Propuesta de Procesos Complementarios para un Sistema de Recuperación de Información

Rey, M., Kuna, H., Rambo, A., Canteros, A., Cantero, A., Martini, E., Corrales, N., Rauber, F.

1. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Naturales, Universidad Nacional de Misiones.

hdkuna@gmail.com

RESUMEN

Con la finalidad de mejorar el proceso de búsqueda de información para investigadores por medio de la utilización de un Sistema de Recuperación de Información (SRI) específico de las ciencias de la computación, es necesario definir métodos que trabajen sobre el análisis de los datos existentes para generar operaciones que incrementen la relevancia de los resultados a presentar a los usuarios. Entre las alternativas de técnicas aplicables se destacan: el tratamiento de tópicos clave, técnicas de clustering y de análisis de probabilidad basadas en Bayes, entre otras. En el presente trabajo se exponen propuestas de procesos a través de las que se busca demostrar que es posible utilizar distintas técnicas para generar procesos que, a partir de datos disponibles en el sistema, proporcionen información para mejorar la calidad de los resultados y operaciones internas de la herramienta, siendo este el objetivo principal de la presente línea de investigación.

Palabras clave: *información científica, meta-buscador, clustering, topic modelling, análisis de probabilidad.*

CONTEXTO

Está línea de investigación articula el "Programa de Investigación en Computación" de la Facultad de Ciencias Exactas Químicas y Naturales (FCEQyN) de la Universidad Nacional de Misiones (UNaM) con el Grupo de Investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

1 INTRODUCCION

1.1 Antecedentes

En trabajos anteriores se ha presentado un meta-buscador orientado a la recuperación de información científica correspondiente al área de ciencias de la computación [1]. A lo largo del proceso de desarrollo de esta herramienta se han integrado componentes y módulos cuya función consiste en maximizar la relevancia de los resultados a presentar al usuario [2].

Sin embargo, se considera que el desarrollo e integración de procesos cuya operatoria explote los datos almacenados por el meta-buscador es el siguiente paso en el desarrollo de la solución. Considerando que el objetivo central de cualquier SRI es maximizar la relevancia de los resultados a presentar a su usuario, la evolución del meta-buscador desarrollado pasa por la integración de funciones anexas a las básicas de búsqueda y evaluación [3]. Procesos como los planteados en el presente trabajo se consideran de gran utilidad para la recuperación de información de mayor utilidad.

1.2 Gestión de datos del SRI

En trabajos anteriores se han detallado algunos de los inconvenientes detectados en el meta-buscador. En la mayoría de los casos se trató de operaciones en las que se identificaba un problema común, la falta de datos almacenados internamente por el SRI tanto como resultado de su operatoria como de la interacción con el usuario. Se determinó que sin esta capacidad el desarrollo de procesos complementarios a los básicos del SRI sería una tarea de gran complejidad [2].

En este sentido, se implementó un esquema para la representación de los meta-datos de

los elementos con los que habitualmente opera el meta-buscador. Además de desarrollar la funcionalidad de almacenamiento de los datos involucrados en cada operación de búsqueda realizada.
Como resultado, el SRI progresivamente almacena datos que posteriormente y en una instancia de procesamiento off-line se complementan a través de procesos de ETL (Extracción, Transformación y Carga por su sigla en inglés). Tales operaciones, parten de los registros básicos de las entidades almacenadas y obtienen los datos necesarios para completar los perfiles definidos para cada una de ellas en la base de datos (BD) del SRI [3].

1.3 Procesos complementarios para el SRI y tecnologías relacionadas

Con un mayor volumen de datos en la BD del meta-buscador, se propuso comenzar con el desarrollo de procesos complementarios al mismo que permitieran incrementar la relevancia de los resultados a presentar al usuario y optimizar su funcionamiento en general.
Esta es una alternativa que se reconoce en otros SRI, operando sobre diferentes contextos y conjuntos de datos. En ellos, se reconocen diversas técnicas que permiten mejorar la experiencia del usuario con el sistema. Un ejemplo es el caso de recomendación de productos en una tienda virtual en base a búsquedas y compras previas del usuario, utilizando técnicas de clustering y filtrado basado en perfiles y datos de opiniones [4]. En las plataformas de contenido como pueden ser Netflix o Spotify se reconocen herramientas para recomendación u armado de listas de reproducción basado en intereses demostrados por el usuario o sus contactos [5, 6]. La estimación de utilidad de resultados con base en una consulta ingresada por el usuario, contextualizando la misma sobre selecciones previas que hubieran resultado satisfactorias para otros usuarios [7] y su revisión a partir de la consideración de los perfiles de usuario [8, 9] son otro tipo de soluciones aplicables en las que se encuentran implementaciones en la actualidad. De igual manera, la detección de patrones de uso o navegación sobre sitios web pueden ser utilizados para ofrecer contenido específico a un usuario [10, 11]. En esta última alternativa, la identificación de outiers e inliers [12] es un aspecto que hace a la calidad de la solución que percibe el usuario, y dada la naturaleza de los datos se deberá evaluar la vinculación a métodos como los prototipos difusos propuestos por Zadeh [13].
En el ámbito de SRI que operan con información científica se reconocen iniciativas similares. En algunas se han planteado métodos para aumentar la precisión en la asignación de algunos meta-datos a las publicaciones [14] permitiendo mejorar su proceso de clasificación, además de la asignación automática de categorías en las que una BD pueda ordenar su catálogo a través del análisis de referencias bibliográficas [15]. Con respecto al tratamiento de términos clave (keywords), su reconocimiento en documentos científicos es de utilidad para la estimación del contenido del mismo y facilitar su clasificación a través de diferentes métodos, como, por ejemplo: la relación con los términos presentes en los títulos y nombres de fuentes de publicación [16]. En otras soluciones se reconoce el uso del coeficiente TF-IDF (Term Frequency - Inverse Document Frequency) [17, 18] junto a técnicas de clustering para generar métodos de clasificación automática para contenido web [19]. Finalmente, se reconocen procesos de recomendación de resultados basados en técnicas de topic modelling en conjunto con otras de análisis de probabilidad [20, 21].
A partir de la enumeración anterior y otros ejemplos similares existentes en la bibliografía, se puede considerar que el desarrollo de procesos complementarios para un SRI forma parte de la evolución del mismo. De esta manera, se plantea al objetivo principal del presente trabajo, el desarrollo de procesos complementarios que generen un impacto directo en la

experiencia del usuario con un meta-buscador que opera sobre documentos científicos del área de ciencias de la computación.

2 LÍNEAS DE INVESTIGACION, DESARROLLO E INNOVACIÓN

Para un investigador científico la búsqueda de información en internet implica la utilización de herramientas especializadas que permitan obtener un mayor volumen de resultados relevantes y de estrecha relación con las consultas que ejecute sobre las mismas. En este sentido, la integración de procesos complementarios en un meta-buscador de propósito específico como el producto de la presente línea de investigación, constituye un área de trabajo de sumo interés ya que guarda estrecha relación con la efectividad de sus operaciones.
En este contexto, la utilización de técnicas y métodos de probada efectividad en otras áreas como inteligencia artificial, explotación de información y análisis de grandes volúmenes de datos, son los pilares del desarrollo de los procesos mencionados. Su integración permitirá incrementar la relevancia y calidad integral de los resultados a presentar al usuario.

3 RESULTADOS Y OBJETIVOS

3.1 Procesos planteados

Inicialmente se han planteado procesos centrados en la presentación de resultados al usuario. La primera propuesta consiste en métodos para agrupar los resultados considerando el área temática sobre la cual el usuario con base en análisis de tópicos y métodos de probabilidad bayesiana. Mientas que la segunda propuesta abarca la presentación de resultados recomendados a partir de los elementos de la BD del SRI, implementando técnicas específicas para este tipo de operaciones.

3.1.1 Proceso 1

La propuesta en este caso comienza por reconocer las palabras clave que se encuentren en los resultados de la ejecución de una búsqueda y determinar cuál es su probabilidad de ocurrencia en las diferentes áreas temáticas con las que opera la herramienta. Concretamente se pretende aumentar la precisión en la clasificación de los resultados a una determinada área temática a fin de poder organizar mejor el listado final en base a las preferencias del usuario.
En este sentido, cobran importancia las keywords de los resultados, ya que pueden tomarse como indicadores temáticos representativos del contenido del documento resultante de la búsqueda. La utilización de éstas como recurso para un proceso de clasificación de los resultados que presenta el meta-buscador permite la utilización de un amplio conjunto de técnicas como ser: clustering, determinación de frecuencia ocurrencia de términos sobre una colección de documentos y topic modelling.
El proceso que se presenta en esta sección se encuentra en una etapa de diseño, se ha determinado que el mismo opere sobre los resultados que se presentan al usuario una vez ejecutada una búsqueda por el SRI. Sobre ese listado se aplicarían técnicas de topic modelling para la extracción de términos clave que serán utilizados en un proceso de clasificación para determinar el área temática a la que pertenecen. Como resultado se podrán aplicar diferentes técnicas para ponderar los resultados que resulten más cercanos al área de interés del usuario.
Con respecto a su implementación, se ha avanzado en la generación de un prototipo de módulo integrado al SRI que captura los resultados obtenidos del proceso de búsqueda y extrae los tópicos más representativos de cada uno de ellos. Actualmente se está generando con base en la taxonomía definida por la ACM una colección de documentos base para cada área y subárea en la que se divide la disciplina. El paso siguiente consiste en la extracción de tópicos de cada conjunto de documentos a fin de establecer un conjunto de keywords relacionado con cada ítem de

la taxonomía. Una vez que se cuente con estos elementos, se procederá con la selección de las técnicas a través de las cuales se realizará la clasificación de los resultados, incluyendo en el análisis a los grupos antes mencionados: clustering, modelos basados en probabilidad y técnicas de análisis de frecuencia de términos.

3.1.2 Proceso 2

Este proceso tiene por objetivo realizar una recomendación de datos de autores relacionados con la temática de la consulta ingresada por el usuario. Para ello utiliza la BD interna del SRI, obteniendo un conjunto de perfiles de autores que se pueden considerar influyentes en el área de la consulta. De esta manera, el meta-buscador podría presentar resultados propios mientras se ejecutan las consultas sobre las fuentes externas, incrementando la interacción con el usuario y disminuyendo los tiempos de espera.

El proceso en cuestión se encuentra en una fase de diseño, evaluando las técnicas aplicables según el estado del arte referido a sistemas de recomendación. En este sentido, entre las opciones disponibles se presentan inicialmente aquellas basadas en contenido, en las que una opción es utilizar perfiles de usuario generados a partir del contenido que reconocen de utilidad en el uso de la herramienta y así recomendar nuevo contenido en base a tales preferencias. Por otra parte, existen métodos denominados colaborativos, en estos la recomendación se genera a partir de una comparación entre perfiles de usuario que sean considerados similares, por lo tanto, un elemento que haya sido de utilidad para un usuario sería de potencial interés para otros usuarios con un perfil análogo, generando así la recomendación. Finalmente, se reconocen métodos que toman características de los mencionados previamente buscando unificar capacidades y disminuir el impacto de problemas que pueden presentar en su operatoria. De esta manera, los métodos híbridos, se presentan como una alternativa flexible en aquellos casos en los que las recomendaciones no pueden producirse únicamente a partir de contenido o en forma colaborativa.

En la actualidad se están evaluando en detalle los escenarios en los que cada método obtiene los mejores resultados, además de analizar las técnicas necesarias para su implementación, esperando obtener una propuesta de método a corto plazo para iniciar su desarrollo.

3.2 Trabajos en curso y a futuro

En el marco de la presente investigación la prioridad actual del desarrollo se centra en la finalización del diseño e implementación de los procesos complementarios. En relación a esta actividad, el desarrollo de un módulo de gestión de usuarios, que permita la definición de perfiles de los usuarios del SRI se presenta como una necesidad en un futuro cercano. Por otro lado, la incorporación de técnicas de detección de outliers e inliers y vinculando estos conceptos a los prototipos difusos propuestos por Zadeh, realizando esta detección en los perfiles tanto de los usuarios como de las otras entidades almacenadas en la BD del meta-buscador, con el objetivo de detectar elementos tales como: intrusos, mal uso del SRI, datos erróneos en los datos de las entidades, así como nuevas tendencias de búsqueda.

4 FORMACION DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del "Programa de Investigación en Computación" de la FCEQyN de la UNaM, con diez integrantes relacionados con las carreras de Ciencias de la Computación de la UNaM. De los cuales tres están realizando su tesis de grado, tres se encuentran realizando una maestría. La línea y el equipo de investigación se vinculan con el Grupo de Investigación SMILe de la Universidad de Castilla-La Mancha, España.

5 BIBLIOGRAFIA

1. Kuna, H., et al: Avances en la Construcción de un Sistema de

Recuperación de Información para Información Científica en Ciencias de la Computación. XVIII Workshop de Investigadores en Ciencias de la Computación (2016).
2. Rey, M., Kuna, H., et al: Propuesta de Esquemas de Perfiles para la Recuperación de Datos Científicos para un Sistema de Recuperación de Información del Área de Ciencias de la Computación. XXII Congreso Argentino de Ciencias de la Computación, San Luis, Argentina (2016).
3. Rey, M., Kuna, H., Martini, E., Canteros, A., Cantero, A., Rambo, A., Biale, C., Corrales, N.: Un Meta-buscador como Plataforma para el Desarrollo de Procesos de Explotación de Datos Científicos. IV Seminario Argentina-Brasil de Tecnologías de la Información y la Comunicación. Corrientes, Argentina (2016).
4. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing. 7, 76–80 (2003).
5. Gomez-Uribe, C.A., Hunt, N.: The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 6, 13:1–13:19 (2015).
6. Sander Dieleman: Recommending music on Spotify with deep learning, http://benanne.github.io/2014/08/05/spotify-cnns.html (2014).
7. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query Recommendation Using Query Logs in Search Engines. In: Current Trends in Database Technology - EDBT 2004 Workshops. pp. 588–596. Springer, Berlin, Heidelberg (2004).
8. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive Web Search Based on User Profile Constructed Without Any Effort from Users. In: Proceedings of the 13th International Conference on World Wide Web. pp. 675–684. ACM, New York, NY, USA (2004).
9. Fawcett, T., Provost, F.J.: Combining Data Mining and Machine Learning for Effective User Profiling. In: KDD. pp. 8–13 (1996).
10. Somlo, G.L., Howe, A.E.: Adaptive Lightweight Text Filtering. In: Advances in Intelligent Data Analysis. pp. 319–329. Springer, Berlin, Heidelberg (2001).
11. Pazzani, M., Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning. 27, 313–331 (1997).
12. Hawkins, D.M.: Identification of outliers. Taylor & Francis (1980).
13. Zadeh, L.A.: A note on prototype theory and fuzzy sets. Cognition. 12, 291–297 (1982).
14. Gómez Núñez, A.J.: Una aproximación multimetodológica para la clasificación de las revistas de Scimago Journal & Country Rank (SJR), (2016).
15. Gómez-Núñez, A.J., Vargas-Quesada, B., de Moya-Anegón, F., Glänzel, W.: Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. Scientometrics. 89, 741 (2011).
16. Garland, K.: An experiment in automatic hierarchical document classification. Information Processing & Management. 19, 113–120 (1983).
17. Salton, G.: Developments in automatic text retrieval. science. 253, 974 (1991).
18. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management. 24, 513–523 (1988).
19. Muñoz, G., del Cisne, M.: Módulo para Clasificación Automática y Temática de Páginas Web., (2012).
20. Hernández, A., Tomás, D., Navarro Colorado, B.: Una Aproximación a la Recomendación de artículos científicos según su grado de especificidad. Procesamiento del Lenguaje Natural. 91–98 (2015).
21. Blei, D.M.: Probabilistic topic models. Communications of the ACM. 55, 77–84 (2012).