conjunto de arcos en el grafo, corresponde a aquellos valores de similaridad entre los documentos que son mayores o iguales a un límite dado σ, llamado límite de agrupamiento. La similaridad entre dos documentos es medida usando la métrica del coseno, la cual es ampliamente usada en el campo de la recuperación de información.

Una vez que los documentos están organizados en el grafo de similaridad G_σ, la formación del agrupamiento entre los documentos consiste en cubrir dicho grafo con subgrafos en forma de estrella. Tales subgrafos están formados por un vértice central llamado star center (centro de estrella), m vértices satélite y arcos entre el vértice centro de estrella y cada uno de los vértices satélite (Ver Figura 3). Cada vértice satélite puede pertenecer a uno o más grupos, es decir, puede ser adyacente a uno o más centros de estrella.

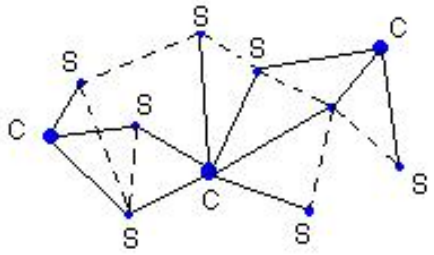

Figura 3. Grafo de similaridad ya agrupado. En el agrupamiento hay tres grupos identificados por sus centros de estrella C y los satélites S de cada grupo están conectados con los centros de estrella de los grupos a los que pertenecen por líneas continuas.

En [1] se presentan dos versiones del algoritmo de agrupamiento, una versión fuera de línea que se usa para colecciones de documentos estáticas y otra versión en línea que se usa para colecciones dinámicas de documentos, en las cuales nuevos documentos llegan de manera incremental en el tiempo y deben ser insertados en la colección, como es el caso del sistema de recomendación desarrollado. Por lo tanto, para el sistema de recomendación presentado en este artículo se implementó la versión en línea del algoritmo de agrupamiento Star Clusters, el cual soporta la inserción y eliminación de documentos del agrupamiento. Así, cada vez que llegan nuevos documentos al sistema, son adicionados con facilidad al agrupamiento ya existente. Luego de actualizar el agrupamiento con los nuevos documentos que llegan al sistema, la técnica seleccionada de filtrado de información basado en contenido, la cual es presentada en [2], usa dicho agrupamiento para determinar si un nuevo documento debe o no ser recomendado al usuario. Cada documento nuevo, además de ser comparado con el perfil del usuario, es comparado también con los documentos que pertenecen al mismo grupo y que han sido relevantes con anterioridad para el usuario; de esta manera, se mejoran los resultados del filtrado. Adicionalmente, para mejorar el desempeño del filtrado, cada nuevo documento es prefiltrado antes de ser comparado con los demás documentos de su grupo. Dicho prefiltrado consiste en que cada documento sólo es tenido en cuenta si su similaridad con el perfil del usuario (calculada también con la métrica del coseno) supera un valor límite θ, llamado límite de prefiltrado. En resumen, la técnica de filtrado de información basado en contenido que fue utilizada en el sistema de recomendación desarrollado, es una variación del algoritmo presentado en [2] y sigue los pasos que se observan en la Tabla 1 cada vez que se van a generar las predicciones de relevancia para un usuario α del sistema.

2.3. Filtrado Colaborativo

Las técnicas más comunes de filtrado colaborativo hacen uso de algoritmos basados en vecindario. Tales algoritmos comprenden tres pasos principales: el cálculo de correlación entre los usuarios, la selección del subconjunto de usuarios más similares al usuario activo y la generación del valor de predicción con base en los usuarios seleccionados.

Para la selección de la técnica de filtrado colaborativo a usar en el sistema, se examinaron diversas técnicas para el cálculo de la correlación entre los usuarios. Entre dichas técnicas se encuentran las dos siguientes. La primera técnica es basada en la medida de correlación de Pearson y es usada en el sistema GroupLens [11], la cual consiste en calcular el promedio ponderado de las desviaciones de la media obtenidas entre el usuario que se está examinando y los demás usuarios que han calificado los mismos ítems que él. La segunda técnica es una variación de la anterior, es llamada correlación de Pearson comprimida y fue usada en el sistema Ringo [14], en la cual se usa la desviación de la mediana en lugar de la desviación de la media entre las calificaciones de los usuarios. El filtro colaborativo implementado en el sistema de recomendación desarrollado utiliza la primera de las técnicas anteriores para el cálculo de las correlaciones. Dicha técnica es la más tradicionalmente usada en los sistemas de recomendación como GroupLens [11] y por ello cuenta con suficiente documentación en sus detalles de funcionalidad. Los perfiles colaborativos de los usuarios se representan con una matriz dispersa $M_{\|U\|\times\|D\|}$ en la que las filas representan los elementos en el conjunto de usuarios U y las columnas representan los elementos en el conjunto de documentos D. En cada celda de la matriz están las calificaciones

que los usuarios han dado a cada documento como retroalimentación ya sea implícita o explícita. Los valores de retroalimentación de los usuarios en el sistema de recomendación desarrollado están entre 0 y 5 inclusive (ver sección 2.5). El calculo de la correlación o similaridad entre cada dos pares de perfiles de usuario del sistema K y L se realiza de manera periódica fuera de línea. Sin embargo, para tal cálculo, no se utiliza la medida original de correlación de Pearson sino que se utiliza una modificación de la misma, la cual consiste en devaluar aquellas correlaciones entre dos perfiles colaborativos de usuarios que sólo tienen un número pequeño de ítems en común [20]. Esta devaluación se aplica multiplicando la correlación por un valor de significancia así:

$$r\prime_{KL} = \begin{cases} r_{KL} & \text{Si} \quad N_{common} > \beta \\ r_{KL} * \frac{N_{common}}{\beta} & \text{Si} \quad N_{common} \leq \beta \end{cases}$$

donde N_{common} es el número de ítems en común en los perfiles de los usuarios K y L (co-rated items) y r_{KL} es el valor de correlación de Pearson dado por:

$$r_{KL} = \frac{Cov(K,L)}{\sigma_K \sigma_L} = \frac{\sum_i (K_i - \bar{K})(L_i - \bar{L})}{\sqrt{\sum_i (K_i - \bar{K})^2}\sqrt{\sum_i (L_i - \bar{L})^2}}$$

Donde $\bar{K}$ es el promedio de valores de retroalimentación del usuario K y todas estas sumatorias y promedios se calculan sólo sobre aquellos documentos para los que ambos usuarios K y L han dado valores de retroalimentación. Esta medida de correlación está en el intervalo $[-1, 1]$.
Como se observa en la fórmula para $r\prime_{KL}$, si dos usuarios tienen más de β ítems en común en sus perfiles, se deja la correlación inalterada, es decir, no se multiplica por el valor de significancia.
En el sistema de recomendación desarrollado, se estableció el valor $\beta = 20$ para el límite de significancia.

Luego de obtener los valores de correlación entre los usuarios del sistema, para calcular la predicción de relevancia para el usuario K sobre el documento h, que dicho usuario no ha visto aún, se utiliza el promedio ponderado de todos los valores de retroalimentación que los usuarios han dado sobre el documento h de acuerdo con la siguiente ecuación:

1. Adicionar los nuevos documentos al agrupamiento usando el algoritmo Star Clusters.
2. Prefiltrar los nuevos documentos alrededor del perfil de usuario usando el límite de prefiltrado θ.
3. Para cada nuevo documento α que tenga similaridad mayor o igual a θ con el perfil de usuario se calcula su predicción de relevancia así:
   - *a*) Si el documento no está en el centro de estrella de un grupo, su predicción de relevancia se calcula como el promedio ponderado de los valores de relevancia de todos sus centros de estrella adyacentes. Los valores usados para las ponderaciones serán las distancias o similitudes entre el documento analizado y sus centros de estrella adyacentes.
   - *b*) Si el documento está en el centro de estrella de un grupo, su predicción de relevancia se calcula como el promedio ponderado de los valores de relevancia de todos sus vértices adyacentes (satélites). De igual manera, los valores usados para las ponderaciones serán las distancias o similitudes entre el documento analizado y sus vértices adyacentes.
   - *c*) Si ninguno de los vértices adyacentes a un documento tiene aún asignado un valor de relevancia porque el usuario aún no ha realizado ninguna acción de retroalimentación sobre ellos, se calcula la predicción de relevancia del documento usando su distancia o similaridad con respecto al perfil del usuario usando la métrica del coseno.

Tabla 1. Pasos de la técnica de filtrado de información basado en contenido usada en el sistema cada vez que se calculan predicciones de relevancia para un usuario. Es una variación del algoritmo presentado en [2].

$$K_h = \bar{K} + \frac{\sum_{J \in Usuarios}(J_k - \bar{J})r\prime_{KJ}}{\sum_J \|r\prime_{KJ}\|}$$

donde J_h es el valor de retroalimentación dado al documento h por uno de los usuarios del sistema que ha dado calificaciones para dicho documento y $\bar{J}$ es el promedio de valores de retroalimentación dados al documento h por los usuarios del sistema que han calificado dicho documento h. Los promedios en esta ecuación se calculan sobre todos los valores de retroalimentación que tengan los documentos y no sólo sobre los que dichos documentos tengan en común, y las sumatorias son para todos los usuarios diferentes de K que ya han dado alguna calificación al documento h.
El cálculo de las predicciones también se realiza periódicamente después de realizar el cálculo de las correlaciones entre los usuarios.

2.4. Filtrado Híbrido

Para la elección de una técnica de filtrado híbrido que combine las bondades del filtrado basado en contenido y del filtrado colaborativo, se estudiaron tres alternativas. La primera de ellas es el sistema Fab presentado en [3]. El sistema híbrido de Fab mantiene los perfiles basados en contenido de los usuarios como vectores de términos, y obtiene la información colaborativa para generar las recomendaciones comparando dichos perfiles entre si para determinar los usuarios similares. Luego, con los valores de similaridad entre los usuarios, los ítems son filtrados de tal manera que se consideran relevantes para un usuario aquellos ítems que son similares al perfil del usuario y aquellos ítems que han recibido altas calificaciones de los usuarios que tienen un perfil similar al de dicho usuario. La segunda alternativa es el sistema de recomendación ACE presentado en [13]. Este sistema híbrido calcula las predicciones de relevancia como una combinación lineal de las predicciones generadas por un filtro colaborativo y un filtro basado en contenido. Cada una de las dos predicciones que se combinan es multiplicada por un factor (k y $k-1$ respectivamente) que determina la influencia de cada tipo de predicción en el valor híbrido obtenido. La tercera alternativa es la presentada en [5]. En esta técnica de filtrado híbrido, las predicciones de relevancia híbridas son calculadas como el promedio ponderado de las predicciones generadas por los tipos de filtro de información utilizados (que pueden ser más de dos). La ventaja de esta alternativa con respecto a la primera es que mantiene la independencia entre cada tipo de filtro permitiendo mayor modularidad, pues en cualquier momento puede ser modificada la implementación de uno de los filtros de información sin afectar a los demás. Adicionalmente, a diferencia de la segunda alternativa, los pesos usados para cada filtro son almacenados independientemente para cada usuario y modificados según el desempeño que presenten los filtros a lo largo del tiempo, permitiendo así que cada usuario vaya a su propio ritmo.

En el sistema desarrollado, acogiendo la tercera de las alternativas, se usaron sólo dos filtros: uno basado en contenido y otro colaborativo, los cuales fueron presentados en las dos subsecciones anteriores. Los pesos de cada uno de los filtros para cada usuario son establecidos inicialmente con valores de 0,5, pues dichos pesos deben sumar uno en todo momento, y luego son ajustados de acuerdo con las acciones de retroalimentación del usuario (ver sección 2.5). Entonces, como la suma de los pesos debe ser igual a uno, el promedio ponderado con el cual se calcula la predicción híbrida de relevancia $p_{u,d}$ para un usuario u acerca de un documento d, está dada por:

$$p_{u,d} = \sum_{i=1}^{n} w_{i,u} q_{i,u,d}$$

donde el número de filtros n, es igual a dos para el caso presentado en este artículo, el valor $q_{i,u,d}$ representa la predicción de relevancia producida por el filtro i para el documento d y el usuario u, y el valor $w_{i,u}$ es el peso que tiene el filtro i en el cálculo de la predicción híbrida para el usuario u.

Para la generación de las recomendaciones que serán presentadas al usuario, las predicciones híbridas son calculadas de manera periódica en las noches y los documentos son ordenados descendentemente en una lista de acuerdo con su valor de predicción híbrida para presentar al usuario sólo los primeros m documentos de dicha lista. De esta manera, cuando el usuario llega a revisar sus nuevas recomendaciones no tiene que esperar por su generación pues estas ya fueron generadas con anterioridad.
Con base en la lista de documentos recomendados que le son presentados, el usuario puede interactuar con el sistema ya sea dando un valor explícito de calificación a los documentos (retroalimentación explícita), descargando o eliminando el documento (acciones de retroalimentación implícita.).

2.5. Procesamiento de las acciones de retroalimentación del usuario

Las acciones de retroalimentación de un usuario en el sistema representan las reacciones que muestra el usuario ante los documentos que el sistema de recomendación le presenta. Algunas de dichas reacciones son detectadas e interpretadas por el sistema, ya sean ellas implícitas o explícitas, para aprender cada vez más acerca de los intereses del usuario, modificar su perfil y generar un listado de documentos recomendados cada vez más acorde con dichos intereses. El sistema de recomendación desarrollado detecta e interpreta tres tipos de acciones de retroalimentación del usuario sobre uno de los documentos recomendados: eliminación, descarga y calificación directa del documento. La eliminación de un documento de la lista de documentos recomendados es una acción de retroalimentación implícita que se interpreta como un signo de que dicho documento no es interesante para el usuario. La descarga de un documento es una acción de retroalimentación implícita que se interpreta como una muestra de interés del usuario hacia el documento. La calificación directa del documento es una acción de retroalimentación explícita que se interpreta según el puntaje de calificación dado por el usuario al documento. Dicho puntaje es un valor discreto perteneciente al conjunto ordenado $\{0, 1, 2, 3, 4, 5\}$, donde 0 indica que el documento no es interesante para el usuario y 5 indica que el documento es altamente interesante para el usuario.
Para la interpretación y procesamiento de las acciones de retroalimentación del usuario, el sistema de recomendación traduce cada una de dichas acciones a un valor de relevancia en el intervalo $[-1, 1]$ como se observa en la Tabla 2.

El procesamiento de las acciones de retroalimentación del usuario consiste en la modificación de los perfiles del usuario, tanto basados en contenido como colaborativos, y en el ajuste de los pesos que tiene cada tipo de filtro de información en el cálculo de la predicción de relevancia híbrida, como se plantea en la sección anterior. La forma en la cual se realizan tales modificaciones y ajustes depende directamente de la representación usada de los perfiles de los usuarios y por lo tanto de las técnicas de cada tipo de filtrado seleccionadas como se observó en las secciones anteriores.

La modificación de los perfiles colaborativos de los usuarios consiste simplemente en adicionar un nuevo valor a la matriz $M_{\|U\|\times\|D\|}$ de usuarios por documentos (ver sección 2.3), en la celda correspondiente al usuario que realizó la acción de retroalimentación y al documento sobre el cual se realizó dicha acción.

Para la modificación de los perfiles basados en contenido del usuario se siguió la misma regla usada en [15], en la cual, dado el valor $f_{u,d}$ de relevancia interpretado a partir de la retroalimentación del usuario u sobre el documento d, se actualiza el perfil $\vec{P}_u$ del usuario, representado como un vector de términos, con base en el vector que representa el documento calificado $\vec{D}$:

$$\vec{P}_u = \vec{P}_u + \alpha_{p,u} \times f_{u,d} \times \vec{D}$$

donde el valor $\alpha_{p,u}$ representa la tasa de aprendizaje del perfil $\vec{P}_u$, indicando la sensibilidad de dicho perfil a las retroalimentaciones del usuario, y $f_{u,d}$ representa el valor de relevancia interpretado por el sistema (ver Tabla 2).
Para el sistema de recomendación desarrollado, el valor de $\alpha_{p,u}$ para todos los perfiles se estableció como $\alpha_{p,u} = 1$.

Acción de retroalimentación del usuario u **ante el documento** d	**Valor de relevancia interpretado por el sistema** $f_{u,d}$
Eliminación de documento	-1
Descarga de documento	0,6
Calificación con valor 0	-1
Calificación con valor 1	$-0{,}6$
Calificación con valor 2	$-0{,}2$
Calificación con valor 3	0,2
Calificación con valor 4	0,6
Calificación con valor 5	1

Tabla 2. Valores de relevancia interpretados por el sistema de recomendación a partir de las acciones de retroalimentación del usuario.

En el ajuste de los pesos de cada filtro de información (basado en contenido y colaborativo), los cuales se usan en el cálculo de las predicciones híbridas de relevancia para un usuario, el objetivo es minimizar el error promedio $\bar{E}_{i,u}$ de las predicciones generadas por cada filtro i para el usuario u:

$$\bar{E}_{i,u} = \frac{E_{i,u}}{l_{i,u}},$$

$$E_{i,u} = \sum_{j=1}^{l_{i,u}} |pr_{i,u,j} - f - u, j|$$

donde $l_{i,u}$ es el número total de documentos para los cuales el filtro de información i ha generado predicciones, $pr_{i,u,j}$ es el valor de la predicción generada por el filtro de información i para el usuario u y el documento j, y $f_{u,j}$ es el valor de relevancia interpretado por el sistema de recomendación con base en una acción de retroalimentación del usuario u sobre el documento recomendado j.
Entonces, a partir del error promedio de predicción $\bar{E}_{i,u}$ calculado para el filtro de información i y el usuario u, los nuevos pesos para los filtros de información son calculados de tal manera que sumen uno y sean inversamente proporcionales a dicho error:

$$w_{i,u} = \frac{1}{\bar{E}_{i,u}}$$

Así, con la modificación de los perfiles del usuario, tanto colaborativos como basados en contenido, y con el ajuste de los pesos de filtros de información usados para el cálculo de las predicciones híbridas, se ejecuta el procesamiento de las acciones de retroalimentación realizadas por los usuarios sobre los documentos que les son recomendados. Con este procesamiento el aprendizaje del sistema es continuo y acorde con los cambios graduales en los intereses de los usuarios.

2.6. Medición de la calidad de las recomendaciones

Después del diseño y desarrollo de un sistema de recomendación, como en cualquier otro sistema, se hace necesaria la medición de la calidad de sus resultados con el fin de analizar los puntos fuertes y débiles de su funcionamiento, visualizar las variables involucradas, diagnosticar las causas de posibles fallas y descubrir la forma de afectar dichas variables para optimizar los resultados. Herlocker y otros presentaron en [8] un estudio acerca de la medición de la calidad en los sistemas de recomendación a través de la evaluación de la exactitud en sus predicciones. Según tal estudio, una métrica para la exactitud de un sistema de recomendación mide, empíricamente, que tan cerca está el ordenamiento de ítems predichos para un usuario por el sistema con respecto al ordenamiento verdadero que el usuario haría, según su preferencia, de los mismos ítems. En [8] se presentan además varias métricas que son categorizadas en tres clases: métricas de exactitud predictiva, métricas de exactitud en clasificación y métricas de exactitud en ordenamiento. En la clase de métricas de exactitud predictiva están aquellas métricas que miden qué tan cerca están los puntajes de relevancia predichos por el sistema con respecto a los puntajes reales dados por un usuario. Entre estas métricas están el error medio absoluto (MAE por sus siglas en inglés) y otras métricas relacionadas como el error medio absoluto normalizado y el error medio cuadrado. En la clase de métricas de exactitud en clasificación están aquellas métricas que miden la frecuencia con la cual un sistema de recomendación toma decisiones incorrectas o correctas acerca de si un ítem es bueno para el usuario. Entre estas métricas están Precision and Recall y las curvas ROC (Receiver Operating Characteristic). En la clase de métricas de exactitud de ordenamiento están aquellas que miden la habilidad de un sistema de recomendación para producir un ordenamiento de ítems recomendados que coincida con el ordenamiento que un usuario haría de los mismos ítems. Entre estas métricas están la correlación predicción-calificación, métrica half-life utility y la medida NDPM.

Para la evaluación del sistema de recomendación desarrollado y presentado en este artículo, como se detalla en la sección 3, se usaron las curvas ROC. El modelo de las curvas ROC (Receiver Operating Characteristic) trata de medir el grado con el cual un sistema de filtrado de información puede distinguir de manera exitosa entre relevancia y ruido [8].Una curva ROC representa una gráfica del *recall* versus el *fallout*, donde *recall* es el porcentaje de ítems relevantes (proporción de verdaderos positivos) recomendados a los usuarios y el *fallout* es el porcentaje de ítems no relevantes (proporción de falsos positivos) recomendados. Cada punto en la curva ROC corresponde a un valor límite de predicción de relevancia a partir del cual el sistema considera que la predicción de un ítem es positiva o negativa (relevante o no relevante respectivamente). Los puntos ubicados en la parte más a la izquierda de la curva corresponden con los valores límite mayores de predicción de relevancia, por lo cual es la parte más exigente de la curva. Los puntos ubicados más a la derecha de la curva corresponden con los valores límite menores, por lo cual es la parte menos exigente de la curva. Los valores de *recall* y *fallout*, se calculan con base en las llamadas matrices de confusión o tablas de contingencia como la presentada en la Tabla 3.

		Verdadera clasificación dada por el usuario	
		Positivo	**Negativo**
Clasificación dada por el sistema de recomendación	**Positivo**	Verdaderos Positivos (**TP**)	Falsos Positivos (**FP**)
	Negativo	Falsos Negativos (**FN**)	Verdaderos Negativos (**TN**)
	Totales	Total Positivos (**P**)	Total Negativos (**N**)

Tabla 3: Tabla de contingencia o matriz de confusión con base en la cual se calculan los valores de recall y fallout para las gráficas ROC.

Un ítem es considerado *verdadero positivo* cuando dicho ítem es en realidad positivo para el usuario y es también clasificado como positivo por el sistema de recomendación. Si tal ítem es positivo para el usuario pero es clasificado como negativo por el sistema de recomendación, es considerado un *falso negativo* (ver Tabla 3).
Un ítem es considerado *verdadero negativo* cuando dicho ítem es en realidad negativo para el usuario y es también clasificado como negativo por el sistema de recomendación. Si tal ítem es negativo para el usuario pero es clasificado como positivo por el sistema de recomendación, es considerado un *falso positivo* (ver Tabla 3).

$$recall = TP\_rate = \frac{TP}{P}$$

$$fallout = FP\_rate = \frac{FP}{N}$$

El área bajo una curva ROC es una medida que resume el desempeño del sistema de filtrado y es equivalente a la probabilidad de que el sistema esté en capacidad de elegir correctamente entre dos ítems (documentos en este caso), uno seleccionado aleatoriamente a partir del conjunto de ítems no relevantes y el otro seleccionado aleatoriamente a partir del conjunto de ítems relevantes. De manera intuitiva, según [8], la medida AUC (area bajo la curva ROC) captura el *recall* del sistema a diversos niveles de *fallout*. Según esta interpretación del área bajo la curva, un sistema de filtrado de información perfecto generaría una curva ROC tal que su área AUC es igual a 1 (ver Figura 4); y un sistema de filtrado que genera predicciones aleatorias, el cual representa el peor caso, produciría una curva ROC tal que su área AUC es igual a 0,5 (ver Figura 5).

Luego de presentar en esta sección la teoría acerca de las curvas ROC y el área bajo la curva AUC, se detalla, en la siguiente sección, la forma en la cual se llevaron a cabo las pruebas del sistema de recomendación desarrollado y el uso de las curvas ROC para evaluar los resultados obtenidos en dichas pruebas.

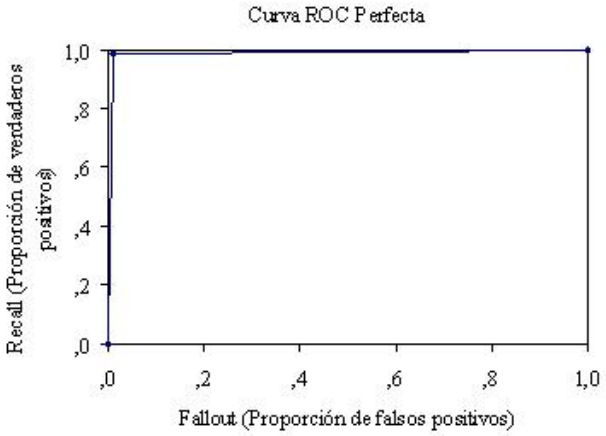

Figura 4. Curva ROC para un sistema de filtrado perfecto.

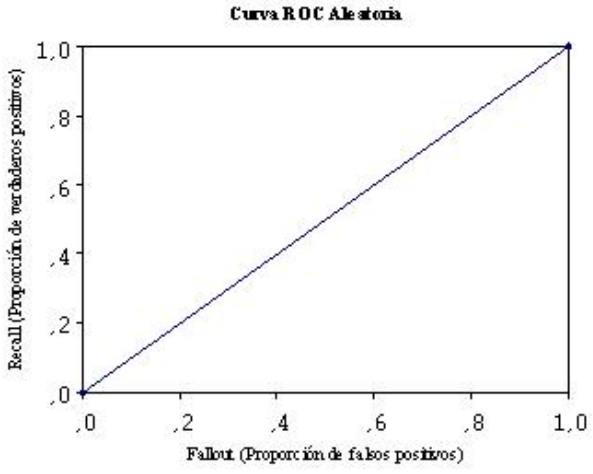

Figura 5. Curva ROC para un sistema de filtrado que genera predicciones aleatorias.

3. Evaluación del sistema

Para evaluar la efectividad y la línea de aprendizaje del sistema de recomendación desarrollado, se realizaron experimentos durante cinco días consecutivos con cuatro usuarios reales involucrados en proyectos de investigación relacionados con el área de ciencias de la computación.

Cada usuario se registró en el sistema con un tema de interés (ver Tabla 4) y proporcionó un conjunto de 23 documentos de entrenamiento para los cuales, a su vez, asignó un valor de calificación de relevancia con respecto al tema de interés en una escala de 0 a 5 inclusive. Con dichos documentos de entrenamiento y con las palabras clave ingresadas por cada usuario en el momento de su registro, se generaron y entrenaron los perfiles iniciales basados en contenido para dichos usuarios.

Usuario	Tema de interés
1	Online Public Access Catalog
2	E-Learning
3	Requirements Engineering
4	Evolutionary Computing

Tabla 4. Temas de interés definidos por los usuarios durante su registro en el sistema para la realización de los experimentos

Luego del entrenamiento, para cada día de los experimentos, el sistema desarrollado ejecutó las tareas de generación de consultas a partir de los perfiles de usuario, búsqueda y descarga de documentos, cálculo de predicciones de relevancia para los nuevos documentos, por parte de cada uno de los filtros de información, y generación del conjunto de recomendaciones del día, el cual está conformado por los primeros 30 documentos con mayor predicción híbrida de relevancia.
Los usuarios por su parte, durante cada día de los experimentos, ingresaron al sistema y examinaron la lista de los 30 documentos recomendados para el día en cuestión, asignando a cada uno de ellos un puntaje de calificación, ya sea de manera explícita o implícita. Con base en dichos puntajes de calificación el sistema realizó los ajustes necesarios a los perfiles para comenzar de nuevo, durante el día siguiente, el ciclo de generación de consultas y generación de nuevas recomendaciones.

El propósito de los experimentos consistió en la medición del cambio en la exactitud de las predicciones híbridas generadas para los documentos recomendados a medida que pasan los días de prueba. Para tal medición se usó la métrica del área Bajo la Curva ROC (AUC por sus siglas en inglés) presentada en la sección anterior.

Para medir la exactitud de las predicciones generadas hasta el momento durante cada día, se construyeron entonces cinco gráficas ROC, una por cada día de los experimentos, y se calculó el área bajo la curva AUC de cada una de ellas. Tales gráficas se pueden observar juntas en la Figura 6 para facilitar su comparación.
Además, para ver con mayor claridad el cambio en el área bajo la curva ROC durante los días en que se llevó a cabo el experimento, los valores de dicha área para cada día son presentados en la Tabla 5 y graficados en la Figura 7.

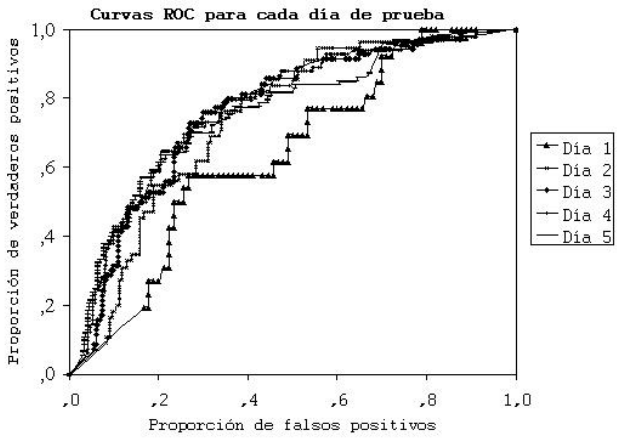

Figura 6. Gráficas ROC para las recomendaciones producidas acumuladas durante cada día de los experimentos.

La parte más interesante de la gráfica representada en la Figura 6 es aquella que corresponde con los puntos del lado izquierdo. Tales puntos corresponden a valores límite mayores de predicción de relevancia tal como fue descrito en la sección anterior. Es decir, dichos puntos corresponden con una ventana más pequeña de documentos recomendados que son presentados al usuario en el tope de la lista ordenada ascendentemente por predicción de relevancia. Es la parte más exigente de la curva.
Examinando entonces la parte más a la izquierda de las curvas ROC de la Figura 6, se observa que a medida que pasan los días de prueba disminuye la proporción de falsos positivos presentes en el tope de la lista de documentos recomendados a los usuarios. Esta disminución es mucho mayor al comienzo de las pruebas, entre el primer y tercer día, y luego se hace más pequeña. Inclusive, si se

sigue la trayectoria de las curvas, se observa que aquellas correspondientes a los dos últimos días (cuarto y quinto) permanecen casi iguales hasta que se alcanza una proporción de falsos positivos de aproximadamente 0,354 (35 %) y una proporción de verdaderos positivos 0,65 (65 %), volviéndose a unir de nuevo al final. Observando el comportamiento del área bajo la curva (AUC) en la Figura 7, también se nota que el mayor aprendizaje del sistema y, por lo tanto, la mayor mejora en su rendimiento se da entre el primer y tercer día de las pruebas. Luego, la variación se hace más pequeña comenzando así una etapa de aprendizaje más lento llegando incluso a disminuir un poco para el último día de pruebas.

Día de Pruebas	Área Bajo la Curva ROC
Día 1	0.639
Día 2	0.739
Día 3	0.761
Día 4	0.771
Día 5	0.764

Tabla 5. Valores del área bajo la curva ROC calculados durante los diferentes días de los experimentos realizados.

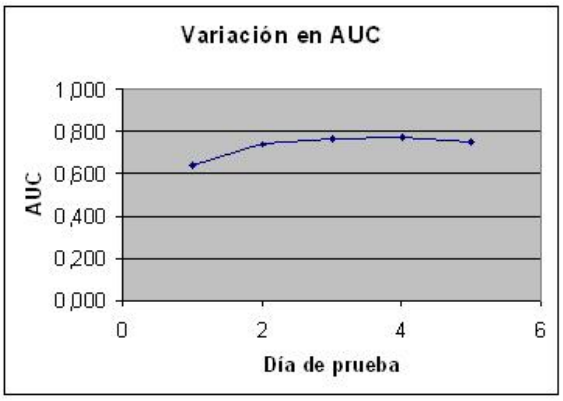

Figura 7. Cambio del área bajo la curva ROC observado durante los días de prueba del sistema.

El mismo comportamiento se observa al graficar el cambio en el error medio absoluto (MAE por sus siglas en inglés) durante los días de la prueba, como se ve en la Figura 8. Las mayores disminuciones en el error se observan entre los primeros dos días de la prueba y luego, el cambio en el error se va haciendo más pequeño.

Los cambios cada vez más pequeños en el desempeño del sistema observados en cada una de las gráficas de las Figuras 6, 7 y 8, podrían explicarse por la poca variación en los términos de las consultas generadas, a partir de cada uno de los perfiles de los usuarios, para la obtención de nuevos documentos durante los últimos dos días de pruebas (ver Tabla 6).

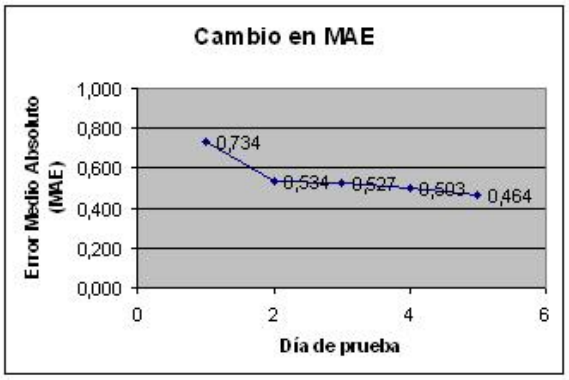

Figura 8. Cambio en el error medio absoluto (MAE) durante los días de pruebas.

Con estas consultas, el número de documentos nuevos obtenidos, a partir de los mismos motores de búsqueda, ya mencionados en la sección 2.1, era cada vez menor, disminuyendo así el número de documentos con una buena predicción de relevancia disponibles para recomendar a los usuarios cada día.

Otro factor que puede influir en la rápida optimización de los perfiles de los usuarios y por lo tanto en la poca variación de las consultas al final de las pruebas, es el valor constante de $\alpha_{p,u} = 1$ establecido para la tasa de aprendizaje de dichos perfiles, como se indicó en la sección 2.5.

Adicional a la poca variación en los términos de las consultas generadas, la frecuencia con la cual se generaron nuevas recomendaciones (una vez cada día) es quizá demasiado alta comparada con la frecuencia de ingreso de nuevos documentos hacia el interior de los motores de búsqueda utilizados, por lo cual se podría explicar también el poco flujo de documentos nuevos que ingresaron al sistema durante los últimos días de pruebas. En consecuencia, sería muy importante determinar una frecuencia de generación de recomendaciones adecuada de acuerdo con los motores de búsqueda utilizados.

Sin embargo, para comprobar las hipótesis anteriores con respecto a las causas del cambio lento en el desempeño del sistema durante los últimos días

de experimentación, se hace necesaria una evaluación mucho más detallada y extensa que permita la obtención de más datos de prueba, contando con un mayor número de usuarios, temas de interés por usuario y motores de búsqueda.

Perfil de Usuario	Día de Prueba	Términos de Consulta Generada
Usuario 1	Día 1	Queri web search term library
	Día 2	Queri search web term document
	Día 3	web queri searcher ir retriev
	Día 4	web queri searcher ir opac
	Día 5	web searcher opac ir manifest
Usuario 2	Día 1	adapt system user model concept
	Día 2	adapt system model knowledg educ
	Día 3	adapt system model educ user
	Día 4	adapt system user educ model
	Día 5	adapt educ knowledg concept learn
Usuario 3	Día 1	scenario requir model system process
	Día 2	requir scenario process goal model
	Día 3	requir scenario goal engin action
	Día 4	scenario requir goal engin action
	Día 5	scenario requir goal engin action
Usuario 4	Día 1	popul genet evolutionari mutat algorithm
	Día 2	algorithm popul genet optim evolutionari
	Día 3	genet popul evolutionari algorithm size
	Día 4	genet popul evolutionari fit jectiv
	Día 5	genet popul evolutionari fit jectiv

Tabla 6. Términos de las consultas generadas durante los cinco días de pruebas para cada usuario en su único tema de interés o perfil.

4. Conclusiones

En el presente artículo se expuso el desarrollo de un sistema de búsqueda personalizada y recomendación de documentación científica. Se presentaron las diversas técnicas usadas para desarrollar cada una de las partes del sistema: generación automática de consultas, búsqueda de documentos, filtrado de información híbrido, colaborativo y basado en contenidos, con el respectivo procesamiento de las acciones de retroalimentación del usuario, tanto implícitas como explícitas.

Adicionalmente, se expusieron los resultados de experimentos preliminares llevados a cabo con usuarios reales durante un período de tiempo de cinco días. Con base en dichos resultados, usando la métrica del área bajo la curva ROC, se pudo demostrar que a lo largo de los días de prueba, el desempeño del sistema presentó cambios positivos (aumento en el desempeño), los cuales fueron mucho mayores durante los primeros tres días de pruebas y luego más pequeños hasta alcanzar un área bajo la curva ROC con valor de 0,754 y un error medio absoluto (MAE) de 0,464. Sin embargo, aunque las gráficas de desempeño presentadas como curvas ROC y error absoluto medio demuestran que durante el escaso tiempo de pruebas se obtuvieron cambios positivos en el aprendizaje del sistema, se requiere la realización de pruebas más exhaustivas, con un mayor número de usuarios, motores de búsqueda y tiempo para recolectar datos que nos permitan analizar el punto de estabilización en aprendizaje del sistema.

Luego de los experimentos realizados, se analizaron de manera intuitiva las posibles causas de la disminución en cambio del desempeño del sistema durante los últimos días de pruebas, llegando a la formulación de una hipótesis preliminar. Tal hipótesis plantea que la frecuencia con la cual se generan nuevas recomendaciones es un factor muy importante en los resultados obtenidos, el cual debe estar de acuerdo con el flujo de nuevos documentos dentro de los motores de búsqueda utilizados y con la frecuencia de generación de consultas. A su vez, las consultas generadas son afectadas también por los cambios en los perfiles de los usuarios para los cuales, se estableció como parámetro, una tasa de aprendizaje constante con valor 1 para los experimentos. Para futuros experimentos que incluyan la realización de pruebas más exhaustivas, será muy importante analizar la variación de dicho parámetro en conjunto con la frecuencia de generación de recomendaciones.

Adicionalmente, es de destacar lo útiles que resultarían sistemas de recomendación como este en el ámbito de la investigación y para la formación de bases de datos de prueba y entrenamiento que

puedan ser usadas en la evaluación de sistemas de recomendación híbridos y basados en contenido, pues hasta el momento son mucho más accesibles los datos para evaluar sistemas colaborativos.

Referencias

[1] Aslam, Pelehov, and Rus. A practical clustering algorithm for static and dynamic information organization. In *SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms)*, 1999.

[2] Javed Aslam, Katya Pelekhov, and Daniela Rus. Using star clusters for filtering. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 306–313, New York, NY, USA, 2000. ACM Press.

[3] Marko Balabanovic and Yoav Shoham. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72, 1997.

[4] Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles. Discovering relevant scientific literature on the web. *IEEE Intelligent Systems*, 15(2):42–47, 2000.

[5] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. Combining content-based and collaborative filters in an online newspaper, 1999.

[6] Daniel Dreilinger and Adele E. Howe. Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, 15(3):195–222, 1997.

[7] P. W. Foltz. Using latent semantic indexing for information filtering. In *Proceedings of the conference on Office information systems*, pages 40–47, New York, NY, USA, 1990. ACM Press.

[8] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[9] Michael J. Pazzani, Jack Muramatsu, and Daniel Billsus. Syskill webert: Identifying interesting web sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.

[10] Stefan Pitzek. Impact of online-availability of scientific literature, March 2002.

[11] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.

[12] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[13] Ramón Sangúesa, Alberto Vázquez, and Javier Vázquez Salceda. The ace recommender system. Technical Report LSI-01-21-R, Universitat Politècnica de Catalunya, 2001.

[14] Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1, pages 210–217, 1995.

[15] Beerud D. Sheth. A learning approach to personalized information filtering. Master's thesis, January 1994.

[16] Gabriel L. Somlo and Adele E. Howe. Using web helper agent profiles in query generation. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 812–818, New York, NY, USA, 2003. ACM Press.

[17] Daniel R. Tauritz. *Adaptive Information Filtering: concepts and algorithms*. PhD thesis, Leiden University, 2002.

[18] Steven Wright. Personalisation: How a computer can know you better than yourself, 2002.

[19] Kai Yu. *Statistical Learning Approaches to Information Filtering*. PhD thesis, LMU Munich: Faculty of Mathematics, Computer Science and Statistics, 2004.

[20] Chun Zeng, Chun-Xiao Xing, and Li-Zhu Zhou. Similarity measure and instance selection for collaborative filtering. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 652–658, New York, NY, USA, 2003. ACM Press.

[21] Byoung-Tak Zhang and Young-Woo Seo. Personalized web-document filtering using reinforcement learning. *Applied Artificial Intelligence*, 15(7):665–685, 2001.