GENERACIÓN DE SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN PARA LA GESTIÓN DOCUMENTAL EN EL ÁREA DE LAS CIENCIAS DE LA COMPUTACIÓN

H. Kuna[1], **M. Rey**[1], **J. Cortes**[1], **E. Martini**[1], **L. Solonezen**[1], **R. Sueldo**[1], **G. Pautsch**[1]

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

hdkuna@gmail.com , m.rey00@gmail.com

RESUMEN

La búsqueda de información en bibliotecas digitales es una actividad en la cual los resultados a presentar al usuario, a partir de una consulta, deben responder a sus expectativas. Los Sistemas de Recuperación de Información buscan optimizar el proceso de búsqueda de contenido en la web a través de diversas herramientas, entre ellas los meta-buscadores. Los mismos amplían el espectro de cobertura en la búsqueda, a partir de la capacidad para utilizar las bases de datos de varios buscadores en simultáneo; además de poder incorporar diversos métodos para el ordenamiento de los documentos, que mejoren la relevancia de los resultados para el usuario. En este trabajo se presenta el desarrollo de un Sistema de Recuperación de Información para la búsqueda de documentos científicos en el área de las ciencias de la computación, haciendo especial hincapié en el algoritmo de ranking utilizado para ordenar el listado final de resultados.

Palabras clave: *recuperación de información, algoritmo de ranking, búsqueda web, indicadores bibliométricos.*

CONTEXTO

Está línea de investigación articula el "Programa de Investigación en Computación" de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones; el Grupo de Investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España; y el Departamento de Bibliotecología de la Universidad Nacional de Mar del Plata, Argentina.

1 INTRODUCCION

1.1 Sistemas de Recuperación de Información

Un Sistema de Recuperación de Información (SRI) se puede definir como un proceso capaz de almacenar, recuperar y mantener información (Salton & Mcgill, 1983) (Kowalski, 1997). Existen en la literatura diversas propuestas sobre la estructura básica que debiera tener un SRI, un ejemplo es la de (Baeza-Yates & Ribeiro-Neto, 1999) que lo considera a partir de la unión de cuatro elementos como son: los documentos que forman parte de la colección sobre la que se realizará la recuperación; las consultas que representan las necesidades de información por parte de los usuarios; la forma en la que la modelan las representaciones de los documentos, consultas y las relaciones presentes entre ellos; y una función de evaluación que determina para cada consulta y documento el orden que ocupará en los resultados a presentar.

En la actualidad los principales modelos de SRI que operan sobre internet son: los directorios, los buscadores y los meta-buscadores (Olivas, 2011). En el contexto de la presente investigación cobran mayor relevancia los meta-buscadores, debido a que posibilitan la utilización de bases de datos de otros buscadores replicando las consultas de los usuarios sobre los mismos y procesando los resultados obtenidos de la manera que se crea conveniente para generar un listado único. Para ello, los meta-buscadores deben contar con

algoritmos de evaluación, que son aplicados para la fusión de las listas de resultados obtenidas de cada buscador (Serrano-Guerrero, Romero, Olivas, & De la Mata, 2009).

1.2 Métricas para la Evaluación de Documentos Científicos

Dada la naturaleza del SRI a generar, los métodos para la evaluación de los resultados deben ser desarrollados en forma particular. Para la evaluación de documentos científicos se debe considerar una serie de características evaluables, como ser (Bollen, Van de Sompel, Hagberg, & Chute, 2009) (Pendlebury, 2009): el tipo de fuente de publicación, la calidad de los autores y del artículo en sí, medida a través de la cantidad de veces que haya sido citado. Para cada una de estas características existen métricas ampliamente aceptadas que pueden aplicarse. En el caso del tipo de fuente de publicación, para aquellas publicaciones realizadas en revistas existen dos índices que se utilizan para estimar su calidad: por un lado el Factor de Impacto (IF, por sus siglas en inglés) generado por la Web Of Knowledge[1] que administra el Institute for Scientific Information (ISI) (Garfield E, 2006) que es parte de la empresa Thomson Reuters; y el índice SJR, SCImago Journal Rank (Gonzalez-Pereira, Guerrero-Bote, & Moya-Anegon, 2009), producido utilizando la base de datos del buscador Scopus[2] de la editorial Elsevier y generado por el grupo de investigación SCImago de la Universidad de Extremadura, España. En ambos casos se trata de métricas que toman las citas que reciben los artículos publicados en una revista y las evalúan tanto en cantidad como en lo referente a la relevancia que tiene la producción que la realiza. Mientras que en caso de que la publicación se realice en un congreso o evento similar existe otro ranking como es el que genera en la web Computer Research & Education (CORE)[3] en donde a diversas conferencias o congresos se los ubica en uno de los cuatro niveles que tiene establecidos: A+, A, B y C sin proveer mayores detalles sobre el método utilizado para realizar el cálculo.

Para estimar la calidad de la producción de un autor existen otras métricas como pueden ser: el índice H (Hirsch, 2005) y el índice G (Egghe, 2006); lo que hacen éstas es tomar la cantidad de citas recibidas por las diferentes publicaciones del autor y la cantidad de publicaciones para calcular un valor que representa la influencia del mismo. Finalmente, para evaluar la calidad de una colección de publicaciones a través del tiempo se puede utilizar un índice como es el AR (Jin, 2007), que toma la cantidad de citas obtenidas por las mismas y las pondera utilizando ese factor en combinación con la antigüedad de cada uno de los artículos que componen la colección.

2 LINEAS DE INVESTIGACION y DESARROLLO

Existen implementaciones de SRI en la web que utilizan diferentes métodos de búsqueda, pero no existen implementaciones de herramientas de este tipo que se apliquen específicamente a bases de datos de documentos científicos en el área de las ciencias de la computación, que además incorporen técnicas de inteligencia artificial o lógica difusa para la mejora de la relevancia de los resultados a presentar al usuario.

La generación de este tipo de Sistemas de Recuperación de Información requiere del desarrollo de diversos componentes, entre los cuales se destaca el algoritmo que se utiliza para establecer un ranking entre los resultados de la consulta a realizar.

[1] Web of Knowledge: http://wokinfo.com. Accedido: 25/02/2013.

[2] Scopus: http://www.scopus.com. Accedido 25/02/2013.

[3] CORE: http://www.core.edu.au. Accedido: 25/02/2013.

3 RESULTADOS OBTENIDOS/ESPERADOS

3.1 Grado de Avance

El presente proyecto ha comenzado sobre la mitad del año 2012. En una primera etapa se analizó el estado del arte, se estudiaron los distintos modelos de SRI y los elementos que deben componerlos. Conjuntamente se seleccionaron las tecnologías que se utilizaron en el desarrollo de un prototipo de meta-buscador, priorizando aquellas que fueran basadas en la filosofía Open Source, como ser: los lenguajes HTML, PHP y SQL, junto al motor de bases de datos MySQL, utilizando como entorno para su implementación al servidor web Apache.

Se han determinado los siguientes componentes a desarrollar del meta-buscador en una primera etapa:

- Módulo para la gestión de las consultas: encargado de adaptar las consultas efectuadas por el usuario para ser utilizadas posteriormente en los buscadores integrados.
- Módulos para la búsqueda en las bases de datos (buscadores): encargado de gestionar la realización de las consultas, adaptadas previamente, sobre los buscadores incorporados. A continuación captura los resultados y los prepara para el próximo componente.
- Módulo para la gestión de los resultados: en primera instancia realiza una pre-selección de los resultados, desechando aquellos que no son relevantes en cuanto a su tipo (informe de una cita sobre el documento, por ejemplo). Posteriormente se procede a la unificación de los resultados en un único listado para la aplicación del algoritmo de ranking.
- Módulo para la mejora de la relevancia de los resultados: en éste se aplican diversas técnicas para que los resultados para que sean de mayor utilidad al usuario.

3.2 Desarrollo del Algoritmo de Ranking

El componente del meta-buscador que realiza el ordenamiento de los resultados obtenidos de las bases de datos, es de los más importantes. En este caso particular, al trabajar con artículos científicos, se requirió que el algoritmo de ranking evaluara las propiedades de los mismos a través de diversas métricas. Entre las características evaluables de los documentos científicos se seleccionaron: la fuente de publicación, la calidad de sus autores y del artículo en sí, medida por la relación entre la antigüedad del documento con la cantidad de veces que ha sido citado.

Para cada una de las propiedades seleccionadas, se determinaron las métricas a considerar para ponderar cada resultado:

- Para la fuente de publicación: se tomaron dos factores para la valoración de este punto, dependiendo si el artículo se publicó en una revista científica o en un congreso del área de conocimiento que corresponda. Para el primer caso, existen a nivel internacional dos métricas: el Factor de Impacto desarrollado por ISI y el índice SJR desarrollado por el grupo de investigación SCImago. En este caso se ha optado por utilizar al segundo ya que presenta diversas ventajas con respecto al primero, como ser (Falagas, Kouranos, Arencibia-Jorge, & Karageorgopoulos, 2008), (Leydesdorff, De Moya-Anegón, & Guerrero-Bote, 2010): es de acceso abierto; en la base de datos de Scopus contiene una mayor cantidad de revistas, incluyendo aquellas que no están escritas en inglés; no sólo hace una evaluación cuantitativa de las citas recibidas por un artículo sino que también lo hace en forma cualitativa, incorporando la calidad de la revista que genera la cita; entre otras. Para el caso de los artículos procedentes de congresos o reuniones científicas se empleó el ranking generado por la Computing Research and Education Association of Australia (CORE). El modelo de

clasificación de este ranking se transformó a un formato numérico para poder operar con él. El valor correspondiente a la fuente de publicación se obtiene mediante la fórmula 1, en caso de haber sido en una revista, y con la fórmula 2, en caso de haber sido en un congreso.

$$fuentePublicacion = \log_{10}(SJR) \quad (1)$$

$$fuentePublicacion = [A^* = 1;\ A = 0.75;\ B = 0.5;\ C = 0.25] \quad (2)$$

- Para la calidad de los autores: en este caso se usó como base el índice H propuesto por (Hirsch, 2005). El mismo representa la cantidad *x* de artículos de un autor que han recibido *x* citas como mínimo. En caso de que se trate de un artículo con más de un autor se considera la posición que ocupa en el mismo. El cálculo del valor que se utilizó se puede ver en la fórmula 3.

$$autores = \log_{10}(\sum(indiceH(autor_i))/i) \quad (3)$$

- Antigüedad y citas: en la última métrica se ponderó la calidad de la publicación teniendo en cuenta la antigüedad y la cantidad de citas obtenidas por la misma. Para ello se utilizó como base el índice AR, en la fórmula 4 se puede observar el resultado de la adaptación realizada.

$$calidadPublicacion = citasRecibidas / antigüedadPublicacion \quad (4)$$

- Finalmente se ponderan los tres factores antes expuestos a través de tres constantes que se utilizaron para dar distinto peso a cada elemento, esto se refleja en la fórmula 5. Los valores planteados para α, β y γ fueron: 0.5, 0.3 y 0.2 respectivamente.

$$valorFinal = \alpha * fuentePublicacion + \beta * autores + \gamma * calidadPublicacion \quad (5)$$

Ese valor final fue el que se utilizó para realizar el orden de los resultados antes de presentarlos al usuario.

3.3 Trabajos Previstos en la Próxima Etapa

Para el año 2013 se tiene previsto:

- Completar el desarrollo de los componentes del meta-buscador.
- Analizar distintas técnicas inteligentes a incorporar en el meta-buscador para la mejora de la relevancia de los resultados a presentar al usuario final.
- Incorporar elementos dentro del SRI que permitan la expansión de la consulta original realizada por el usuario.
- Optimizar el algoritmo de ranking implementado para la gestión de los resultados.
- Incorporar técnicas de clustering y de lógica difusa para una mejor organización de los resultados.

4 FORMACION DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del “Programa de Investigación en Computación” de la Facultad de Ciencias Exactas Quimicas y Naturales dela UNaM, con siete integrantes (todos ellos alumnos, docentes y egresados de la carrera de Licenciatura en Sistemas de Información de la facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones) de los cuales cuatro están realizando su tesis de grado y dos están realizando un doctorado. Esta línea de investigación vincula al “Programa de Investigación en Computación” del Departamento de Informática de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones, al Grupo de Iventigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España; y el Departamento de

Bibliotecología de la Universidad Nacional de Mar del Plata, Argentina.

5 BIBLIOGRAFIA

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE*, *4*(6).

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131-152.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal*, *22*(8), 2623-2628.

Garfield E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90-93.

Gonzalez-Pereira, B., Guerrero-Bote, V., & Moya-Anegon, F. (2009). The SJR indicator: A new indicator of journals' scientific prestige. *arXiv:0912.4141*.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569-16572.

Jin, B. (2007). The AR-index: complementing the h-index. *ISSI newsletter*, *3*(1), 6.

Kowalski, G. (1997). *Information Retrieval Systems: Theory and Implementation* (1st ed.). Norwell, MA, USA: Kluwer Academic Publishers.

Leydesdorff, L., De Moya-Anegón, F., & Guerrero-Bote, V. P. (2010). Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for Information Science and Technology*, *61*(2), 352–369.

Olivas, J. A. (2011). *Búsqueda Eficaz de Información en la Web*. La Plata, Buenos Aires, Argentina: Editorial de la Universidad Nacional de La Plata (EDUNLP).

Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis*, *57*(1), 1-11.

Salton, G., & Mcgill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Serrano-Guerrero, J., Romero, F. P., Olivas, J. A., & De la Mata, J. (2009). BUDI: Architecture for fuzzy search in documental repositories. *Mathware & Soft Computing*, *16*(1), 71–85.