Generación de un Algoritmo de Ranking para Documentos Científicos del 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]

[1]Depto. de Informática, Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones.
{hdkuna, m.rey00}@gmail.com

Resumen. La generación de un algoritmo de ranking para el ordenamiento de documentos científicos pertenecientes al área de ciencias de la computación es un requerimiento fundamental para el desarrollo de Sistemas de Recuperación de Información que sean capaces de operar sobre tal tipo de elementos. Estos sistemas 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 algoritmo de ranking para ordenar el listado de resultados que retorne un Sistema de Recuperación de Información para la búsqueda de documentos científicos en el área de las ciencias de la computación.

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

1 Introducción

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 [1], [2]. Existen en la literatura diversas propuestas sobre la estructura básica que debiera tener un SRI, un ejemplo es la que lo considera a partir de la unión de cuatro elementos como son [3]:

- 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.
- La 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 [4]. Considerando tal clasificación se puede afirmar que existen diversas implementaciones de SRI en la web que utilizan diferentes métodos de búsqueda sobre contextos generales o particulares, como se puede observar en distintas publicaciones [5], [6].

1.2 SRI para Documentos Científicos del Área de Ciencias de la Computación

No se ha encontrado evidencia de la existencia de implementaciones de SRI que sean aplicadas específicamente a bases de datos de documentos científicos pertenecientes al área de ciencias de la computación, que además implementen diversos métodos para la mejora del listado de resultados a presentar al usuario en base a la relevancia que puedan tener los mismos con respecto a la consulta efectuada.

En el contexto del presente trabajo cobran una mayor notoriedad los meta-buscadores, debido a que posibilitan la utilización de bases de datos de otros buscadores, replicando las consultas de los usuarios sobre cada una de ellas y, posteriormente, procesar los resultados obtenidos de la manera que se crea conveniente para generar un único listado de resultados a presentar al usuario.

La generación de un SRI que opere sobre documentos científicos del área de ciencias de la computación, requiere directamente el desarrollo de diversos componentes, entre los cuales se destaca el algoritmo a utilizar para la evaluación de cada resultado obtenido de las búsquedas con el objetivo de fusionar y ordenar el listado final de resultados [5].

1.3 Métricas para la Evaluación de Documentos Científicos

Dada la naturaleza del SRI planteado y el algoritmo de ranking a generar para el mismo, 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 [7], [8]:

- El tipo de fuente de publicación, distinguiendo si el mismo se publica en una revista científica o en un congreso científico o evento similar.
- La calidad de los autores, considerando en este caso la cantidad de publicaciones que ha realizado el mismo y la relevancia de las mismas, medida a través de la cantidad de citas que hubieran generado.
- La calidad del artículo en sí, en este caso, medida a través de la cantidad de veces que haya sido citado a lo largo del tiempo.

Para cada una de estas características existen métricas ampliamente aceptadas que pueden aplicarse, algunas de ellas pueden observarse con claridad en la tabla 1.

Table 1. Métricas relevadas para la evaluación de artículos científicos

Característica a evaluar	Métricas disponibles		Origen de la métrica
Tipo de fuente de publicación	Publicación en Revista Científica	Factor de Impacto (IF) [9]	Web of Knowledge[1] – Institute for Scientific Information (ISI)
		SCImago Journal Rank (SJR) [10]	Scopus[2] – Grupo SCImago, Univ. De Extremadura, España
	Publicación en Congreso Científico	Ranking CORE [11]	Computer Research & Education of Australia[3]
Calidad de los autores	Índice H [12]		Artículo científico
	Índice G [13]		Artículo científico
Calidad del artículo	Índice AR [14]		Artículo científico
	Cantidad de citas		-

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) [9]; y el índice SJR, SCImago Journal Rank [10]. 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 un ranking como es el que genera en la web de la Computer Research & Education of Australia (CORE) [11] en donde a diversas conferencias o congresos se los ubica en uno de los cuatro niveles que tiene establecidos: A*, A, B y C, listado que se establece en la mencionada web que será reformado y actualizado a la brevedad.

Para estimar la calidad de la producción de un autor se dispone de métricas como pueden ser: el índice H [12] y el índice G [13]; 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.

Para evaluar la calidad de una colección de publicaciones a través del tiempo se puede utilizar un índice como es el AR [14], que toma la antigüedad de las mismas y las pondera utilizando ese factor en combinación con la cantidad de citas obtenidas por cada uno de los artículos que componen la colección; siendo este último factor

[1] www.wokinfo.com – Accedido: 16/07/13
[2] www.scopus.com – Accedido: 16/07/13
[3] www.core.edu.au – Accedido: 16/07/13

otra de las métricas disponibles para evaluar la calidad de un documento científico particular.

El objetivo del presente trabajo es el de desarrollar un algoritmo de ranking para documentos científicos específico para el área de ciencias de la computación, de manera de generar un componente que pudiera ser incluido en un SRI, puntualmente un meta-buscador, cuyo propósito sea la recuperación de contenidos de esta área de conocimiento en la web. Para tal tarea se pretende incluir en el algoritmo diversas métricas que permiten la evaluación de un artículo científico desde diversos aspectos como son: el tipo de fuente de publicación, la calidad de los autores que lo suscriben y la calidad del artículo en sí.

2 Materiales y Métodos

2.1 Estructura del SRI

Dado que el algoritmo de ranking se debe incorporar a un SRI, se debió considerar cuál sería la estructura del mismo, estando la misma conformada por módulos para realizar las siguientes operaciones, un esquema de los mismos puede observarse en la figura 1:

- Tratamiento de las consultas introducidas por el usuario para ser utilizadas sobre las fuentes de datos integradas;
- Realización de la búsqueda replicando la consulta del usuario en las diferentes fuentes de datos;
- Captura, selección y unificación de los resultados obtenidos de las diferentes fuentes, siendo éste el módulo en el que el algoritmo a desarrollar se incorporaría;
- Mejoramiento de los resultados a presentar al usuario a través de diversas técnicas inteligentes.

Una cuestión más a considerar fueron las fuentes de datos a las que accedería el SRI para obtener los documentos científicos que coincidieran con los requerimientos del usuario, considerando que las mismas deberían permitir la obtención, directa o indirectamente, de los valores correspondientes a las métricas que se determinara utilizar en el algoritmo de evaluación. Los buscadores seleccionados inicialmente fueron: el buscador académico de Google, Google Scholar[4], y el buscador Scopus de la editorial Elsevier. Esta selección se debió a que ambas herramientas se encuentran en constante actualización tanto en sus componentes como en los documentos a los que acceden y cumplen con el requisito de disponer de diversas métricas que se pueden utilizar al evaluar a las publicaciones que recuperan, constituyendo alternativas de gran calidad para dar cumplimiento a los objetivos planteados [15], [16].

Con la definición del contexto en el cual operaría el algoritmo de ranking se prosiguió con el diseño, desarrollo y posterior validación del mismo.

[4] www.scholar.google.com – Accedido: 16/07/13

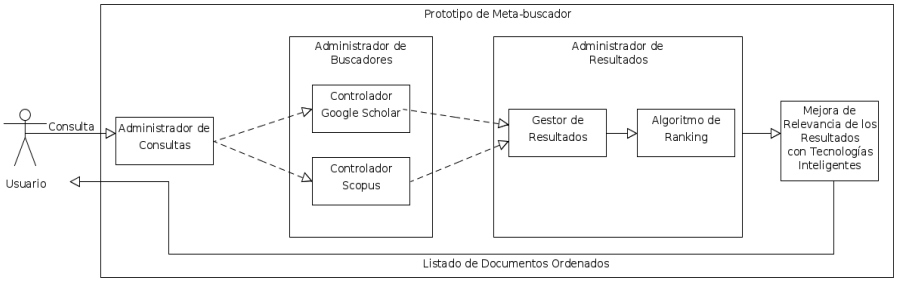

Figura 1. Componentes del meta-buscador

2.2 Diseño del Algoritmo

En una primera instancia se seleccionaron las métricas a utilizar dentro del algoritmo de ranking, priorizando diversos aspectos en cada una. Considerando nuevamente las características evaluables de los documentos científicos, la fuente de publicación, la calidad de sus autores y la calidad del artículo en sí; se determinaron las métricas a considerar para ponderar cada resultado:

- Para el tipo de fuente de publicación: se consideraron 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 correspondiente. Para el primer caso, se ha optado por utilizar el índice SJR [10] desarrollado por el grupo de investigación SCImago; esta selección se debió a las ventajas que presenta con respecto al IF de ISI [9], como ser [17], [18]: 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). En resumen:

  – *Si* (tipo_publicación = revista_científica) *Entonces* usar_SJR
  – *Si* (tipo_publicación = congreso_científico) *Entonces* usar_CORE

- Para la calidad de los autores: en este caso se optó por utilizar el índice H [12], aun considerando algunas críticas que puedan realizarse sobre el mismo, ya que es ampliamente aceptado y utilizado para la evaluación de la producción científica de un determinado autor [7], [8]. El índice concretamente representa la cantidad *X* de artículos de un autor que han recibido *X* citas como mínimo.
- Para la calidad del artículo: en este caso se determinó considerar ambas métricas relevadas previamente, el índice AR [14] y la cantidad de citas recibidas por el artículo, remarcando la necesidad de la primera de ser adaptada para trabajar sobre un único documento en vez de una colección como se plantea originalmente.

2.3 Desarrollo del Algoritmo de Ranking

Una vez seleccionadas las métricas que conformarían el algoritmo se procedió con el desarrollo concreto del mismo. Para tal actividad se definieron inicialmente las fórmulas a través de las cuales se calcularían los valores correspondientes para cada propiedad de los documentos:

- Para el factor correspondiente a la propiedad del tipo de fuente de publicación: en caso de que se trate de una publicación en una revista científica se calcula el logaritmo en base 10 del valor del índice SJR de la revista, esto con la finalidad de homogeneizar los valores de este factor con respecto al resto de los componentes del algoritmo, ya que el rango de valores presentes en el índice es mayor a la decena en un gran número de revistas. Mientras que para el caso de que la publicación se realice en un congreso o evento científico se debió adaptar el modelo de clasificación que otorga el ranking CORE, transformando a un formato numérico la clasificación del congreso para poder operar con él. El valor correspondiente al factor de 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.

$$\text{fuentePublicacion} = \log10(\text{SJR}) \; . \tag{1}$$

$$\text{fuentePublicacion} = [A^* = 1;\ A = 0.75;\ B = 0.5;\ C = 0.25] \; . \tag{2}$$

- Para el factor correspondiente a la calidad de los autores: se considera el índice H del autor del artículo en evaluación. En caso de que se trate de un artículo con más de un autor se pondera el valor del índice con respecto a la posición que ocupa en el listado de autores del documento. Además se vuelve a utilizar el logaritmo en base 10 para el valor resultante de la sumatoria ponderada de los valores del índice H de los autores del artículo. El cálculo del factor se puede ver en la fórmula 3.

$$\text{autores} = \log10(\sum(\text{indiceH}(\text{autor}_i))/i) \; . \tag{3}$$

- Para el factor correspondiente a la calidad del documento en evaluación: en este caso, dado en enfoque combinado al emplear como base al índice AR y anexar al mismo la cantidad de citas recibidas por la publicación, se determinó que la el factor ponderaría la calidad de la misma a través del cociente resultante entre ambos elementos: la antigüedad y la cantidad de citas. Dando origen a la fórmula 4 en la que se puede observar el resultado de la adaptación realizada.

$$\text{calidadPublicacion} = \text{citasRecibidas} \,/\, \text{antigüedadPublicacion} \; . \tag{4}$$

Una vez determinados los componentes correspondientes a cada una de las características a evaluar de un documento científico, se determinó que se adicionaría al cálculo final del algoritmo un factor de ajuste, el cual tendría la función de permitir que uno de los factores tuviera más importancia que los otros. El cálculo de los factores asociados a cada propiedad multiplicados por los factores de ajuste resulta en el valor final que se utiliza para realizar el orden de los resultados antes de presentarlos al usuario.

- Con la inclusión de los factores de ajuste asociados a los componentes correspondientes a las propiedades evaluadas se da forma al valor final correspondiente a cada documento en evaluación por parte del algoritmo de ranking, esto se refleja en la fórmula 5. Los valores establecidos, en forma conjunta con los expertos en la temática, para los factores de ajuste fueron: 0.5, 0.3 y 0.2 respectivamente.

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

3 Experimentación

3.1 Desarrollo del Prototipo de SRI para la Experimentación

Con el objetivo de validar el correcto funcionamiento del algoritmo de ranking propuesto se ha incluido al mismo dentro de un prototipo de meta-buscador, el cual constituye la implementación parcial del SRI descripto en las secciones anteriores.

El prototipo mencionado fue desarrollado priorizando el uso de tecnologías 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.

El proceso de implementación del prototipo se descompuso en los siguientes pasos:

1. Desarrollo de los métodos para acceder, consultar y extraer los resultados de los buscadores Google Scholar y Scopus.
2. Implementación del algoritmo de ranking con el acceso a las fuentes de datos que almacenan los valores de las diferentes métricas involucradas.
3. Desarrollo de los componentes visuales del prototipo, es decir, de las interfaces para captura de las consultas del usuario y la correspondiente a la presentación del listado de resultados unificado.
4. Integración de todos los componentes en un único producto software.

3.2 Validación del Algoritmo Desarrollado

El proceso de validación constó de dos etapas que evaluaron los resultados desde dos perspectivas, inicialmente se ha considerado a los resultados desde la óptica de un experto en bibliotecología y posteriormente se ha evaluado al algoritmo de ranking

como componente del prototipo de SRI encargado de la mejora de la relevancia de los resultados a presentar al usuario final, para lo cual se ha contado con la colaboración de tres expertos en la temática de desarrollo de métodos de recuperación de información a partir de la web.

Para la primera instancia de validación, cuyo detalle puede observarse en la tabla 2, se han realizado diversas consultas, utilizando el prototipo de meta-buscador descripto en la sección anterior, operando sobre un número reducido de documentos, y exportando los resultados de los cálculos correspondientes al algoritmo de ranking a un archivo externo al SRI. Considerando tales datos el experto en el área de bibliotecología, ha determinado que las métricas empleadas han sido calculadas en forma correcta, generando un valor numérico que permite establecer un orden entre los documentos, que forman parte del listado resultante de la búsqueda, en base a la relevancia de los mismos evaluada a partir de las propiedades seleccionadas.

Table 2. Resultados de la primera instancia de validación

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por el experto
data mining AND outliers	20 (10 Google Scholar + 10 Scopus)	74%
fuzzy sets AND clustering	20 (10 Google Scholar + 10 Scopus)	87%
alphanumeric data AND outliers	20 (10 Google Scholar + 10 Scopus)	81%
scientific production AND metrics	20 (10 Google Scholar + 10 Scopus)	77%
text mining AND ontologies	20 (10 Google Scholar + 10 Scopus)	96%

Posteriormente se procedió con la valoración del algoritmo de ranking como componente del prototipo de SRI por parte de los expertos en la temática, el detalle de la experimentación se observa en la tabla 3, en la que se incrementaron la cantidad de consultas y la cantidad de resultados a obtener. En este caso el modo de trabajo de los expertos consistió en la evaluación de la calidad de los resultados con respecto a los diversos requerimientos del usuario. Como resultado, se ha determinado que el componente de gestión de los resultados, a través del algoritmo de ranking desarrollado, cumple satisfactoriamente con el objetivo de evaluar la calidad de los resultados para la generación del listado final a presentar al usuario, logrando que el mismo presente en sus primeros lugares a aquellos documentos científicos de mayor calidad.

Table 3. Resultados de la segunda instancia de validación

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por los expertos
data mining AND outliers	100 (50 Google Scholar + 50 Scopus)	72%
fuzzy sets AND clustering	100 (50 Google Scholar + 50 Scopus)	93%
alphanumeric data AND outliers	100 (50 Google Scholar + 50 Scopus)	84%
scientific production AND metrics	100 (50 Google Scholar + 50 Scopus)	90%
text mining AND ontologies	100 (50 Google Scholar + 50 Scopus)	94%
data mining AND systems audit	100 (50 Google Scholar + 50 Scopus)	69%
scientific articles AND ranking algorithms	100 (50 Google Scholar + 50 Scopus)	83%
fuzzy controllers AND robotics	100 (50 Google Scholar + 50 Scopus)	79%
fuzzy sets AND document processing	100 (50 Google Scholar + 50 Scopus)	75%
web agents AND document analysis	100 (50 Google Scholar + 50 Scopus)	86%

4 Conclusiones y Trabajos Futuros

Con el presente trabajo se ha conseguido desarrollar y validar un algoritmo de ranking específico para la evaluación de documentos científicos pertenecientes al área de ciencias de la computación. Para el mismo se han tomado en consideración distintos indicadores bibliométricos, con la finalidad de obtener valores para la evaluación de las distintas propiedades a fin de determinar la calidad de documentos científicos del área de ciencias de la computación. Además se ha incluido al mismo dentro de un prototipo de SRI, concretamente un meta-buscador, cuyo campo de aplicación son los documentos antes mencionados, constituyendo un avance significativo en lo que respecta a los objetivos del presente trabajo.

Como trabajos a futuro se pueden mencionar: evaluar la incorporación de otros indicadores bibliométricos que puedan ser de utilidad para el algoritmo de ranking, considerando en todo momento la especificidad propia del área a la que pertenecen

los documentos a evaluar; evaluar la incorporación de elementos propios de la lógica difusa y/o inteligencia artificial para automatizar adaptación de los factores de ajuste del algoritmo de ranking; incorporar al análisis de cada artículo resultante una evaluación de la reputación de sus autores para la sub área temática sobre la que se realice la búsqueda; entre otros.

5 Bibliografía

1. Salton, G., Mcgill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., (1983)
2. Kowalski, G.: Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers, Norwell, MA, USA (1997).
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. ACM press, New York (1999)
4. Olivas, J. A.: Búsqueda Eficaz de Información en la Web. Editorial de la Universidad Nacional de La Plata (EDUNLP), La Plata, Buenos Aires, Argentina (2011)
5. Serrano-Guerrero, J., Romero, F. P., Olivas, J. A., de la Mata, J.: BUDI: Architecture for fuzzy search in documental repositories. Mathw. Soft Comput., 16, 1, 71–85 (2009)
6. de la Mata, J., Olivas, J. A., Serrano-Guerrero, J.: Overview of an Agent Based Search Engine Architecture, en Proc. Of the Int. Conf. On Artificial Intelligence IC-AI'04, 62-67. Las Vegas, USA (2004)
7. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A Principal Component Analysis of 39 Scientific Impact Measures. Plos One, (2009)
8. Pendlebury, D. A.: The use and misuse of journal metrics and other citation indicators. Arch. Immunol. Ther. Exp. (Warsz.), 57(1), 1-11 (2009)
9. Garfield, E.: The history and meaning of the journal impact factor. JAMA, 295(1), 90-93 (2006)
10. Gonzalez-Pereira, B., Guerrero-Bote, V., Moya-Anegon, F.: The SJR indicator: A new indicator of journals' scientific prestige, arXiv:0912.4141, (2009)
11. CORE Conference Ranking, Computer Research & Education of Australia, http://www.core.edu.au
12. Hirsch, J. E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A., 102(46), 16569-16572 (2005)
13. Egghe, L.: Theory and practise of the g-index. Scientometrics. 69(1), 131-152 (2006)
14. Jin, B.: The AR-index: complementing the h-index. Issi Newsl. 3(1), p. 6 (2007)
15. Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., Herrero-Solana, V.: Coverage analysis of Scopus: A journal metric approach. Scientometrics. 73(1), 53-78 (2007)
16. Meho, L. I., Yang, K.: A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar. arXiv e-print cs/0612132, (2006)
17. Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., Karageorgopoulos, D. E.: Comparison of SCImago journal rank indicator with journal impact factor. Faseb J. 22(8), 2623-2628 (2008)
18. Leydesdorff, L., Moya-Anegón, F., Guerrero-Bote, V. P.: Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. J. Am. Soc. Inf. Sci. Technol., 61(2), 352–369 (2010)