AVANCES EN LA CONSTRUCCIÓN DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN PARA INFORMACIÓN CIENTÍFICA EN CIENCIAS DE LA COMPUTACIÓN

H. Kuna[1], **M. Rey**[1], **E. Martini**[1], **A. Canteros**[1], **A. Cantero**[1], **A. Rambo**[1], **C. Biale**[1], **E. Zamudio**[1]

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

{hdkuna}@gmail.com

RESUMEN

La búsqueda de información en internet es una actividad en donde la relevancia de los resultados cobra una importancia cada vez mayor a medida que la cantidad, variedad y disponibilidad de los datos aumentan. En los últimos años han aparecido herramientas de recuperación de información para diferentes ámbitos, constituyendo aplicaciones para el procesamiento de datos de un área en particular, integrando métodos de evaluación de la relevancia adaptados a su contexto, inclusive para el área de Ciencias de la Computación.

La recuperación de información científica se considera un área en la cual el desarrollo de herramientas de este tipo se podría considerar de gran utilidad en la medida que provean de información confiable y de diversas fuentes a un usuario-investigador. La posibilidad de que los resultados a mostrar al usuario integren datos de publicaciones, autores y medios de publicación, enriquecidos a través de meta-datos relacionados al impacto generado con sus aportes al tema de búsqueda; además de establecer el orden de presentación a partir de métodos de evaluación con un enfoque multi-dimensional e integral con base en ese mismo impacto, se define como el objetivo principal de la presente línea de investigación.

Palabras clave: *información científica, meta-buscador, producción científica, evaluación de investigación, métricas.*

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 Sistemas de Recuperación de Información e Información Científica

Los Sistemas de Recuperación de Información (SRI) se reconocen como herramientas de gran utilidad para la gestión de información en internet [1], [2]. Actividades como el almacenamiento, representación, análisis y mantenimiento de grandes volúmenes de datos son llevadas a cabo por herramientas de este tipo en diferentes contextos de aplicación [3]–[5].

En los últimos años, se han generado y publicado numerosos indicadores definidos para la evaluación del impacto generado por publicaciones científicas [6]. Esta situación ha conllevado al desarrollo de herramientas especializadas, con similares funciones a las de los SRI, en la recopilación de información y aplicación de uno o más de tales indicadores para la evaluación de un documento en particular, además de sus autores y los medios seleccionados para su divulgación [7]–[9].

Sin embargo, solo una fracción de los indicadores y las herramientas que los utilizan lo hacen en forma complementaria, es decir que no es habitual poder ejecutar búsquedas de contenido de utilidad para un

investigador científico haciendo uso de la información obtenida a partir del cálculo de indicadores sobre el objetivo de la búsqueda [10]–[12]. Por otra parte, el tipo de evaluación predominante en las soluciones disponibles se basa en el análisis unidimensional de la producción científica, sin considerar la interacción entre indicadores para la evaluación de más de una característica de una publicación en un momento determinado [13], [14].

1.2 Antecedentes

En los últimos años, se han producido avances en el desarrollo de diversos componentes de un SRI de propósito específico para la recuperación de publicaciones científicas del área de Ciencias de la Computación (CC) [15]. Se ha logrado desarrollar un SRI funcional con módulos que permiten su aplicación, entre los que se destacan el módulo de expansión de consultas [16], [17] y el de aplicación de un algoritmo de ranking para el ordenamiento de los resultados a presentar al usuario [18], [19], inclusive se han planteado diferencias en cuanto a los métodos de estimación del impacto de la producción científica [20].

El mencionado algoritmo de ranking ha sido diseñado considerando las principales limitaciones de los modelos de evaluación de la actualidad [8], [21], [22] y, a través de su aplicación, se evalúa a una publicación científica a partir de los valores de diversas métricas, provenientes de fuentes diferentes y que evalúan a cada documento desde tres propiedades o dimensiones: la calidad de la fuente de su publicación, la calidad de sus autores y la calidad del documento en sí mismo [19].

1.3 Inconvenientes detectados en el SRI desarrollado

El SRI desarrollado no se encuentra exento de limitaciones o inconvenientes, entre los que se pueden mencionar:

- Los elementos a recuperar desde las diferentes fuentes a las que accede el SRI se limitan a publicaciones científicas.
- Las bases de datos sobre las cuales se replican las consultas se corresponden únicamente a buscadores de librerías digitales.
- Los resultados recuperados de cada fuente de datos no son re-analizados una vez que se han ordenado según los criterios del algoritmo de ranking y se han presentado al usuario. De esta manera se descartan sus meta-datos al finalizar el proceso de búsqueda.
- No se mantienen datos correspondientes al feedback del usuario al completar la operación con el SRI, de modo que no se puede brindar una experiencia personalizada en base a sus preferencias de búsqueda.
- No se cuenta con datos históricos correspondientes a búsquedas ejecutadas, lo que dificulta la aplicación y/o generación de nuevos indicadores para la evaluación de resultados en base a un historial de clasificaciones realizadas.

A partir de estos problemas se han planteado cambios en la arquitectura y en el funcionamiento general del SRI, principalmente con el objetivo de recuperar una mayor cantidad de información, incluyendo no solo documentos científicos sino también datos de autores y de fuentes de publicación relevantes para la consulta ingresada por el usuario. De esta manera, se pretende desarrollar un SRI para la recuperación de información científica en un sentido más general para las CC.

2 LÍNEAS DE INVESTIGACION, DESARROLLO E INNOVACIÓN

La búsqueda de información en internet es una actividad habitual para un investigador científico. El uso de buscadores y herramientas especializadas para la recuperación de contenido podría

considerarse de suma utilidad dados los volúmenes de información disponibles en la actualidad. Las CC son consideradas como un área en constante expansión e innovación, por lo que la premisa anterior cobra mayor relevancia.
Es en este contexto que se reconoce como objetivo de interés para la presente línea de investigación, el desarrollo de un SRI que facilite a un usuario investigador la ejecución de búsquedas en forma transparente hacia diferentes fuentes, recuperando no solo datos de documentos, sino también de autores y medios de publicación relacionados. Por otra parte, el mencionado SRI deberá contar con componentes que permitan establecer el orden de los resultados a partir de un análisis multi-dimensional de cada elemento recuperado, integrando toda la información disponible; y previendo el almacenamiento de diversos datos de cada operación con el objetivo de mejorar la experiencia de un usuario en base a resultados calificados como relevantes en búsquedas previas.

3 RESULTADOS Y OBJETIVOS

3.1 Objetivos de la investigación

Dados los inconvenientes detectados, en la presente línea de investigación se propone la refactorización del SRI presentado anteriormente, integrando al mismo la recuperación de información científica de autores y medios de publicación, además de los documentos que se obtienen en la actualidad. Actualizando el método de evaluación de cada tipo de resultado con base en el modelo de evaluación planteado previamente, al contar con mayor cantidad de información.
Además se propone la modificación de la arquitectura planteada para el SRI a fin de incorporar un middleware para la gestión de la recuperación de datos, que se base en las bases de datos a las que accede el SRI, y no en el tipo de resultado a recuperar. De esta manera, se facilitaría la integración de fuentes que no se correspondan únicamente a bibliotecas digitales, pudiendo acceder a una fuente específica para recuperar un tipo de resultado específico, como podría ser de autores.
El cumplimiento de ambos objetivos permitiría el procesado de datos de documentos, autores y fuentes de publicación tanto para su presentación al usuario, ordenados según el algoritmo de ranking, como así también su post-procesamiento para contar con un mayor volumen de datos para mejorar la calidad de la evaluación de futuras consultas. De igual manera, se podrían persistir los resultados y su evaluación a modo de historial operativo del SRI, pudiendo agregarse datos correspondientes al feedback del usuario.
Con estos cambios, se considera que el SRI constituiría una herramienta de mayor calidad para la presentación de información a un investigador científico.

3.2 Actividades en curso

En el marco de la presente línea de investigación se encuentran en ejecución las siguientes actividades:

- Refactorización del código fuente del SRI para la incorporación en el mismo de resultados del tipo autor y fuente de publicación.
- Desarrollo del middleware para la gestión de la recuperación de datos, definiendo elementos específicos para cada fuente, variando el tipo de resultado a recuperar.
- Desarrollo del componente encargado de la persistencia de los resultados de las búsquedas ejecutadas para su posterior procesamiento.
- Adaptación de los componentes de la interfaz gráfica del SRI para la presentación de los tipos de resultados integrados en esta iteración.
- Diseño e implementación de los perfiles de datos para la persistencia de documentos científicos, autores y fuentes de publicación. Para ello se

encuentra en curso un análisis de las diferentes fuentes de información a fin de extraer los atributos comunes a las mismas y unificando los formatos a utilizar para su implementación.

3.3 Objetivos a corto plazo

A corto plazo los objetivos de la presente investigación son:

- Adaptar los métodos de aplicación del algoritmo de ranking del SRI para la evaluación de los diferentes tipos de resultados. Incorporando en estas operaciones las métricas que pudieran utilizarse a partir de los meta-datos obtenidos de cada resultado a partir de la ejecución de las búsquedas.
- Unificar una taxonomía para las sub-áreas incluidas dentro de las CC a fin de proponer al usuario la selección de una de ellas para la orientación de la búsqueda hacia resultados de mayor relevancia dentro de la misma.
- Diseñar y desarrollar un componente que permita al SRI la gestión de perfiles de usuario, incluyendo un feedback sobre los resultados de las búsquedas que ejecute y parametrizaciones específicas orientadas a sus áreas de mayor interés.

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 (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, cuatro se encuentran realizando una maestría y uno se encuentra realizando un doctorado. La línea y el equipo de investigación se vinculan con el Grupo de Investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

5 BIBLIOGRAFIA

[1] 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.

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

[3] «Semantic Scholar — Allen Institute for Artificial Intelligence». [En línea]. Disponible en: http://allenai.org/semantic-scholar.html. [Accedido: 03-mar-2016].

[4] 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.

[5] J. de la Mata, J. A. Olivas, y J. Serrano-Guerrero, «Overview of an Agent Based Search Engine Architecture», en *Proc. Of the Int. Conf. On Artificial Intelligence IC-AI'04*, Las Vegas, USA, 2004, vol. I, pp. 62-67.

[6] J. Bollen, H. Van de Sompel, A. Hagberg, y R. Chute, «A Principal Component Analysis of 39 Scientific Impact Measures», *PLoS ONE*, vol. 4, n.º 6, p. e6022, jun. 2009.

[7] A. N. Guz y J. J. Rushchitsky, «Scopus: A system for the evaluation of scientific journals», *Int Appl Mech*, vol. 45, n.º 4, pp. 351-362, abr. 2009.

[8] P. Jacso, «As we may search-Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases», *CURRENT SCIENCE-BANGALORE-*, vol. 89, n.º 9, p. 1537, 2005.

[9] L. I. Meho y K. Yang, «A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar», arXiv e-print cs/0612132, dic. 2006.

[10] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, y G. Pappas, «Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses», *FASEB J*, vol. 22, n.º 2, pp. 338-342, ene. 2008.

[11] L. Leydesdorff, «How are new citation-based journal indicators adding to the bibliometric toolbox?», *J. Am. Soc. Inf. Sci.*, vol. 60, n.º 7, pp. 1327-1336, jul. 2009.

[12] Y. Ding, G. G. Chowdhury, S. Foo, y W. Qian, «Bibliometric information retrieval system (BIRS): A web search interface utilizing bibliometric research results», *J. Am. Soc. Inf. Sci.*, vol. 51, n.º 13, pp. 1190-1204, ene. 2000.

[13] R. Monastersky, «The number that's devouring science», *Chronicle of Higher Education*, vol. 52, n.º 8, p. 14, 2005.

[14] J. Ewing, «Measuring journals», *NOTICES-AMERICAN MATHEMATICAL SOCIETY*, vol. 53, n.º 9, p. 1049, 2006.

[15] H. Kuna, M. Rey, E. Martini, L. Solonezen, y L. Podkowa, «Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación», *Revista Latinoamericana de Ingeniería de Software*, vol. 2, n.º 2, pp. 107-114, 2013.

[16] M. Rey, H. D. Kuna, E. Martini, L. Podkowa, J. G. A. Pautsch, y E. Zamudio, «Generación de un método de expansión de consultas basado en ontologías para un sistema de recuperación de información», presentado en XX Congreso Argentino de Ciencias de la Computación (Buenos Aires, 2014), 2014.

[17] H. D. Kuna, M. Rey, L. Podkowa, E. Martini, y L. Solonezen, «Expansión de consultas basada en ontologías para un sistema de recuperación de información», presentado en XVI Workshop de Investigadores en Ciencias de la Computación, 2014.

[18] H. Kuna, M. Rey, J. Cortes, E. Martini, y L. Solonezen, «Generating a Ranking Algorithm for Scientific Documents in the Computing Science Area», en *XIX Argentine Congress of Computer Science Selected Papers*, La Plata, Buenos Aires, Argentina: EDULP, 2014, pp. 185-195.

[19] H. Kuna, E. Martini, y M. Rey, «Evolution of a Ranking Algorithm for Scientific Documents in the Computer Science Area», en *XX Argentine Congress of Computer Science Selected Papers*, La Plata, Buenos Aires, Argentina: EDULP, 2015, pp. 145-155.

[20] «Modelos de evaluación de producción científica para el área de Ciencias de la Computación». [En línea]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/45837. [Accedido: 04-mar-2016].

[21] R. Adler, J. Ewing, y P. Taylor, «Joint committee on quantitative assessment of research: citation statistics», *Australian Mathematical Society Gazette*, vol. 35, n.º 3, pp. 166–88, 2008.

[22] J. Bar-Ilan, «Which h-index? — A comparison of WoS, Scopus and Google Scholar», *Scientometrics*, vol. 74, n.º 2, pp. 257-271, nov. 2007.