Desarrollo de un Sistema de Recuperación de Información como herramienta para el investigador en Ciencias de la Computación

M. Rey[1], **H. Kuna**[1], **E. Martini**[1], **L. Solonezen**[1], **A. Rambo**[1], **L. Podkowa**[1]

1. Programa de Inverstigación en Computación. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Naturales Universidad Nacional de Misiones.

{m.rey00, hdkuna} @gmail.com

RESUMEN

Realizar búsquedas de información es una actividad diaria en todos los ambientes que reviste especial importancia. En áreas de investigación es una actividad que puede insumir un tiempo importante entre operaciones que van desde: definir los criterios, iniciar la búsqueda, filtrar los resultados, encontrar la información y luego tenerla a disposición de una manera clasificada y ordenada. La fase de búsqueda se inicia sobre las expectativas planteadas por el investigador y finaliza cuando el mismo encuentra finalmente el o los documentos que poseen la información de interés. Además cada área posee sus características en cuanto a la relación existente entre los términos comúnmente utilizados lo cual puede ser definitorio al momento de encontrar documentos relevantes para la consulta en cuestión.

Los meta-buscadores permiten consultar las bases de datos de varios buscadores ampliando el espectro inicial de la búsqueda y permitiendo delimitarlo en un campo específico. En este caso se presenta un meta-buscador dedicado a la recuperación de información por parte de investigadores dentro del área de las ciencias de la computación, conjuntamente al detalle de una primera fase de evaluación del mismo por parte de investigadores del área.

Palabras clave: *recuperación de información, investigación científica, ciencias de la computación.*

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); el Grupo de Investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

1 INTRODUCCION

1.1 Sistemas de Recuperación de Información

La gran cantidad de información disponible, especialmente desde el surgimiento de Internet, en formato digital trajo aparejada la inquietud en cuanto a la búsqueda de información textual en una serie de documentos. Un Sistema de Recuperación de Información (SRI) es el responsable de encontrar información relevante la cual está presente en un registro documental que debe ser almacenado, representado, analizado (manipulado) y mantenido [1], [2].

Su estructura se compone de elementos básicos, como ser: los documentos en sí, la representación de los mismos, el tratamiento de las consultas sobre estos documentos y el tratamiento de los resultados [3].

Dentro de los diferentes grupos de SRI como son los directorios, los buscadores y los meta-buscadores son estos últimos los que amplían su potencialidad dado que generan una misma consulta sobre varios motores de búsqueda en simultaneo. Cuando un meta buscador recibe una consulta de un usuario, invoca los motores de búsqueda subyacentes para recuperar información útil para el mismo. Permitiendo de esta manera el aumento de la cobertura de búsqueda de la Web y la mejora de la escalabilidad de la misma [4].

Con la finalidad de lograr que los resultados del proceso de búsqueda sean relevantes para el usuario, se dispone de varias herramientas, entre ellas: el uso de concatenadores lógicos, el uso de metadatos en un área específica y la utilización de la relación sintática o semántica que existe entre los términos de las consutas, por mencionar algunos. La manera en que se relacionan todos estos elementos y el rol que juegan son cuestiones a ser consideradas al momento de diseñar un meta-buscador siempre con el objetivo de brindar un resultado satisfactorio al usuario [5].

1.1 Un meta-buscador como herramienta para un Investigador

Al momento de realizar una búsqueda de información, el investigador debe identificar cuáles serán aquellas herramientas que permitirán obtener documentos de la mayor relevancia posible, pero también contemplar otras consideraciones sobre la misma herramienta como la accesibilidad a la mima, la velocidad de respuesta, la calidad de sus fuentes de consulta y el criterio que será aplicado para ordenar los resultados obtenidos. Con la herramienta seleccionada, se inicia la búsqeuda accediendo a diferentes fuentes de información, en el caso de los meta-buscadores estos disparan sus consultas a varias bases de datos.

En el marco de la presente investigación se ha generado un SRI que accede a fuentes de documentación científica, ya que el mismo se focaliza en un perfil orientado a la investigación en ciencias de la computación. Se ha desarrollado una arquitectura pensada para optimizar la consulta realizada, por lo que se cuenta con una propuesta modular tendiente a cubrir varias fuentes de documentos, entre ellas: Google Scholar[1], ACM Digital Library[2], IEEE Xplore[3] [6]. Y para ordenar los resultados obtenidos se definió un algoritmo de ranking particular que evalúa la calidad de cada artículo científico [7].

Para mejorar la actividad del investigador y minimizar de manera óptima la búsqueda de información existen diferentes iniciativas de generación de herramientas software [8], inclusive se reconocen algunas que trabajan de manera específica para el área de ciencias de la salud [9] y otras que tienen un perfil general en cuanto al área temática de aplicación pero que limitan su búsqueda a referencias bibliográficas [10]. Pero no se ha encontrado evidencia de meta-buscadores que operen sobre repositorios documentales del área de ciencias de la computación y que incorporen soluciones particulares para el tratamiento de documentos científicos de tal área, pudiendo ser considerados como una herramienta de ayuda para un investigador.

En el presente trabajo se relata el desarrollo de un meta-buscador y su utilización y evaluación de rendimiento por parte de investigadores, específicamente aplicado al área de ciencias de la computación y en especial desde la perspectiva de generar las búsquedas teniendo en cuenta la aplicación de conceptos relacionados por pertenecer a un mismo campo de aplicación científica.

2 LÍNEAS DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

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 tratamiento de resultados específicamente orientadas para la mejora de la relevancia de los resultados a presentar al usuario.

Considerando tal escenario, se considera que una herramienta de tales carácterísticas constituye un insumo de gran valor para un investigador del área de ciencias de la computación, siendo un instrumento que facilite la búsqueda de información

[1] scholar.google.com – Accedido 20/03/14.
[2] dl.acm.org – Accedido 18/03/14.
[3] ieeexplore.ieee.org/ - Accedido 18/03/14.

científica de calidad y que sea relevante para la consulta planteada.

3 RESULTADOS OBTENIDOS Y OBJETIVOS

3.1 Descripción del SRI desarrollado

El trabajo abordado previamente por los autores ha consistido en el planteo de lineamientos generales sobre la estructura que debería tener el SRI [6], para posteriormente comenzar el desarrollo de sus diversos componentes [7].

Conjuntamente se seleccionaron las tecnologías a utilizar en el desarrollo de la herramienta, priorizando por un lado aquellas que fueran basadas en la fisolofía Open Source y por otro aquellas que permitieran el uso del SRI adecuadamente desde la web, siendo seleccionadas: los lenguajes JSP, Javascript, xHTML, Java y SQL, junto al motor de bases de datos MySQL, utilizando como plataforma para su implementación el servidor web Tomcat.

La estructura general del meta-buscador se compone de los siguientes módulos:

- Módulo para la gestión de las consultas (MC): 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) (MB): encargado de gestionar la realización de las consultas, adaptadas previamente, sobre los buscadores incorporados. Los buscadores a los que accede actualmente el SRI son: Google Scholar[4], ACM Digital Library[5], IEEE Xplore[6].
- Módulo para la gestión de los resultados (MR): encargado de procesar los resultados para su posterior presentación al usuario final. El módulo posee como componente primordial al algoritmo de ranking, específicamente desarrollado para el SRI [7], cuya función es clasificar cada documento recuperado a partir de metricas que ponderan varios factores que hacen a la calidad de la publicación entre ellos: la calidad de su fuente de publicación (revista o evento científico donde se ha publicado), la calidad de sus autores y su calidad en sí a través del tiempo de antigüedad que tenga.

El proceso de la búsqueda, ver figura 1, se compone de los siguientes pasos:

1. El usuario ingresa la consulta.
2. El MC captura el texto de la consulta y la adapta según los requerimientos determinados por cada fuente de datos a consutlar.
3. El MB captura cada consulta adaptada y la efectúa sobre la fuente de datos correspondiente.
4. El MR caputa los listados de documentos generados por cada buscador y los filtra, eliminando aquellos que no se consideran relevantes en cuanto a su tipo (informe de una cita sobre el documento, por ejemplo).
5. Se unifican los listados de resultados obtenidos de las fuentes eliminando aquellos documentos duplicados.
6. Se aplica el algoritmo de ranking, asignando el valor correspondiente a cada documento del listado.
7. Se ordena el listado de resultados en base al valor obtenido por cada documento al aplicarse el algoritmo de ranking.
8. Se genera el listado formateado para su presentación al usuario final, el cual visualiza los siguientes datos de cada documento: título, fuente de publicación, listado de autores, descripción general, cantidad de citas, link de acceso al documento, descripción del valor asignado por el algoritmo de ranking.

[4] scholar.google.com – Accedido 18/03/14.
[5] dl.acm.org – Accedido 18/03/14.
[6] ieeexplore.ieee.org/ - Accedido 18/03/14.

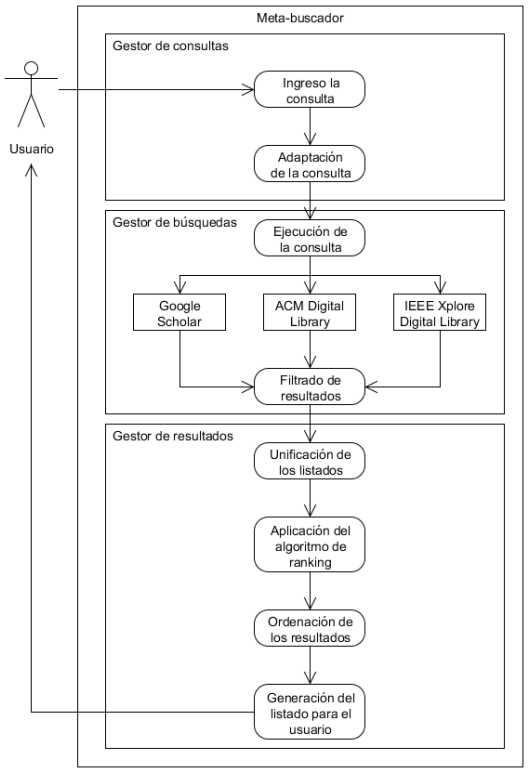

Figura 1 – Proceso de búsqueda

3.2 Experiencia de uso

Durante el segundo semestre del año 2013 el meta-buscador ha sido utilizado en modo de beta cerrada por los investigadores que forman parte del Programa de Investigación en Computación de la FCEQyN, aproximadamente 20 personas. De esta etapa de testeo se han obtenido datos de gran valor para la corrección de diversos bugs por parte del equipo de desarrollo. Además de diversas recomendaciones basadas en el uso de los investigadores que han generado modificaciones en el software, algunas ya implementadas, otras por implementar, como ser:

- Inclusión de detalles sobre los valores de cada propiedad evaluada por el algoritmo de ranking para cada documento.
- Posibilitar la realización de un feedback por parte del usuario sobre cada resultado, a modo de voto positivo o negativo, y que eso sea utilizado para caracterizar determinados resultados modificando el ranking a establecer.
- Incluir más fuentes de consulta (buscadores).
- Permitir personalizar la ponderación que se hace de cada propiedad evaluada (calidad de fuente de publicación, calidad de los autores, calidad del documento).
- Gestionar perfiles de usuario de modo tal de poder guardar búsquedas así como también poder adaptar el funcionamiento en base a preferencias del usuario.
- Desarrollar métodos para exportar meta-datos de los resultados para ser utilizados en algún gestor de bibliografía, por ejemplo: Zotero, Mendeley, entre otros.

Con el objetivo de ampliar la fase de testeo, el meta-buscador estará disponible desde el mes de abril del presente año para ser utilizado por los alumnos de la carrera Licenciatura en Sistemas de Información de la FCEQyN en la cátedra Tesis de Grado, con el objetivo de constituir una herramienta para agilizar el relevamiento bibliográfico necesario para el planteo y posterior desarrollo de sus trabajos de tesis. Con esta actividad se espera contar con mayores datos y otro nivel de feedback que permita la mejora de la calidad integral del SRI desarrollado.

3.3 Trabajos Previstos en la Próxima Etapa

Para el año 2014 se tiene previsto:

- Agregrar más fuentes de datos al SRI, incluyendo alguna del orden nacional, como podría ser el SEDICI[7] de la UNLP.
- Incorporar más métricas al algoritmo de ranking para el análisis de documentos científicos.
- Continuar con el testeo de la herramienta por parte de usuarios para depurarla lo máximo posible.

[7] sedici.unlp.edu.ar – Accedido 18/03/14

- Incorporar técnicas de lógica difusa y/o de inteligencia artificial para la mejora de la experiencia de uso del SRI.
- Incorporar métodos que permitan contar con información relativa a la reputación de cada autor en relación al área temática en la que se haya desarrollado la consulta.

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 siete integrantes (todos ellos alumnos, docentes y egresados de la carrera de Licenciatura en Sistemas de Información de la FCEQyN – UNaM) de los cuales tres están realizando su tesis de grado, uno se encuentra realizando una maestría y dos están realizando un doctorado. Esta línea de investigación vincula al "Programa de Investigación en Computación" del Departamento de Informática de la FCEQyN de la UNaM, al Grupo de Iventigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

5 BIBLIOGRAFIA

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

[2] G. Salton y M. Mcgill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[3] R. Baeza-Yates y B. Ribeiro-Neto, *Modern information retrieval*, vol. 463. ACM press New York., 1999.

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

[5] J. Serrano-Guerrero, F. P. Romero, J. A. Olivas, y J. de la Mata, «BUDI: Architecture for fuzzy search in documental repositories», *Mathware & Soft Computing*, vol. 16, n.º 1, pp. 71–85, 2009.

[6] H. Kuna, M. Rey, E. Martini, L. Solonezen, R. Sueldo, y J. G. A. Pautsch, «Generación de sistemas de recuperación de información para la gestión documental en el área de las Ciencias de la Computación», presentado en XV Workshop de Investigadores en Ciencias de la Computación, 2013.

[7] H. Kuna, M. Rey, E. Martini, L. Solonezen, y R. Sueldo, «Generación de un algoritmo de ranking para documentos científicos del área de las ciencias de la computación», presentado en XVIII Congreso Argentino de Ciencias de la Computación, 2013.

[8] J. L. Orihuela, «Guía de recursos en Internet para Investigadores», *eCuaderno. Recuperat*, vol. 30, n.º 10, 2009.

[9] S. Sastre-Suárez y E. Pastor-Ramon, «Evaluación de metabuscadores gratuitos especializados en ciencias de la salud», *El profesional de la información*, vol. 20, n.º 6, pp. 639–644, 2011.

[10] S. Jung, J. L. Herlocker, J. Webster, M. Mellinger, y J. Frumkin, «LibraryFind: System design and usability testing of academic metasearch system», *J. Am. Soc. Inf. Sci.*, vol. 59, n.º 3, pp. 375-389, feb. 2008.