Evolución de un Algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación

H. Kuna[1], E. Martini[1], M. Rey[1]

[1]Dpto. de Informática, Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones.
hdkuna@gmail.com

Resumen. Un metabuscador es un tipo de sistema software utilizado para la recuperación de información, se caracteriza por operar con resultados obtenidos a partir de motores de búsqueda tradicionales y por estructurarse con elementos desarrollados a medida para un contexto de funcionamiento específico. Una de las cuestiones de diseño que resulta clave, es la fusión y el ordenamiento de los resultados provenientes de las diferentes fuentes de documentos. Este trabajo presenta una evolución de un algoritmo de ranking, desarrollado para un metabuscador y que tiene por finalidad ponderar publicaciones científicas, correspondientes particularmente al área de las ciencias de la computación, ámbito de aplicación del sistema en desarrollo.

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

1 Introducción

En esta sección se describen los ejes teóricos del presente trabajo: Sistemas de Recuperación de Información y Métricas para la Evaluación de documentos científicos utilizadas para la generación del algoritmo de ranking.

1.1 Sistemas de Recuperación de Información

Un SRI (Sistema de Recuperación de Información) es un proceso capaz de almacenar, recuperar y mantener información [1], [2]. En [3] se propone que la estructura básica de un SRI se compone de 4 elementos fundamentales: Los documentos, las consultas del usuario, la manera en que se representan estos elementos y una función de evaluación. Los modelos de SRI más difundidos y que más se extienden sobre internet son los directorios, los buscadores y los meta-buscadores [4]. Dentro de la literatura existen publicaciones, entre ellas [5], [6] donde se extiende el uso de SRI tanto a contextos generales como particulares.

En trabajos anteriores [7], [8] se ha puesto de manifiesto la ausencia de SRI específicos, que se orienten a documentos científicos del área de las CS (ciencias de la computación), a partir de ello se han realizado avances en la construcción de un SRI,

en particular un metabuscador y sus componentes considerando en todos los casos las adaptaciones necesarias para su correcto funcionamiento en un contexto como el planteado. Los principales componentes del metabuscador desarrollado son [8]:

- Módulo para la gestión de consultas: encargado de capturar optimizar y ejecutar las consultas que ingresa el usuario sobre las bases de documentos integradas al SRI.
- Módulo para la gestión de búsquedas: encargado de ejecutar las consultas sobre las fuentes a las que accede el metabuscador y recuperar los resultados de cada una de ellas.
- Módulo para la gestión de resultados: encargado de obtener los listados de resultados recuperados en las búsquedas y procesarlos para su presentación al usuario final en un listado unificado.

Entre los componentes del último de los módulos se destaca el algoritmo de ranking utilizado para establecer el orden de los artículos científicos que serán presentados al usuario como resultado de la ejecución de una búsqueda concreta [7]. En esta nueva presentación se ha extendido el conjunto de métricas a utilizar y se han realizado correcciones en las fórmulas de su aplicación para mejorar la calidad del algoritmo.

1.2 Métricas para la evaluación de documentos científicos

Existe una serie de características que son ampliamente utilizadas para evaluar la calidad de los documentos científicos [9], [10]. Estas características se cuantifican a partir de métricas o indicadores, la mayoría de éstos se enfocan en una de las siguientes propiedades:

- La calidad de la fuente de publicación.
- La calidad de los autores de la publicación.
- La calidad de la publicación.

A través de tales propiedades se puede evaluar a un documento científico en primera instancia con base en la calidad de la fuente o lugar en el que haya sido publicado, distinguiendo si ésta ha sido en una revista o un evento científico de la disciplina; posteriormente considerando la calidad de sus autores, medida a partir del reconocimiento que hayan obtenido publicaciones previas de los mismos; y finalmente la calidad de la publicación en sí, entendida la misma como la repercusión que haya tenido desde su presentación.

En un primer relevamiento de métricas se detectaron aquellas con mayor grado de uso en los últimos años [7]. En el presente trabajo se ha ampliado el relevamiento con el objetivo de incluir métricas que permitan mermar las limitantes de las primeras o por lo menos ampliar el espectro de cobertura y de esa manera mejorar la calidad del algoritmo de ranking desarrollado. Las métricas relevadas pueden observarse en la tabla 1, en la misma se discriminan aquellas previamente relevadas y las que se incorporaron en el marco del trabajo actual.

En el caso de las métricas que permiten valorar la calidad de una fuente de publicación determinada, puntualmente revistas científicas, el IF (Impact Factor) [11]

ha sido por mucho tiempo la opción más extendida, en los últimos años han proliferado una serie de métricas alternativas y/o complementarias. El SJR (SCImago Journal Rank) [12], los indicadores SNIP (Source Normalized Impact per Paper) y RIP (Raw Impact per Paper) [13], el EI (Eigenfactor) y el AI (Article Influence) [14]. También se han encontrado para este criterio, métricas que originalmente fueron concebidas para la evaluación otros aspectos, esto surge debido a la practicidad de dichas métricas, y ha llevado a su implementación, haciendo las adaptaciones necesarias, para constituir un indicador a utilizar fuera del contexto para el cual fueron pensadas en un primer momento. Tal es el caso de la implementación del índice H que es realizado por MAS (Microsoft Academic Research) [15]. En todos los casos se trata de métricas que evalúan a una revista a partir de la influencia de los artículos de la misma en una ventana de tiempo determinada, incluyendo en algunos casos una ponderación de la revista en base a su relación con otras. Para la valoración de eventos o reuniones científicas, no existe una variedad similar en cantidad y calidad de métricas, como alternativa se destacan una serie de rankings de congresos que se calculan a partir de un análisis similar al realizado por las métricas de evaluación de revistas. En lo que respecta a las CS se encuentran el ranking CORE [16] y el ranking ERA [17]; a éstos se suman, nuevamente, implementaciones de métricas originalmente definidas para la evaluación de otros aspectos, pero que han sido adaptadas para la evaluación de congresos por la fuente de datos a partir de la cual son calculadas. Ejemplos de esto son el MAS del cual se puede obtener una implementación del índice H para congresos [18]; y CiteSeerX utilizando sus datos en forma combinada con los de DBLP para generar un IF para eventos científicos [19].

Tabla 1. Métricas relevadas para la evaluación de artículos científicos

Propiedad a evaluar		**Métricas originales**	**Nuevas métricas relevadas**
Tipo de fuente de publicación	Publicación en Revista Científica	IF	
		SJR	
			SNIP
			RIP
			EI
			AI
			Indice H
	Publicación en Congreso Científico	Ranking CORE	
			Ranking ERA
			Índice H
			IF
Autores		Índice H	
			Índice G
			Indice E
			Índice W
Artículo		Índice AR	
		Cantidad de citas	

Para el caso de aquellas métricas que se utilizan para la evaluación de la calidad de los autores, el índice H es el pionero en este aspecto y es el indicador más utilizado, pero el mismo no está libre de limitaciones y a partir de él han surgido cerca de una docena de variantes [20] que tienen por objetivo cubrir los inconvenientes de la métrica original, entre estas variantes se destacan los índices G [21], E [22] y W [23].

Para evaluar la calidad de un artículo científico se propuso mantener las métricas previas, ajustando al índice AR [24] para que opere sobre un único documento, evaluando la cantidad de citas obtenida por el documento al momento de su recuperación considerando su antigüedad, definiendo de esa manera una métrica para la calidad de la publicación.

El resto del artículo se compone de la siguiente manera, en la sección 2 se describe el diseño del algoritmo de ranking, en la sección 3 su implementación y en la sección 4 las conclusiones a las que se han arribado.

2 Diseño del algoritmo de ranking

2.1 Modelo conceptual

El ámbito de los indicadores bibliométricos, sobre todo a la hora de tener que escoger alguno, es un terreno amplio y un tanto difuso. Esto está dado porque la mayoría de las métricas existentes, poseen en mayor o menor medida adeptos y detractores. Tanto [10] como [13] concuerdan de que no existe un indicador perfecto, y que el impacto de la producción científica al ser una construcción multidimensional no puede ser correctamente medida por un solo indicador.

Existen varios inconvenientes que se presentan al tener que utilizar indicadores bibliométricos para evaluar un aspecto de calidad específico de una publicación. Uno de éstos problemas es saber si es lo mismo utilizar un indicador u otro, en [25] se aborda brevemente esta cuestión y se estudia la correlación que existe entre un grupo de indicadores de revistas científicas, la conclusión es que el comportamiento de los mismos varía en función de las áreas del conocimiento. Es decir, algunos indicadores son equivalentes en ciertas áreas y en otras no.

Otra cuestión a tener en cuenta es saber cuál es el valor "real" de una métrica. Esto surge a partir de que la información desde la que se calculan algunos indicadores no es homogénea, por lo tanto, una misma métrica calculada a partir de bases de datos diferentes pueden poseer puntajes diferentes para un mismo objeto de estudio. En [26] por ejemplo, se estudian las variaciones que existen en el índice H de un conjunto de autores en función de la fuente de la que se toman los datos. Un problema estrechamente vinculado al tema anterior, es el grado de solapamiento que hay entre los datos utilizados para calcular ciertos indicadores. Esto es un limitante muy difícil de tratar ya que los indicadores están en continuo desarrollo [27]. A partir de las cuestiones planteadas previamente, resulta por lo menos inexacto o incompleto definir la calidad de algún aspecto de una publicación científica en función de una única métrica.

A partir de estas situaciones, se construyó un modelo (ver figura 1) que hace uso de un conjunto de indicadores para evaluar una publicación específica. La evaluación se hace considerando los 3 aspectos evaluables descriptos en la sección anterior: el lugar de publicación, los autores del documento y al artículo en sí mismo. Para ponderar cada resultado del SRI se obtendrán los valores de las métricas que se encuentren disponibles para esa publicación, pudiendo ser consideradas como diferentes aquellas que se obtengan de fuentes diferentes; además esta condición permite que una métrica podría ser utilizada para la evaluación de más de una propiedad de cada documento.

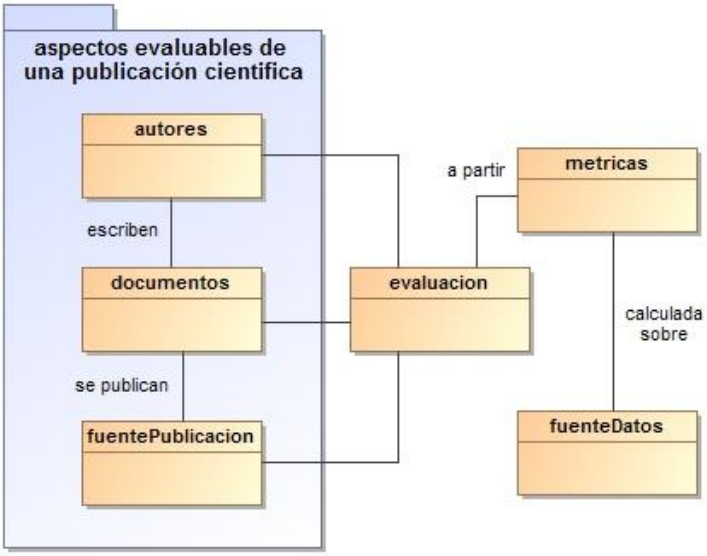

Fig.1 Estructura utilizada para la evaluación de publicaciones científicas

De esta manera, el modelo planteado contempla los problemas presentados anteriormente y a través de sus características permite minimizar sus efectos, ya que el enfoque integrador que se utiliza, permite la evaluación de las propiedades de un documento científico a partir de una serie de métricas con diferentes orígenes, sin que la ponderación final sea dependiente de alguna métrica en particular. Además el modelo es fuertemente escalable, ya que la integración de un aspecto diferente para la evaluación de un documento podría ser realizada sin necesidad de modificar la estructura general del modelo. Asimismo, y de manera análoga al metabuscador, que a medida que incorpora bases de datos en donde buscar artículos se hace más robusto, el algoritmo de ranking se hace más robusto a medida que incorpora un mayor número de métricas y fuentes de datos desde donde obtener sus valores.

2.2 Cálculo de los parámetros

En esta evolución del algoritmo se mantiene la formula general en donde se define que la calificación de una publicación *Q*, resulta de la suma de 3 parámetros: *FP* (Fuente de publicación), *A* (Autores), y *D* (documento), cada parámetro es

multiplicado por un factor de ajuste α, β y γ (ver Formula 1). Los factores de ajuste sirven para variar el peso de los términos a partir de la importancia que se desee dar a cada uno.

$$Q = \alpha * FP + \beta * A + \gamma * D \quad (1)$$

Cada parámetro es calculado internamente a partir de un número de métricas que no está establecido en un primer momento, sino que varía en función de aquellas que se encuentren para la publicación que se esté evaluando. Entonces cada parámetro *P* toma su valor a partir del cociente de la suma de todas las métricas *m* que se encuentren para dicha publicación sobre la cantidad *n* de métricas utilizadas (ver Fórmula 2).

$$P = \frac{\sum_1^n m}{n} \quad (2)$$

Dado que los rangos de valores de las métricas varían considerablemente de una a otra, no resulta posible utilizarlas en conjunto de manera directa. Para poder integrar todas las métricas dentro de un mismo cálculo y mantener un equilibrio de pesos entre ellas dentro de la formula, resulta necesario normalizar sus valores. La normalización propuesta tiene dos variantes y se aplican a partir del tipo de métrica que se esté manejando. La primera variante es la de utilizar el cociente entre el valor que dicha métrica posee para la publicación a evaluar y el valor máximo existente para esa métrica dentro de la misma fuente de datos desde la cual se haya obtenido (ver Fórmula 3). Éste tipo de normalización es usada mayormente en las métricas de los parámetros *FP* y *A*. La variante restante se logra utilizando el logaritmo en base 10 del valor de la métrica (ver Fórmula 4).

$$m = {}^{m}/_{\max(m)} \quad (3)$$

$$m = \log_{10} m \quad (4)$$

3 Implementación del algoritmo de ranking

3.1 Selección y recuperación de métricas a incluir en el modelo

Con el modelo de evaluación y el método de cálculo de los parámetros definidos, se procedió a la implementación de la evolución del algoritmo de ranking. Para ello se comenzó por realizar una selección de las métricas a utilizar para la evaluación de cada uno de los aspectos de un documento científico.

En base al relevamiento descripto en la sección 1.2 se determinó la existencia de un conjunto de métricas que se encuentran definidas de manera teórica y que están destinadas a la evaluación de algún aspecto particular de una publicación científica. Dichas métricas se calculan a partir de información que no es de sencilla recopilación y recuperación. Dada esta realidad y por cuestiones de rendimiento del metabuscador, se ha optado por utilizar en esta instancia, una serie de métricas que estuvieran previamente calculadas a partir de fuentes de datos ampliamente reconocidas como son Scopus, ISI y DBLP entre otras. Se han encontrado de algunas métricas,

diferentes versiones de la misma, que son calculadas a partir de bases de datos diferentes. En estas situaciones se decidió incorporar ambas. En otros casos se observó que algunas bases de datos, por ejemplo MAS, han optado por utilizar para calificar revistas o congresos indicadores que no fueron concebidos originalmente para dicho ámbito, como el Índice H, además de utilizar indicadores nativos de revistas.

Identificadas las implementaciones de las métricas disponibles, se debió evaluar la factibilidad de generar un método para obtener sus valores de manera que puedan ser usadas en el algoritmo de ranking. En todos los casos en los que fue posible hacer la extracción de los valores de las métricas se decidió utilizarlas en forma local, es decir, a partir de una base de datos interna del SRI con el objetivo de optimizar los tiempos de consulta y consecuentemente los de procesamiento del conjunto de resultados.

Para lograr esto, se inició por desarrollar componentes de software que permitieron extraer los valores de las mismas desde la página web en donde estuvieran publicados. Con los datos disponibles, se desarrolló una serie de procesos de transformación y carga de los mismos para lograr homogeneidad en la base de datos.

El conjunto de métricas y fuentes de datos finalmente seleccionadas para la evaluación de los artículos científicos se puede visualizar en la Tabla 2.

Tabla 2. Métricas incorporadas al modelo para el cálculo del algoritmo y su origen

Característica a evaluar		**Métrica utilizada**	**Origen de los datos**
Tipo de fuente de publicación	Revista científica	SJR	Scopus
		RIP	Scopus
		SNIP	Scopus
		Índice H	Scopus
		AI	ISI
		EI	ISI
		EI	MAS
		Índice H	MAS
	Congreso o Evento científico	ERA	ERA
		CORE	CORE
		IF	CiteSeerX + DBLP
		Indice H	MAS
Calidad de los autores		Índice H	ArnetMinner
		Indice G	ArnetMinner
		Índice H	GS
Calidad del artículo		Cantidad Citas	(*)
		Índice AR	

(*) fuentes utilizadas por el agente de búsqueda del metabuscador

Las herramientas software que se utilizaron en las tareas mencionadas fueron: los lenguajes Java, HTML, XML y JSON para la extracción de contenido desde la web, el módulo de integración de datos de la suite de inteligencia de negocios Pentaho (Pentaho Data Integration) y el motor de bases de datos PostgreSQL.

Los valores de las métricas correspondientes a los factores de evaluación de la calidad de la fuente de publicación y la calidad de los autores han sido extraídos, transformados y cargados a la base de datos del SRI. Para el caso del factor de evaluación de la calidad de la publicación se trata de métricas que se obtienen a partir de los metadatos de cada resultado que se recupera a raíz de las búsquedas ejecutadas por el metabuscador por lo tanto, al menos inicialmente, no es posible cargar las mismas en la base de datos.

3.2 Validación del algoritmo desarrollado

Una vez implementado el algoritmo de ranking se debió comenzar el proceso de validación del mismo para su posterior incorporación al SRI. El proceso se planteó en dos etapas, una inicial realizada por expertos y una posterior que tiene por objetivo evaluar estadísticamente la eficiencia del algoritmo, utilizando como medida de rendimiento, la correlación que hay entre el comportamiento de las métricas utilizadas y los resultados obtenidos.

Para el caso de la primera instancia de validación se planteó un esquema de trabajo similar al utilizado en trabajos anteriores [7, 8]. Se han ejecutado una serie de consultas utilizando el metabuscador y se han exportado los conjuntos de resultados junto al detalle de los cálculos llevados a cabo para la aplicación del algoritmo de ranking. Estos resultados han sido evaluados por expertos en la temática de la consulta que han determinado un porcentaje de efectividad de la clasificación realizada sobre los documentos recuperados. Con el objetivo de comparar la aplicación de esta evolución del algoritmo de ranking con respecto a la versión anterior, se comparan los porcentajes de efectividad obtenidos con respecto a los obtenidos con la misma consulta en la publicación original. El detalle y los resultados de la experimentación se pueden observar en la tabla 3.

Tabla 3. Resultados de la primera instancia de validación

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por el experto	Comparación con la versión original
data mining AND outliers	60 (20 Google Scholar + 20 IEEEXplore + 20 ACM Digital Library)	78%	+4%
alphanumeric data AND outliers	60 (20 Google Scholar + 20 IEEEXplore + 20 ACM Digital Library)	86%	+5%
scientific production AND metrics	60 (20 Google Scholar + 20 IEEEXplore + 20 ACM Digital Library)	80%	+3%

La segunda fase de la evaluación se planteó a partir de la solución que el algoritmo podría proporcionar a un problema que se presenta en la evaluación de documentos

científicos como es la correlación entre diversas métricas que se utilizan, por ejemplo: entre la cantidad de citas y el índice SJR de una revista en particular. Con respecto a esto se buscó determinar si la integración en el algoritmo de un conjunto de métricas heterogéneas resultaría en que el valor a otorgar para un documento en especial, y por consiguiente, su evaluación no está sesgada por un indicador en particular.

Esta etapa de validación se encuentra en desarrollo al momento del cierre de la edición del presente trabajo, debido a la cantidad de métricas y documentos a evaluar los resultados se encontrarán disponibles en futuras publicaciones.

4 Conclusiones y trabajos futuros

En el presente trabajo se ha planteado una evolución de un algoritmo de ranking específico, utilizado para ponderar y ordenar documentos científicos pertenecientes al área de las ciencias de la computación. Entre las principales mejoras alcanzadas se puede destacar la construcción de un modelo de evaluación de carácter genérico. Dicho modelo aplaca algunos de los inconvenientes que surgen a la hora de trabajar con indicadores bibliométricos y hace posible la integración de varios de ellos. Permitiendo evaluar un documento de una manera más integral y robusta. Si bien el algoritmo se encuentra todavía en segunda fase de evaluación, posee características que lo hacen flexible y escalable, y ha sido validado positivamente por expertos en la temática. Permitiendo fácilmente la incorporación de otras métricas al cálculo.

Como trabajos futuros se destacan: completar la fase de validación y en caso de ser necesario ajustar el algoritmo. Construir el componente que aplique el algoritmo a los resultados obtenidos por el metabuscador y ordene los resultados. Construir componentes que capturen y actualicen las métricas utilizadas en forma automática.

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, 71–85 (2009).
6. De la Mata, J., Olivas, J.A., Serrano-Guerrero, J.: Overview of an Agent Based Search Engine Architecture. Presented at the , Las Vegas, USA (2004).
7. Kuna, H., Rey, M., Martini, E., Solonezen, L., Sueldo, R.: Generación de un algoritmo de ranking para documentos científicos del área de las ciencias de la computación. XVIII Congreso Argentino de Ciencias de la Computación (2013).

8. Kuna, H., Rey, M., Martini, E., Solonezen, L., Podkowa, L.: Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. ReLAIS. 2, 107–114 (2014).
9. Pendlebury, D.A.: The use and misuse of journal metrics and other citation indicators. Arch. Immunol. Ther. Exp. (Warsz.). 57, 1–11 (2009).
10. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A Principal Component Analysis of 39 Scientific Impact Measures. PLoS ONE. 4, e6022 (2009).
11. Garfield, E.: Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science. 178, 471–479 (1972).
12. Gonzalez-Pereira, B., Guerrero-Bote, V., Moya-Anegon, F.: The SJR indicator: A new indicator of journals' scientific prestige. arXiv:0912.4141. (2009).
13. Moed, H.F.: Measuring contextual citation impact of scientific journals. J. Informetr. 4, 265–277 (2010).
14. Bergstrom, C.: Measuring the value and prestige of scholarly journals. Coll Res Libr News. 68, 3146 (2007).
15. Help Center - Microsoft Academic Search, http://academic.research.microsoft.com/About/Help.htm, (2014).
16. CORE: CORE Conference Ranking. Computer Research & Education of Australia, http://core.edu.au/index.php/conference-rankings, (2008).
17. The Australian Research Council: ERA 2012 Journal and Conference Lists, http://www.arc.gov.au/era/era_2012/era_journal_list.htm, (2012).
18. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. 102, 16569–16572 (2005).
19. CiteSeerX — Statistics - Venue Impact Factors, http://citeseerx.ist.psu.edu/stats/venues, (2014).
20. Van Noorden, R.: Metrics: A profusion of measures. Nat. News. 465, 864–866 (2010).
21. Egghe, L.: Theory and practise of the g-index. Scientometrics. 69, 131–152 (2006).
22. Zhang, C.-T.: The e-Index, Complementing the h-Index for Excess Citations. PLoS ONE. 4, e5429 (2009).
23. Wu, Q.: The w-index: A significant improvement of the h-index. ArXiv Prepr. ArXiv08054650. (2008).
24. Jin, B., Liang, L., Rousseau, R., Egghe, L.: The R- and AR-indices: Complementing the h-index. Chin. Sci. Bull. 52, 855–863 (2007).
25. Torres-Salinas, D., Jiménez-Contreras, E.: Introducción y estudio comparativo de los nuevos indicadores de citación sobre revistas científicas en Journal Citation Reports y Scopus. El Prof. Inf. 19, 201–208 (2010).
26. Bar-Ilan, J.: Which h-index? — A comparison of WoS, Scopus and Google Scholar. Scientometrics. 74, 257–271 (2007).
27. Sicilia, M.-A., Sánchez-Alonso, S., García-Barriocanal, E.: Comparing impact factors from two different citation databases: The case of Computer Science. J. Informetr. 5, 698–704 (2011).