Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación

H. Kuna[1,2], M. Rey[1], E. Martini[1], L. Solonezen[1], L. Podkowa[1]
[1]Departamento de Informática, Fac. de Cs. Exactas, Químicas y Naturales. Universidad Nacional de Misiones. Argentina
[2]Facultad de Ingeniería – Universidad Nacional de Itapúa, Paraguay
hdkuna@gmail.com

***Resumen* — La recuperación de información desde la web es una actividad recurrente para el investigador en general que, dada su complejidad, puede insumir una cantidad de timpo considerable. En este contexto, la relevancia y la calidad de los resultados obtenidos en cada búsqueda realizada son cruciales. Los Sistemas de Recuperación de Información se presentan como herramientas de gran utilidad para optimizar el proceso de búsqueda de información en la web, específicamente cuando se generan para contextos de funcionamiento acotados. En el presente trabajo se presenta el desarrollo de un meta-buscador orientado exclusivamente a la recuperación de publicaciones científicas del área de ciencias de la computación. Se incluye la descripción de los componentes desarrollados a medida, considerando las particularidades del tipo de documentos a recuperar, para la mejora de la relevancia y calidad de los resultados a retornar al investigador, comenzando por el proceso de expansión de consultas y finalizando por el algoritmo de ranking que establece el orden final de los resultados para su presentación al usuario.**

***Palabras Clave* — Recuperación de Información, Ontología, Algoritmo de Ranking, Búsqueda Web, Indicadores Bibliométricos, Meta-buscador.**

I. Introducción

En la actualidad internet facilita el acceso a grandes volúmenes de información desde cualquier parte del mundo. La administración de tal cantidad de información es una tarea cuya complejidad ha crecido durante los últimos años, determinado esto por la cantidad y complejidad de los documentos disponibles en la web y su constante cambio.

La información en la web se busca y recupera de diversas maneras, una de ellas es a través de los motores de búsqueda, como por ejemplo: Google[1], Yahoo[2] o Bing[3]. Éstos han mejorado progresivamente y se puede decir que son bastante eficientes, aunque los resultados que proporcionan al usuario no sean del todo eficientes en cuanto a cantidad, calidad y, principalmente, correlación con el objetivo de la búsqueda realizada [1].

El objetivo del presente trabajo es desarrollar de un Sistema de Recuperación de Información de dominio específico, en particular un meta-buscador que se oriente exclusivamente a la recuperación de documentos científicos del área de ciencias de la computación. Se plantea el desarrollo de todos sus componentes, incluyendo aquellos cuya finalidad es realizar la expansión de las consultas ingresadas por el usuario final haciendo uso de una ontología y la evaluación de los documentos recuperados para establecer un ranking considerando su calidad como artículo científico.

[1] www.google.com
[2] www.yahoo.com
[3] www.bing.com

II. Antecedentes

A. Sistemas de Recuperación de Información

Un Sistema de Recuperación de Información (SRI) es un proceso que posee capacidad suficiente para gestionar información, entendiéndose por esta gestión a su recuperación, almacenamiento y mantenimiento para diversos fines según el contexto de su aplicación [2], [3]. En la literatura del área de ciencias de la computación se pueden observar diversas propuestas sobre la organización interna de un SRI, en el marco del presente trabajo se utilizará aquella que se basa en los siguientes elementos [4]:

- Los documentos, que constituyen la fuente de información sobre la cual se pretende realizar búsquedas.
- Las consultas, generadas por los usuarios del SRI que tienen por objetivo recuperar información a la cual el sistema provee acceso.
- La representación de los documentos, las consultas y las relaciones que se definan entre ellos que sean definidas teniendo en cuenta el ámbito de aplicación del SRI.
- La función de evaluación, que determina la pertinencia de cada documento recuperado para dar solución a la consulta que haya ingresado el usuario.

Los principales tipos de SRI que en la actualidad operan sobre internet son: los directorios, los buscadores y los meta-buscadores [1]. Se puede afirmar que existen implementaciones de SRI en la web que utilizan diferentes métodos de búsqueda sobre contextos generales o particulares, contando con implementaciones desarrolladas a medida para el incremento de la relevancia de los resultados a presentar al usuario [5], [6].

En el contexto del presente trabajo cobran una mayor notoriedad los meta-buscadores, ya que gracias a su modularidad, permiten que los componentes del SRI sean desarrollados a medida para las necesidades particulares de su contexto de aplicación. En este caso se presenta el desarrollo realizado particularmente sobre los siguientes componentes:

- El componente que captura la consulta del usuario y la expande, generando consultas similares para expandir el espectro de búsqueda.
- El componente que accede a las fuentes de datos sobre las que serán realizadas las búsquedas y recupera de cada una de ellas los documentos resultantes de la ejecución de una consulta.
- El componente que aplica la función de evaluación de los documentos obtenidos desde cada fuente para ordenar el listado integral a presentar al usuario.

En todos los casos se trata de desarrollos específicos determinados por el ámbito de uso propuesto para la herramienta.

B. SRI para documentos científicos del área de ciencias de la computación

Si bien se han detectado diversas iniciativas tendientes a la generación de SRI de propósito específico en áreas particulares, como por ejemplo en ciencias de la salud [7], u otras con un perfil de aplicación más general [8], inclusive algunas que realizan búsquedas de referencias bibliográficas [9]; 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. Tampoco de productos de este tipo que implementen soluciones complementarias para resolver aspectos clave como son la expansión de las consultas del usuario considerando el contexto de la búsqueda y la aplicación de métodos de evaluación de los documentos en base a la calidad de los mismos con la finalidad de mejorar la relevancia de los elementos del listado de resultados a presentar al usuario.

Explotando las capacidades de los meta-buscadores antes mencionadas, se considera factible generar un SRI que utilice bases de datos de otros buscadores que sean específicos para la recuperación de documentos científicos del área de ciencias de la computación. Así como también el desarrollo de componentes complementarios, tanto para el tratamiento de las consultas introducidas por el usuario, como para la aplicación de un algoritmo de ranking específico para evaluar el tipo de resultados con el que se desea operar, según diferentes métricas, ampliamente aceptadas por la comunidad científica considerando las características propias del área, asignando a cada resultado una calificación que servirá de referencia para establecer el orden de los resultados en el listado final a presentar al usuario [5].

C. Expansión de consultas en un SRI y ontologías

Un SRI cuenta con varias alternativas para lograr optimizar el proceso de búsqueda de información, una de ellas consiste en tomar la consulta que ingresa el usuario y ampliarla a partir de agregar diversos términos, usualmente obtenidos a través de fuentes externas, manteniendo coherencia con el dominio de la consulta. Este método es conocido como expansión de consultas (QE por su sigla en inglés); los términos adicionales generan nuevas consultas, denominadas expansiones [10], [11]. De esta manera el SRI adquiere la capacidad de acceder a una mayor cantidad de documentos relevantes para el usuario, para ello se ejecutan, en una misma sesión de búsqueda, las diferentes consultas sobre cada base de datos a la que posea acceso el SRI, obteniendo listados de resultados individuales por cada expansión que posteriormente son unificados y ponderados antes de ser presentados al usuario [5], [12].

Existen diferentes opciones para la implementación de un proceso de expansión de consultas para un SRI, entre las cuales se pueden mencionar: uso de tesauros, diccionarios, sistemas expertos, entre otros [10], [13], [14]. En el caso particular de este trabajo se hace uso de una ontología de dominio específica para un sub-área temática de las ciencias de la computación.

En general se define a una ontología como una forma de representación del conocimiento de un ámbito específico, que utiliza los términos y relaciones que conforman su vocabulario base, agregando elementos que permiten extender el vocabulario, como son las relaciones entre conceptos, permitiendo organizarlos jerárquicamente, facilitando su utilización para la generación de conocimiento en forma automática [15].

Adaptando la definición anterior al área de ciencias de la computación, una ontología se puede considerar como un esquema conceptual correspondiente a un dominio acotado, que permite la comunicación y transmisión de información entre sistemas, tanto interna como externamente [16]. Constituye una herramienta de gran utilidad para la recuperación de información dado que facilita el tratamiento y análisis del conocimiento a través de una estructura de clases y subclases que adquiere sentido con las relaciones, propiedades y reglas definidas entre las instancias de las mismas [17].

D. Métricas para la Evaluación de Documentos Científicos

El SRI planteado requiere el desarrollo de un método particular para la evaluación de los documentos con los que trabajaría. Al tratarse de artículos científicos se hizo necesario determinar aquellas características de los documentos que serían evaluadas, quedando seleccionadas las siguientes [18], [19]:

- La calidad de la fuente de publicación, que hace referencia a dónde se ha publicado el artículo, pudiendo ser una revista científica o un congreso o reunión científica de similares características.
- La calidad de los autores, valorando la importancia que hubieran tenido las publicaciones que hayan realizado a lo largo de su carrera.
- La calidad del artículo en sí, considerando la antigüedad del mismo y la cantidad de veces que haya sido citado en otros documentos.

Para cada una de las características enunciadas se distinguen diversos indicadores bibliométricos, ver tabla 1, que han sido validados por la comunidad científica.

Para el caso en que el tipo de fuente de publicación sea una revista existen dos índices que se utilizan para estimar su calidad: por un lado el Factor de Impacto (IF, por sus siglas en inglés) [20] desarrollado por la Web of Knowledge, que permite medir la importancia que ha tenido una revista a partir de las citas que han recibido los artículos que se han publicado en ella en un año en particular; y el índice SJR (SCImago Journal Rank) [21] desarrollado por el grupo homónimo que tiene en cuenta las publicaciones científicas de revistas listadas en la base de datos de Scopus.

TABLA I. MÉTRICAS ANALIZADAS PARA LA EVALUACIÓN DE ARTÍCULOS CIENTÍFICOS

Propiedad a evaluar	Métricas		Entidad que da soporte a la métrica
Calidad de la fuente de publicación	Publicación en revista científica	Factor de Impacto (IF) **[20]**	Web of Knowledge[a] – Institute for Scientific Information (ISI)
		SCImago Journal Rank (SJR) **[21]**	Scopus[b] – Grupo SCImago, Univ. De Extremadura, España
	Publicación en Congreso o Evento Científico	Ranking CORE **[22]**	Computer Research & Education of Australia[c]
Calidad de los autores	Índice H **[23]**		-
	Índice G **[24]**		-
Calidad del artículo	Índice AR **[25]**		-
	Cantidad de citas		-

a. www.wokinfo.com – Accedido: 30/03/14
b. www.scopus.com – Accedido: 30/03/14
c. www.core.edu.au – Accedido: 30/03/14

Está inspirado en el PageRank de Google [26] para evaluar el impacto de una publicación de acuerdo al número de citas recibidas con respecto a relevancia de las publicaciones que la citan. Este establece una clasificación de acuerdo a ciertos parámetros, como ser: área de conocimiento, categoría y país.

Mientras que en caso de que la publicación se realice en un congreso o evento similar se cuenta con el ranking CORE [22] desarrollado por la Computer Research & Education of Australia. Un congreso o conferencia es clasificado, según su importancia, en un determinado nivel preestablecido: A*, A, B y C, respectivamente. Este listado de congresos que es de carácter público, es accesible desde la propia web de dicho instituto.

Para estimar la calidad de la producción de un autor se dispone del índice H [23]. Este se calcula en base en la distribución de las citas que han recibido las publicaciones científicas de un determinado autor. De tal manera que, para hallarlo, solo basta ordenar de forma descendente las publicaciones de un autor por el número de veces que ha sido citada cada publicación, y, de esta manera, se identifica el punto en el que el número de orden coincide con el de citas recibidas por publicación, este número representa el índice H. Otra métrica igual de valida es el índice G [24]. Para calcularlo, primeramente los artículos de un autor son ordenados de manera descendente de acuerdo con el número de citas recibidas por cada uno de ellos, al igual que lo hace el índice H. Aquel número mayor en el orden del ranking donde la sumatoria de citas recibidas por el autor sea mayor o igual al cuadrado del número de orden, es considerado como el índice G de dicho autor.

Para evaluar la calidad de una colección de publicaciones se puede utilizar un índice como es el AR [25], dicho indicador combina la cantidad de citas con la antigüedad de cada publicación, para así establecer una valoración de la colección. La cantidad de citas recibidas, por sí solas, también es utilizada como métrica para evaluar la calidad de un documento científico en particular.

III. MATERIALES Y MÉTODOS

A. Estructura del SRI

El trabajo abordado previamente por los autores ha consistido en el planteo inicial de lineamientos generales sobre la estructura del SRI a desarrollar [27], basándose en desarrollos similares que fueron tomados como referencia [5], [6], [28].

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

- Módulo para la gestión de las consultas (MC): encargado de realizar inicialmente la expansión de consultas utilizando el método basado en el uso de una ontología, para posteriormente adaptar las expansiones obtenidas y posibilitar su uso en los buscadores integrados.
- Módulos para la búsqueda en las bases de datos (buscadores) (MB): encargado de tomar las consultas adaptadas y controlar su ejecución sobre las diferentes fuentes de documentos incorporadas al SRI. Actualmente los buscadores consultados son: Google Scholar[4], ACM Digital Library[5], IEEE Xplore[6].
- Módulo para la gestión de los resultados (MR): encargado de procesar los resultados a través del componente que permite la aplicación del algoritmo de ranking desarrollado a medida para el SRI dado su ámbito de trabajo [29]. Como resultado obtiene un valor representativo de la calidad de cada documento recuperado a partir de las búsquedas, que se utiliza como referencia para establecer el orden de aparición del documento en el listado final a presentar al usuario.

B. Funcionamiento del SRI

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

1) El usuario ingresa la consulta al sistema.
2) El MC toma el texto de la consulta y ejecuta el proceso de expansión basado en la ontología.
3) Para cada una de las expansiones generadas, el MC las adapta según los requerimientos determinados por cada fuente de datos a consultar, considerando cantidad de resultados y conectores entre términos.
4) El MB toma todas las consultas adaptadas y las ejecuta sobre la fuente de datos correspondiente.
5) Por cada búsqueda iniciada, el MR recibe el listado de documentos generados de cada fuente y los filtra, eliminando aquellos que no se consideran relevantes en cuanto a su tipo (por ejemplo: resumen de citas de un documento).
6) Se unifican las colecciones de resultados obtenidas de los buscadores consultados, incluyendo un paso de limpieza de resultados duplicados.
7) Para cada documento individual del listado unificado, se aplica el algoritmo de ranking.
8) Se ordena el listado de documentos en base al valor obtenido por cada resultado particular al aplicarse el algoritmo de ranking.
9) Se formatea el listado ordenado para su presentación al usuario final, permitiendo al usuario visualizar los

[4] scholar.google.com – Accedido 30/03/14.
[5] dl.acm.org – Accedido 30/03/14.
[6] ieeexplore.ieee.org/ - Accedido 30/03/14.

siguientes datos por 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.

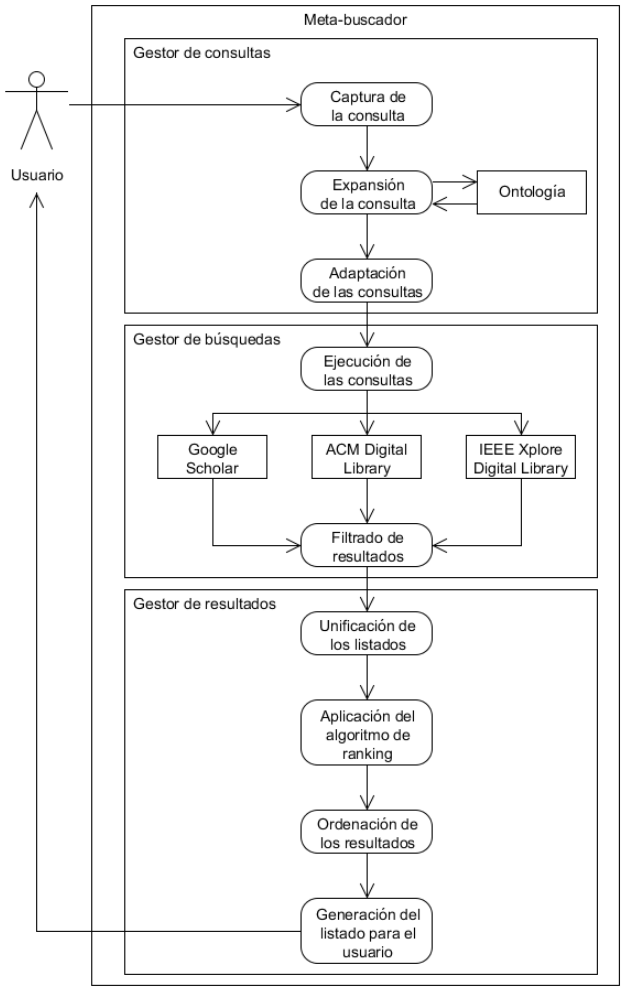

Figura 1. Proceso de búsqueda del SRI

C. Construcción de la ontología para la expansión de consultas

El componente principal del método de expansión de consultas desarrollado para el meta-buscador es la ontología en base a la cual se realiza la expansión. Su diseño requirió de las siguientes actividades [30]:

1) *Definición del dominio de la ontología:* dado el contexto de aplicación del SRI se comenzó por seleccionar un sub-área temática dentro de las ciencias de la computación a partir de la cual se realizaría la ontología. Se determinó que se comenzaría por el sub-área de Inteligencia Artificial (IA).
2) *Determinación de los términos a incluir en la ontología:* se generó un listado de términos a partir de una revisión del estado del arte de la disciplina, obteniendo un conjunto de conceptos que caracterizan a las subdivisiones de la IA, conjuntamente se definieron sinónimos de cada concepto que pudieran ser utilizados en el método de expansión [31]–[33].
3) *Definición de la jerarquía de clases:* utilizando el método *top-down*, se definió como clase a todo concepto que representara una categoría en la que la disciplina pudiera subdividirse, definiendo dentro de cada clase aquellas instancias que serían consideradas como subclases, con el objetivo de que la ontología represente de la mejor manera posible la taxonomía original del área.
4) *Definición de propiedades para las clases:* para esta actividad se consideró como propiedad a aquellos atributos que pudieran describir a los conceptos, teniendo en cuenta las relaciones que pueden existir entre los mismos. Se definieron relaciones implícitas y explícitas, las primeras están dadas por la estructura subyacente de la ontología, es decir, que se definen a partir de las conexiones entre los conceptos, como ser:
   - "Es un (padre)", simbolizando la situación de que una clase contiene a otra.
   - "Es un (hijo)", relación inversa a la anterior, que representa que una clase es contenida por otra.
   - "Es un (hermano)", simbolizando a un conjunto de clases que comparten un mismo "padre".

   En cuanto a las relaciones explícitas, aquellas definidas para dar mayor valor a la ontología considerando su objetivo de aplicación, se determinó que sería representada la relación de sinonimia, utilizando para ello los términos relevados al inicio del proceso.
5) *Creación de instancias para las clases:* utilizando los términos de mayor atomicidad, se definieron los conceptos que conformarían las instancias de las clases de la ontología.

Concluido el diseño de la ontología, se seleccionó la herramienta software Protègè [34] para su implementación, reflejando en la misma el resultado del proceso descripto en la sección anterior.

D. Desarrollo del método de expansión de consultas

El método para realizar la expansión de consultas tendría por objetivo obtener de la ontología aquel concepto que guardara mayor similitud con la consulta ingresada por el usuario (*consulta_original* en adelante) y, a partir de ese concepto, obtener a los conceptos relacionados al mismo en forma implícita y explícita, es decir, su "padre", sus "hermanos" y los sinónimos del término en sí.

Con tal objetivo en mente, el algoritmo se divide en dos etapas, inicialmente se busca al concepto de la ontología más similar a la *consulta_original* (*concepto_candidato* en adelante) y posteriormente se efectiviza la expansión al disponer de los conceptos relacionados.

La búsqueda del *concepto_candidato* consta de los siguientes pasos:

1) Para cada término de la *consulta_original*: se recorre la ontología en su totalidad almacenando en una colección temporal aquellos conceptos con mayor cantidad de coincidencias.
2) En base al contenido de la colección resultante del paso anterior:
   a) Si no contiene ningún elemento: se finaliza la operación de expansión sin resultados válidos.
   b) Si contiene un único elemento: se considera al mismo como el *concepto_candidato*.
   c) Si contiene más de un elemento: se analiza cada concepto cuantificando la cantidad de coincidencias sintácticas con respecto a la *consulta_original*. En caso de empate, se procede a evaluar al elemento en base a las relaciones del mismo en la ontología:

- Si todos tienen el mismo "padre": tal concepto es seleccionado como *concepto_candidato*.
- Si no tienen el mismo "padre": se evalúa la cantidad de instancias contenidas por cada "padre", aquel que presente la mayor cantidad será el *concepto_candidato*. En caso de un nuevo empate: cada uno de los "padres" involucrados será considerado *concepto_candidato* conformando una colección nueva.

Con el o los candidatos determinados se inicia la segunda etapa del método de expansión, que se compone de los siguientes pasos:

3) Se obtiene el concepto "padre" (concepto_padre en adelante) del concepto_candidato.
4) Se obtienen los conceptos del mismo nivel que el concepto_candidato, dando lugar a la colección [conceptos_hermanos].
5) Se obtienen, en caso de existir, los sinónimos del concepto_candidato, dando lugar a la colección [sinónimos_concepto].
6) Se expande la consulta_original haciendo uso de los elementos obtenidos a través de los pasos anteriores, las expansiones quedan conformadas según las fórmulas 1 a 4:

Expansión_1= consulta_original AND concepto_candidato (1)

Expansión_2= concepto_candidato AND concepto_padre (2)

Expansión_3= concepto_candidato OR [conceptos_hermanos] (3)

Expansión_4= concepto_candidato OR [sinónimos_concepto] (4)

Al finalizar la ejecución del algoritmo se cuenta con las cuatro expansiones de la *consulta_original* ingresada por el usuario al sistema.

E. Diseño del algoritmo de ranking

Uno de los puntos claves a la hora de desarrollar el SRI, fue definir la manera en que se ponderaría a cada uno de los resultados obtenidos, ya que su ubicación en la lista final de resultados, surgiría a partir de dicha ponderación. Para ello se optó por diseñar un algoritmo que evalúe diversos aspectos que hacen a la calidad de una publicación académica.

En primera instancia se definió que las características a evaluar serían: el lugar en donde se ha publicado o fuente de publicación, los autores de la publicación, y la calidad propia del artículo. Luego se relevaron, se analizaron y se incorporaron al algoritmo distintas métricas que valoren alguno de los 3 criterios. Las métricas que finalmente se utilizaron fueron:

- *Calidad del lugar de publicación:* Para definir este criterio hay que tener en cuenta que un trabajo de investigación tiene dos grandes maneras de divulgarse, por un lado se encuentran las revistas científicas y por otro lado los congresos. En el caso de que el resultado se haya publicado en una revista científica, se optó por utilizar como métrica el índice SJR [21], este índice es desarrollado por Elsevier en base a los datos del buscador Scopus y posee ciertas ventajas [35], [36] frente a su principal alternativa, el IF de ISI [20], entre las que se destacan: que es de acceso abierto; que la cantidad de revistas que incluye, en combinación con la base de datos de Scopus es superior; que incluye revistas en idiomas distintos al inglés; que utiliza una evaluación tanto cuantitativa, dado por la cantidad de citas recibidas, como cualitativa ya que las citas de revistas más prestigiosas poseen mayor valor respecto a las que no lo son. Y en el caso de que el resultado sea procedente de un congreso la métrica utilizada fue el ranking creado por la Computing Research and Education Association of Australia (CORE) [22].
- *Calidad de los autores:* existen varias alternativas a la hora de elegir métricas para evaluar la producción científica de un autor, entre ellas se destaca el índice H [23], ya que es la pionera en este aspecto y pese a ser resistida por cierta parte de la comunidad científica, es ampliamente aceptada y utilizada [18], [19], además ha sido la base para la creación de otras métricas [24], [37], [38].
- *Calidad de la publicación:* para el presente criterio se decidió incorporar al algoritmo una combinación adaptada de las 2 métricas relevadas, tanto el índice AR [25], autoproclamado como complemento del índice H, como la cantidad de citas, remarcando que la primera se adaptó para evaluar los documentos de manera independiente y no una colección como lo hace el índice original.

F. Desarrollo del algoritmo de ranking

Luego de seleccionar las métricas que serían la base para el armado del algoritmo, fue necesario normalizar sus valores, ya que en cada una de las métricas originales, los valores máximos y mínimos pueden llegar a ser muy diferentes, por ejemplo: un autor prolífico puede poseer un índice H que ronda un valor de 100, mientras que la revista más prestigiosa difícilmente alcance un índice SJR superior a 40. En otras palabras lo que se hizo fue adaptar las métricas para que sus valores estén en función de una escala común. Así las fórmulas correspondientes a cada uno de los parámetros quedaron conformadas de la siguiente manera:

- Para el factor de la fuente de publicación, como se mencionó anteriormente, hay 2 métricas utilizadas: el índice SJR y el ranking CORE. En el caso de que la fuente haya sido una revista, se toma el índice SJR y se aplica a dicho índice el logaritmo en base 10 (fórmula 5) con lo que queda definido el factor. En el caso de que la fuente haya sido un congreso, se busca la categoría del congreso en el ranking y a partir de esa categoría se establece el valor según lo expresado en la fórmula 6.

$$fuentePublicación = \log_{10}(SJR) \quad (5)$$

$$fuentePublicación \in [A^* = 1; A = 0.75; B = 0.5; C = 0.25 \quad (6)$$

- Para el caso de los autores se considera el índice H de cada uno de ellos de la siguiente manera (fórmula 7): de cada autor se toma una fracción de su índice H según la ubicación que tengan en el artículo, así, del primer autor se toma la totalidad, del segundo la mitad, del tercero un tercio y así sucesivamente. Luego se suman todos los valores obtenidos y se aplica el logaritmo en base 10 a dicha sumatoria.

$$autores = \log(\sum indiceH(autori)/i) \quad (7)$$

- Para el factor correspondiente a la calidad del documento en evaluación, se utilizaron los lineamientos propuestos

por el índice AR, en el mismo se trata de definir la vigencia de un trabajo a través del tiempo, esto se hizo, como puede verse en la fórmula 8, a partir del logaritmo en base 10 del cociente entre la cantidad de citas recibidas y la antigüedad del trabajo, la diferencia radica en que el índice mencionado somete dicho cociente a la raíz cuadrada.

$$calidadPublicación = log_{10}(citasRecibidas / antiguedadPublicación). \quad (8)$$

Una vez definida la manera de calcular cada parámetro, se le agregó a cada uno un factor de ajuste (alfa, beta y gamma), a partir del cual se podría hacer variar el peso de un parámetro en el puntaje final. Estos factores de ajuste pueden tomar valores entre 0 y 1, de esta manera cuando un factor de ajuste toma valor 0 el parámetro asociado a dicho factor se anula. Contrariamente cuando el factor toma valor 1 el parámetro aporta toda su ponderación al puntaje final.

En definitiva, el puntaje que se le otorga a un documento y a partir del cual se define su posición entre los resultados, viene dado por (fórmula 9): la suma de 3 parámetros (fuente de publicación, autores y calidad de la publicación) cada uno de ellos multiplicado por su factor de ajuste, el cual va a determinar la importancia de éste en el puntaje final. Para la fase de experimentación, y en forma conjunta con expertos en la temática, los valores establecidos para cada factor de ajuste fueron 0.5, 0.3 y 0.2 respectivamente.

$$valorFinal = \alpha * [fuentePublicación] + \beta * [autores] + \gamma * [calidadPublicación] \quad (9)$$

IV. Experimentación

A. Desarrollo del prototipo de SRI para la experimentación

Una vez diseñado el meta-buscador y sus componentes, se requirió contar con una implementación del mismo para poder desarrollar la experimentación que condujera a su validación. La tecnología utilizada para el desarrollo de la herramienta consistió en: los lenguajes Java, JSP, Javascript, xHTML y SQL, junto al motor de bases de datos MySQL, utilizando como plataforma para su implementación el servidor web Tomcat, habiéndose priorizado aquellas que permitieran usar el SRI desde la web y que fueran de código abierto.

El proceso de implementación del meta-buscador se descompuso en los siguientes pasos:

1) Diseño y desarrollo de los métodos para acceder, consultar y extraer los resultados de los sitios web de los buscadores Google Scholar, ACM Digital Library e IEEE Xplore.
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 meta-buscador, es decir, la interfaz de captura de las consultas del usuario y la correspondiente para la visualización del listado de resultados unificado y ordenado según el valor obtenido por cada documento al aplicar el ranking.
4) Integración de todos los componentes en un único producto software.

B. Validación del SRI desarrollado

Para el proceso de validación se ha contado con la colaboración de un grupo de expertos en el desarrollo de métodos de recuperación de información quienes han evaluado al SRI considerando la relevancia de los documentos que son retornados por el mismo a partir de diferentes consultas. Se ha evaluado el proceso completo de búsqueda, desde la expansión de la consulta ingresada, la ejecución de las expansiones generadas en cada fuente de datos y el procesado posterior de los resultados a través del algoritmo de ranking.

El detalle de la experimentación se observa en la tabla 2, la cantidad de consultas ha sido la recomendada por los expertos, y la cantidad de resultados a obtener fue configurada a niveles adecuados para el correcto seguimiento de todo el proceso en forma manual por parte de los evaluadores. Se ha verificado la pertenencia de cada documento con respecto a la consulta ingresada, teniendo en cuenta la relación directa entre sus términos y aquellos resultantes a partir del proceso de expansión realizado a través de la ontología, además de analizar la posición que cada documento obtuvo en el listado de resultados final en base a su calificación por el algoritmo de ranking. A partir de la combinación de esos factores se ha determinado qué porcentaje de los resultados serían considerados como efectivamente relevantes para el usuario, sirviendo tal métrica como medida de desempeño del meta-buscador.

Como resultado de la experimentación, se ha determinado que el SRI desarrollado, a través de sus diversos componentes, constituye una herramienta para recuperar documentos científicos de calidad comprobable a partir de una consulta del usuario.

TABLA II. (A) RESULTADOS DE LA VALIDACIÓN DEL SRI POR PARTE DE LOS EXPERTOS

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por los expertos
intelligent agents AND web information retrival	120 (40 por buscador)	68%
search methods AND deep-first-search	120 (40 por buscador)	64%

TABLA II. (B) RESULTADOS DE LA VALIDACIÓN DEL SRI POR PARTE DE LOS EXPERTOS

unsupervised learning AND backpropagation networks	120 (40 por buscador)	80%
genetic algorithms AND distributed methods	120 (40 por buscador)	62%
natural language processing AND ontologies	120 (40 por buscador)	72%
artificial intelligence AND computer vision	120 (40 por buscador)	68%
classes of agents AND deliberative agents	120 (40 por buscador)	60%
fuzzy controllers AND robotics	120 (40 por buscador)	76%
fuzzy sets AND expert systems	120 (40 por buscador)	74%
neural networks AND self organizing maps	120 (40 por buscador)	88%

V. Conclusiones

En la actualidad la cantidad de información a la que se puede tener acceso a través de internet, se ha vuelto irrisoria. Es por esto que cada vez cobra mayor importancia la eficiencia con la que se accede a ella y la relevancia de los resultados obtenidos con respecto a la consulta ingresada por el usuario. En este artículo se ha presentado el diseño, la implementación y validación de un SRI, específicamente un meta-buscador para un área de conocimiento específica como son las ciencias de la computación, limitando los documentos a recuperar a publicaciones científicas. Para poder alcanzar mejores resultados se trabajó sobre diferentes componentes como el acceso a múltiples fuentes, la realización de una expansión de la consulta del usuario utilizando ontologías, y la evaluación de cada resultado a través de un algoritmo de ranking específicamente desarrollado para cuantificar la calidad de cada documento recuperado y así poder ordenar los resultados a presentar al usuario en base a su relevancia.

Como trabajos a futuro se pueden mencionar: generar las ontologías restantes para cubrir todas las sub-áreas dentro de las ciencias de la computación de modo de completar el método de expansión de consultas, incorporar otros indicadores bibliométricos al algoritmo de ranking que sean adecuados para el tipo de documentos en evaluación, evaluar la incorporación de elementos que permitan automatizar la adaptación de los factores de ajuste del algoritmo de ranking en base a las preferencias del usuario, considerar la reputación de los autores de un determinado documento en el área temática en la que se haya producido la publicación como otro factor en el algoritmo de ranking, entre otros.

Agradecimientos

Los autores del presente trabajo agradecen a los miembros del Grupo de Investigación SMILe dirigido por el Dr. J. A. Olivas Varela, perteneciente a la Universidad de Castilla-La Mancha, España, su colaboración en calidad de expertos en la evaluación del SRI desarrollado.

Referencias

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

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

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

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

[6] 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.

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

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

[9] 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. Technol.*, vol. 59, n.º 3, pp. 375-389, feb. 2008.

[10] M. de la Villa, S. García, y M. J. Maña, «¿De verdad sabes lo que quieres buscar? Expansión guiada visualmente de la cadena de búsqueda usando ontologías y grafos de conceptos», *Proces. Leng. Nat.*, vol. 47, n.º 0, pp. 21-29, sep. 2011.

[11] Y. Chang, I. Ounis, y M. Kim, «Query reformulation using automatically generated query concepts from a document space», *Inf. Process. Manag.*, vol. 42, n.º 2, pp. 453-468, mar. 2006.

[12] J. Ruiz-Morilla, J. Serrano-Guerrero, J. Olivas, y E. Viñas, «Representación Múltiple de Consultas: Una alternativa a la Expansión de Consultas en Sistemas de Recuperación de Información», en *Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF*, 2010, pp. 531–536.

[13] S. Gauch y J. B. Smith, «An expert system for automatic query reformation», *J. Am. Soc. Inf. Sci.*, vol. 44, n.º 3, pp. 124-136.

[14] J. C. French, D. E. Brown, y N.-H. Kim, «A classification approach to Boolean query reformulation», *J. Am. Soc. Inf. Sci.*, vol. 48, n.º 8, pp. 694-706, ago. 1997.

[15] S. Delisle, «Towards a better integration of data mining and decision support via computational intelligence», en *Sixteenth International Workshop on Database and Expert Systems Applications, 2005. Proceedings*, 2005, pp. 720 - 724.

[16] A. Muñoz y J. Aguilar, «Ontología para bases de datos orientadas a objetos y multimedia», *Av. En Sist. E Informática*, vol. 6, n.º 2, pp. 167–184, 2009.

[17] S. E. S. Sánchez López, «Modelo de indexacion de formas en sistemas VIR basado en ontologias», Maestría, Universidad de las Américas Puebla, México, 2007.

[18] 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.

[19] D. A. Pendlebury, «The use and misuse of journal metrics and other citation indicators», *Arch. Immunol. Ther. Exp. (Warsz.)*, vol. 57, n.º 1, pp. 1-11, feb. 2009.

[20] Garfield E, «The history and meaning of the journal impact factor», *JAMA*, vol. 295, n.º 1, pp. 90-93, ene. 2006.

[21] B. Gonzalez-Pereira, V. Guerrero-Bote, y F. Moya-Anegon, «The SJR indicator: A new indicator of journals' scientific prestige», *arXiv:0912.4141*, dic. 2009.

[22] CORE, *CORE Conference Ranking*. Computer Research & Education of Australia, 2008.

[23] J. E. Hirsch, «An index to quantify an individual's scientific research output», *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, n.º 46, pp. 16569-16572, nov. 2005.

[24] L. Egghe, «An improvement of the H-index: the G-index», *ISSI Newsl.*, vol. 2, n.º 1, pp. 8–9, 2006.

[25] B. Jin, «The AR-index: complementing the h-index», *ISSI Newsl.*, vol. 3, n.º 1, p. 6, 2007.

[26] L. Page, S. Brin, R. Motwani, y T. Winograd, «The PageRank Citation Ranking: Bringing Order to the Web.», 11-nov-1999. [En línea]. Disponible en: http://ilpubs.stanford.edu:8090/422/. [Accedido: 05-abr-2014].

[27] 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.

[28] J. A. Olivas, P. J. Garcés, y F. P. Romero, «An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy

synonymy and fuzzy generality for representing concepts in documents», *Int. J. Approx. Reason.*, vol. 34, n.º 2-3, pp. 201-219, nov. 2003.

[29] 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.

[30] N. F. Noy y D. L. Mcguinness, «Ontology Development 101: A Guide to Creating Your First Ontology», 2001.

[31] E. A. Feigenbaum, A. Barr, y P. R. Cohen, *The handbook of artificial intelligence*. Addison-Wesley New York, 1989.

[32] E. Rich y K. Knight, «Artificial intelligence», *McGraw-Hill New*, 1991.

[33] N. J. Nilsson, *Principles of artificial intelligence*. Springer, 1982.

[34] Stanford Center for Biomedical Informatics Research, *Protègè*. Stanford University, 2014.

[35] L. Leydesdorff, F. de Moya-Anegón, y V. P. Guerrero-Bote, «Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI», *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, n.º 2, pp. 352–369, 2010.

[36] M. E. Falagas, V. D. Kouranos, R. Arencibia-Jorge, y D. E. Karageorgopoulos, «Comparison of SCImago journal rank indicator with journal impact factor», *FASEB J.*, vol. 22, n.º 8, pp. 2623-2628, ene. 2008.

[37] C.-T. Zhang, «The e-Index, Complementing the h-Index for Excess Citations», *PLoS ONE*, vol. 4, n.º 5, p. e5429, may 2009.

[38] A. Sidiropoulos, D. Katsaros, y Y. Manolopoulos, «Generalized Hirsch h-index for disclosing latent facts in citation networks», *Scientometrics*, vol. 72, n.º 2, pp. 253–280, 2007.

Horacio D. Kuna es Licenciado en Sistemas egresado de la Universidad de Morón, Master en Ingeniería del Software egresado del ITBA y la Universidad Politécnica de Madrid. Profesor Titular, Director del Departamento de Informática y del Programa de Investigación en Computación de la Fac. De Cs.Exactas Químicas y Nat. De la Universidad Nacional de Misiones. Doctorando de Ingeniería en Sistemas y Computación, Universidad de Málaga – Tesis depositada en 2014. Docente investigador, Facultad de Ingeniería, Universidad Nacional de Itapua, Paraguay.

Rey Martin es Licenciado en Sistemas de Información. Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.

Esteban Martini es Analista en Sistemas de Computación y tesista de la Licenciatura en Sistemas de Información. Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.

Lisandro Solonezen es Analista en Sistemas de Computación y tesista de la Licenciatura en Sistemas de Información. Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.

Lucas Podkowa es tesista de la Licenciatura en Sistemas de Información. Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.