Avances en el 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], A. Rambo[1], L. Podkowa[1]
[1]Departamento de Informática, Fac. de Cs. Exactas, Químicas y Naturales
Universidad Nacional de Misiones
Posadas, Misiones, Argentina
[2]Facultad de Ingeniería – Universidad Nacional de Itapúa, Paraguay
hdkuna@gmail.com

***Resumen* — Para el investigador, la búsqueda de información en la web se ha convertido en una actividad básica en los últimos años. Esto se debe a la gran cantidad de material disponible, situación que conlleva complejidad a la hora de establecer una selección del material a utilizar en base a su calidad. En este contexto, la incorporación en lo cotidiano de herramientas como son los Sistemas de Recuperación de Información puede resultar de utilidad para la optimización del proceso de búsqueda de información en la web, en especial cuando son desarrollados para contextos de aplicación específicos. La constante actualización y mejora de los procesos implementados en este tipo de herramientas de ayuda al investigador, se plantea como una actividad que debe sostenerse en el tiempo. En el presente trabajo se presentan avances en el desarrollo de los componentes principales de un meta-buscador cuyo objetivo es la recuperación de documentos científicos del área de ciencias de la computación. Las mejoras introducidas corresponden al método para la expansión de las consultas ingresadas por los usuarios y al algoritmo de ranking, el cual es utilizado para la valoración de la calidad de cada documento recuperado por el SRI y para el establecimiento del orden de cada resultado en el listado a presentar al usuario.**

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

I. Introducción

La diversidad de contenido de la web y la necesidad selección de información de calidad constituyen exigencias sobre las cuales se sustenta el desarrollo y uso de herramientas tales como los Sistemas de Recuperación de Información (SRI). La investigación científica no se encuentra exenta de este tipo de requerimientos, ya que la complejidad inherente de la actividad se ve incrementada al momento de buscar artículos científicos en la web.

Se reconocen diversas herramientas que permiten hacer frente a tales dificultades, pero aquellas de acceso más general, como son los buscadores comerciales, pueden proporcionar resultados que no cumplan con los parámetros de calidad necesarios para la actividad científica. Y es en este contexto, que la implementación de soluciones para este tipo de actividades en entornos acotados se presenta como una alternativa viable para mejorar la experiencia

El objetivo del presente trabajo es presentar los avances alcanzados en el desarrollo de un Sistema de Recuperación de Información de dominio específico, concretamente un meta-buscador orientado a la recuperación de artículos científicos del área de ciencias de la computación. Se plantean modificaciones al método de expansión de consultas basado en el uso de una ontología y en el algoritmo de ranking que permite ordenar los resultados en base a su calidad.

II. Antecedentes

A. Un meta-buscador para las ciencias de la computación

Un meta-buscador es una variante de un Sistema de Recuperación de Información (SRI) cuya construcción se realiza modularmente, facilitando que sus componentes se adapten a necesidades particulares del dominio en el cual se pretende que funcionen [1], [2]. Esto permite que la búsqueda, recuperación, almacenamiento y gestión de la información a realizar sea completamente adaptada al ámbito para el cual se ha desarrollado el meta-buscador [3].

En trabajos anteriores [4], [5], los autores han presentado una propuesta de SRI para la recuperación de documentos científicos del área de ciencias de la computación, el mismo se planteó como una herramienta que, a través de diversas técnicas y tecnologías relacionadas con la Inteligencia Artificial (AI), posibilita mejorar la calidad de las búsquedas y, consecuentemente, de los resultados a presentar al usuario. Se ha presentado la estructura básica del meta-buscador, conformada a partir de tres módulos que engloban sus funcionalidades básicas, un esquema general del funcionamiento del SRI se puede observar en la figura 1:

- Módulo para la gestión de las consultas: su función reside en realizar la expansión de la consulta ingresada por el usuario a partir del uso del contenido de una ontología de dominio específico correspondiente a una sub-área de las ciencias de la computación [6].
- Módulos para la búsqueda en las bases de datos (buscadores): su función es la de capturar las consultas resultantes del proceso de expansión, adaptarlas y ejecutarlas sobre las bases de documentos integradas al SRI, recuperando sus resultados para luego ser procesados. Actualmente las fuentes sobre las que se realizan las búsquedas son: Google Scholar (GS), ACM Digital Library, IEEE Xplore.
- Módulo para la gestión de los resultados: su función consiste en procesar los resultados obtenidos desde el componente anterior y aplicar sobre los mismos el algoritmo de ranking para evaluarlos y establecer el orden en el cual serán presentados al usuario [7].

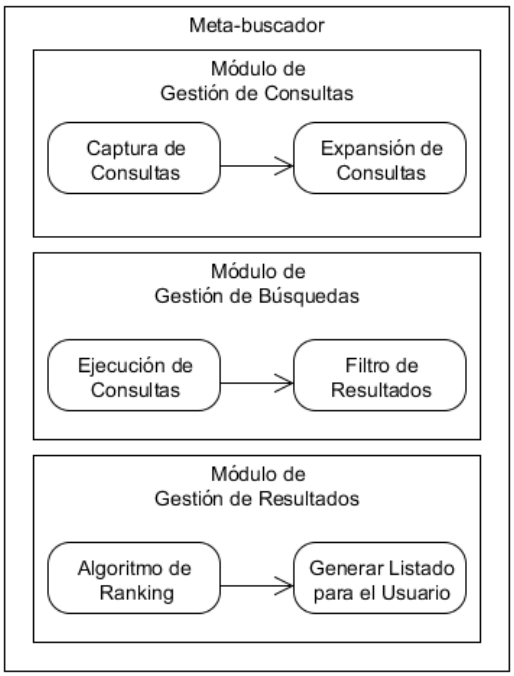

Fig. 1. Esquema general del funcionamiento del meta-buscador desarrollado

Entre los componentes de cada módulo se destacan el método para la expansión de consultas y el algoritmo de ranking para la evaluación de los resultados. En el primer caso, se trata de una operación a través de la cual se analiza el texto de la consulta ingresada por el usuario y se busca en una base de conocimiento, en este caso una ontología de una sub área específica dentro de las ciencias de la computación, un conjunto de términos que permitan ampliar el detalle de la consulta [6]. Estos términos son combinados con los de la consulta original del usuario para generar nuevas consultas, denominadas expansiones, con un mayor nivel de especificidad que favorecen la recuperación de documentos de mayor relevancia para la temática buscada por el usuario [8], [9].

Mientras que en el caso del algoritmo de ranking, se trata de un método que evalúa diferentes propiedades de los documentos científicos recuperados en base a un conjunto de métricas para calcular un valor que permite ordenar el conjunto de resultados a partir de una estimación de la calidad de los mismos [10]. Las propiedades de los documentos evaluadas para el establecimiento del ranking son: la calidad de la fuente de publicación del documento, distinguiendo si la misma es un congreso o una revista científica de la disciplina; la calidad de los autores del documento, medida a través del impacto que los trabajos anteriores de los mismos hayan generado; y finalmente, la calidad del documento en sí mismo, determinada a partir del impacto generado por el mismo desde el momento de su publicación [11], [12].

Ambos componentes han presentado diversos inconvenientes que hacen necesaria una revisión de su definición, por ejemplo: cómo debe reaccionar el SRI a consultas generadas en un idioma con el cual no operen los repositorios documentales a los que se accede al realizar las búsquedas; cómo se debe tratar la ocurrencia de diferentes valores para una métrica en particular de un objeto basados en la consulta a bases de datos diferentes, entre otros. El tratamiento y propuesta de solución para estos inconvenientes se discuten a continuación.

B. Tratamiento de lenguaje natural en un SRI

Entre los principales desafíos que se plantean al momento del desarrollo de un SRI se encuentra el tratamiento que se debe hacer del lenguaje en el que el usuario ingresa las consultas [13]. Esto se manifiesta principalmente cuando el usuario desea ejecutar una consulta en un idioma diferente al correspondiente a los documentos a los que accede el SRI. En el caso particular del presente trabajo, el tratamiento del lenguaje debe realizarse para asegurar que la petición del usuario sea ejecutada correctamente en las diferentes bases de documentos sobre las que realiza las búsquedas el sistema, indiferentemente del idioma en el que se hayan ingresado.

Entre las alternativas para la gestión del lenguaje de las consultas en SRI [14], aquellas que realizan una detección y traducción automática del texto requerido son las que presentan mayor aplicabilidad dados los tiempos de respuesta requeridos. Entre las estrategias más utilizadas se encuentran [15]:

- Emparejamiento de términos entre consultas y documentos sin traducción
- Traducción de las consultas al idioma de los documentos
- Traducción de los documentos al idioma de las consultas
- Traducción de los documentos y las consultas a un lenguaje común

En el contexto de este trabajo en particular, dado que el tipo de SRI desarrollado es un meta-buscador, y que el mismo no realiza un tratamiento interno de los documentos recuperados de las diferentes fuentes, se ha optado por la alternativa que plantea la traducción de las consultas al idioma de los documentos disponibles en cada fuente a examinar [15]. En línea con esta determinación se han limitado los idiomas a considerar en la operación del SRI al castellano y el inglés. De esta manera, se solucionan algunos de los problemas en la búsqueda de documentos, por ejemplo: bases de datos que no admiten la ejecución de consultas en idioma castellano, no indexan contenido en otro idioma que no sea el inglés, o no ejecutan ningún tratamiento sobre el lenguaje de las consultas ingresadas, limitando el espectro de la recuperación de documentos más allá de que se encuentren indexados ítems en ambos idiomas.

Dada esta situación se plantea una modificación en el método de expansión de consultas inicialmente desarrollado para el SRI [6]: en un primer paso se reconoce el lenguaje en el cual fue escrita la consulta, luego se aplica el uso de la ontología para poder hacer la expansión y se obtienen expansiones tanto en castellano como en inglés. De esta manera se cuenta con un conjunto ampliado de consultas que permite la búsqueda en todas las bases documentales a las que accede el meta-buscador, sin importar el idioma original de la consulta del usuario.

C. Evaluación de calidad de documentos científicos

Uno de los componentes principales del meta-buscador es el algoritmo de ranking que se utiliza para analizar, ponderar y ordenar el listado de documentos que se presenta al usuario como resultado del procesado de su consulta [3], [4]. Al tratarse de un SRI que opera con documentos científicos, el algoritmo de ranking debió ser desarrollado en forma particular, utilizando distintas métricas que permitieran la evaluación de una publicación científica. Para ello, en primera instancia se seleccionaron aquellos atributos de los documentos que serían involucrados en la evaluación, siendo elegidos [11], [12]:

- 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 realizar la evaluación de un artículo científico comenzando por la calidad de la fuente en donde se haya publicado, entendiendo por la misma a la revista científica o el congreso o evento similar de la disciplina donde el o los autores hayan presentado el documento, determinando su calidad en base al impacto generado por sus números anteriores; la calidad de los autores, que se puede determinar a partir de la relevancia que hayan obtenido publicaciones previas de los mismos; y por último, la calidad de la publicación que se puede medir a partir del impacto en la comunidad de pares que haya generado el documento desde el momento de su publicación.

Posteriormente se relevaron y seleccionaron métricas para evaluar cada una de las características antes mencionadas. Una vez finalizada la selección, se obtuvieron valores asociados a las mismas desde diferentes bases de datos y se desarrolló el algoritmo que otorga una valoración a cada documento científico que haya sido recuperado durante la etapa de búsqueda del SRI.

La versión original del algoritmo [10] se planteó en base a las métricas con el mayor grado de uso en los últimos años. En el presente trabajo se ha ampliado el relevamiento realizado con el objetivo de incorporar un nuevo conjunto de indicadores que permitan subsanar las dificultades de los originalmente seleccionados, ampliar el espectro de cobertura y mejorar la calidad del algoritmo de ranking del meta-buscador. El resultado de este nuevo relevamiento se puede observar en la tabla I, comparando el conjunto con aquellas de la primera selección.

Es por esto que, en el apartado de métricas para evaluar la calidad de la fuente de publicación para la evaluación de revistas científicas, a los índices IF (Impact Factor) [16] y SJR (SCImago Journal Rank) [17] se agregan los indicadores SNIP (Source Normalized Impact per Paper), RIP (Raw Impact per Paper) [18], EI (Eigenfactor) y AI (Article Influence) [19]. Además se incorporan métricas que, habiendo sido definidas para la evaluación de otro aspecto de una publicación, se han aplicado sobre diferentes bases de datos, generando una implementación para la evaluación de fuentes de publicación, por ejemplo: el índice H publicado por MAS (MicroSoft Academic Search) [20]. En la mayoría de los indicadores relevados, el modo de funcionamiento se basa en la cantidad de citas que reciben los artículos que son publicados por las revistas en una ventana de tiempo variable según la definición de la métrica. Para el caso de la evaluación de una fuente de publicación como lo son los congresos o eventos de la disciplina previamente se utilizaba en forma unitaria el Ranking CORE [21], como resultado del relevamiento se agrega inicialmente el Ranking ERA [22] y, al igual que sucediera con métricas para la evaluación de revistas, se incorporan implementaciones de otros indicadores que fueron modificados para la evaluación de congresos como son: el índice H generado por MAS [23] y el IF publicado por CiteSeerX utilizando información propia y del repositorio DBLP [24].

De igual manera, el conjunto base de métricas a utilizar para evaluar la calidad de un autor ha pasado de estar compuesto solamente por el índice H a estar integrado por otras métricas similares, pero cuyo objetivo es la solución o mejora de alguna característica particular de tal indicador, entre las numerosas variantes [25] disponibles se han seleccionado los índices G [26], W [27] y E [28].

Finalmente, para la evaluación de la calidad de un artículo científico se ha propuesto continuar con el mismo esquema planteado en la versión original del algoritmo de ranking, el mismo está planteado a partir de una modificación del índice AR [29], utilizando la cantidad de citas de un documento en particular.

TABLA I. MÉTRICAS ANALIZADAS PARA LA EVALUACIÓN DE ARTÍCULOS CIENTÍFICOS

Propiedad a evaluar		**Métricas originales**	**Nuevas métricas relevadas**
Calidad de la fuente de publicación	Publicación en revista científica	IF	
		SJR	
			SNIP
			RIP
			EI
			AI
			Índice H
	Publicación en Congreso o Evento Científico	Ranking CORE	
			Ranking ERA
			Índice H
			IF
Calidad de los autores		Índice H	
			Índice W
			Índice G
			Índice E
Calidad del artículo		Índice AR	
		Cantidad de citas	

III. MATERIALES Y MÉTODOS

A. Método para la identificación del lenguaje de las consultas

El método seleccionado para la identificación del lenguaje de las consultas que ingresa el usuario al meta-buscador se basa en una medida de similitud entre el texto ingresado y uno, almacenado internamente, que sirve como referencia para cada idioma que se pretende detectar, castellano e inglés en este caso. Esta estrategia se reconoce en otras publicaciones relacionadas con la temática [30], [31] y requiere de la aplicación de un método a través del cual se representen los textos a comparar y de una función que permita medir la similitud entre las representaciones que sean obtenidas a fin de establecer el idioma inicialmente detectado. En el caso del presente trabajo se ha optado por el modelo vectorial y la ecuación del coseno del ángulo entre los vectores a generar para ser empleados como método de representación y función de similitud, respectivamente.

El modelo de representación vectorial se eligió como alternativa para la representación tanto del texto de la consulta del usuario como de los textos de referencia de cada idioma que almacena el SRI. Este modelo, utilizado en otras publicaciones del área de recuperación de información [3], [32], se basa en que cada documento se representa a través de un vector en un espacio de n dimensiones, guardando en cada posición del vector un valor numérico que guarda relación con la proporción de ocurrencias de diversos componentes del documento. Tales componentes son los que definen la dimensión del vector, ya que si se almacena la frecuencia de unigramas se estaría hablando de un espacio uni-dimensional, para 2-gramas uno bi-dimensional y así sucesivamente. En el contexto del presente trabajo se ha optado por un vector de 2 dimensiones, por lo que en cada una de sus posiciones se

almacenará la frecuencia de ocurrencia de n-gramas, siendo n = 2; de esta manera en la posición i-ésima del vector se tendrá la frecuencia de aparición en el texto del i-ésimo n-grama del texto a ser representado [3].

Una vez seleccionado el método para la representación de los documentos se debió contar con una métrica que permitiera establecer a cuál de los textos de referencia se podía considerar más similar al texto ingresado a modo de consulta por parte del usuario. Considerando el modelo de representación elegido, se optó por utilizar la fórmula del coseno, ver ecuación 1, del ángulo entre los vectores a ser generados.

$$\cos \theta = d1 * d2 / |d1| |d2| \quad (1)$$

Los componentes de la ecuación se describen de la siguiente manera: d1 y d2 son los vectores correspondientes al texto de la consulta y a cada uno de los textos de referencia de cada idioma, el * representa el producto escalar entre los vectores y |d| representa el módulo del vector d, aplicable para cada uno de los vectores. Nótese que la ecuación será aplicada una vez para cada idioma, castellano e inglés, que se desee detectar, para luego comparar los valores obtenidos y determinar a qué texto, e idioma, se corresponde la consulta ingresada por el usuario.

B. Evolución del método de expansión de consultas

La evolución del método para realizar la expansión de consultas, consistió en dividirlo en dos etapas, por un lado una instancia inicial de detección del idioma en el que fuera ingresada la consulta del usuario al SRI y posteriormente la búsqueda de los términos en la ontología correspondiente para generar las expansiones. Cabe destacar que el método en su nueva versión cuenta con dos ontologías idénticas pero inversas, idénticas porque conceptualmente contienen el mismo contenido, inversas porque una se encuentra en idioma castellano y otra en idioma inglés. Además ambas ontologías cuentan con una nueva propiedad en cada concepto contenido que permite obtener la traducción del mismo al idioma opuesto al que sea general de la ontología.

La primera etapa del proceso de expansión, ver figura 2, está basada en el método de detección del lenguaje de la consulta ingresada por el usuario, en adelante "*consulta_original*", descripto en la sección anterior. Una vez determinado el idioma de la consulta, se procede a instanciar la ontología correspondiente para dar inicio a la siguiente etapa del proceso.

La segunda etapa del proceso mantiene la base de las publicaciones anteriores de los autores, es decir, su objetivo es buscar dentro de la ontología el o los conceptos que guarden mayor similitud con respecto a la *consulta_original*, y posteriormente utilizar las relaciones y propiedades definidas para los mismos en la ontología a fin de obtener los términos con los cuales se procede a generar el conjunto de expansiones que ejecuta el SRI sobre las diferentes bases documentales. El procedimiento de búsqueda en la ontología del concepto más similar a la *consulta_original* se descompone en los siguientes pasos:

1) Para todos los términos de la *consulta_original*: se recorre la ontología en todos sus elementos, clases e instancias, para buscar el o los conceptos con la mayor similitud con el término, almacenando cada resultado en una colección auxiliar.

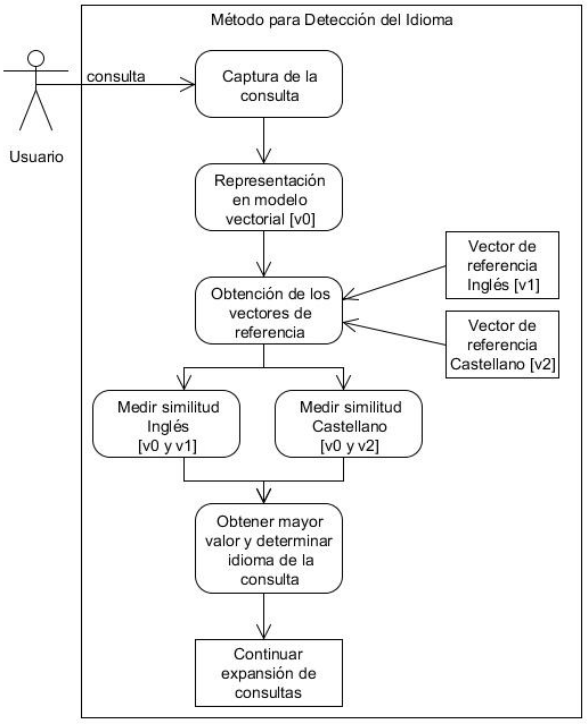

Fig. 2. Método para la detección del idioma de las consultas

2) Se examina la colección resultante del paso anterior:
   a) Si está vacía, finaliza el proceso de expansión sin resultados.
   b) Si contiene un único elemento, éste se denomina "*término_candidato*" y será utilizado para generar las expansiones.
   c) Si contiene n elementos se busca seleccionar al candidato sobre la base de las coincidencias con la *consulta_original*. En caso de que sea un conjunto de elementos los seleccionados, se procede de la siguiente manera, utilizando las relaciones definidas en la ontología:
      i. Si los conceptos comparten a su clase "padre", la misma se establece como *término_candidato*.
      ii. Si no comparten clase "padre" entonces se analizan tales conceptos, aquel que referencie una mayor cantidad de instancias será el *término_candidato*.
      iii. En el caso de que todos los "padres" involucrados referencien igual cantidad de elementos, se genera una colección de *términos_candidatos* con todos ellos.

Al finalizar la búsqueda se cuenta con uno o varios *términos_candidatos* para generar las expansiones, el proceso para hacerlo utiliza las relaciones y propiedades definidas en la ontología, y se compone de los siguientes pasos:

1. Se extrae el concepto "padre" del *término_candidato* y se lo denomina "concepto_padre".
2. Se extraen los "hermanos" del *término_candidato*, es decir, los conceptos de su mismo nivel, y se pasan a almacenar en una nueva colección denominada "*conceptos_hermanos*[]".
3. Se extraen los sinónimos, en caso de existir, del *término_candidato*, y se genera otra colección denominada "*sinónimos_concepto*[]".

4. Se extrae, utilizando la propiedad definida para esta nueva versión de la ontología, la traducción de los conceptos extraídos en los pasos anteriores, y se genera una nueva colección denominada "*traducciones*[]".
5. Se generan las expansiones de la consulta del usuario, para ello se utilizan los elementos obtenidos en los pasos anteriores, iniciando por las expansiones que son realizadas en el mismo idioma de la consulta original del usuario:
   - Expansión_1 = consulta_original AND término_candidato
   - Expansión_2 = término_candidato AND concepto_padre
   - Expansión_3 = término_candidato OR conceptos_hermanos[]
   - Expansión_4 = término_candidato OR sinónimos_concepto[]

Posteriormente se generan las expansiones correspondientes al segundo de los idiomas considerados:

- Expansión_traducida_1 = traducciones[candidato] AND traducciones[padre]
- Expansión_traducida_2 = traducciones[candidato] OR traducciones[hermanos]
- Expansión_traducida_3 = traducciones[candidato] OR traducciones[sinónimos]

Finalizado el proceso de expansión de consultas, se han generado diversas consultas alternativas, complementarias a la originalmente ingresada por el usuario, con el objetivo de mejorar la recuperación de documentos relevantes para la misma a través de términos relacionados con la temática de la búsqueda que fueran propios de la disciplina. El resultado, además, se presenta como solución para los problemas de búsqueda en diferentes bases documentales al contar con un conjunto de expansiones que permiten ejecutar todas las búsquedas sin importar el idioma base de la consulta y los requerimientos impuestos por cada fuente.

C. Modelo conceptual para la evaluación de documentos científicos

La evaluación de publicaciones científicas es un ámbito en el cual convive un gran número de métricas para la evaluación de diferentes características de un documento. En este contexto la gran mayoría de esas métricas acumulan tanto adeptos como detractores, sin embargo, existen iniciativas tendientes a establecer que la evaluación de la producción científica no puede realizarse desde un único indicador sino que debe ser valorada desde diferentes dimensiones [12], [18].

Entre los inconvenientes que se presentan al momento de seleccionar y utilizar indicadores bibliométricos para evaluar un documento científico se pueden mencionar:

- Determinar si utilizar uno u otro indicador resultará en un resultado similar, esto se presenta dada la correlación que existe entre indicadores de revistas científicas [33]. Fenómeno que se presenta de diferente manera a medida que varían las áreas de conocimiento, haciendo que la elección de indicadores a utilizar deba contemplar las características propias del área en la que haya sido generada la publicación [34].
- Determinar el valor "real" de una métrica, esto se debe a que si bien el método para calcular el valor de un determinado indicador es fijo, los datos en base a los que se calcula pueden variar de una base de datos a otra en función de su capacidad de indexación. Por lo tanto, un mismo objeto de estudio podría obtener valores diferentes de una misma métrica según la base de datos que sea utilizada [35].
- Determinar el grado de solapamiento que existe entre las bases de datos utilizadas para el cálculo de indicadores. Las grandes bases de datos, como Scopus de Elsevier o la versión Scholar de Google, se encuentran en constante expansión, lo que dificulta determinar la exactitud con la que ambas bases se solapan y por tal motivo no es recomendable determinar la calidad de alguna propiedad de una publicación científica en base a un único indicador [36].

Considerando estas situaciones, se determinó refactorizar el modelo utilizado por el algoritmo de ranking para la evaluación de los documentos científicos que fueran recuperados por el SRI. En este sentido se empleó el mismo concepto básico de un meta-buscador, manteniendo las propiedades a evaluar antes determinadas, pero integrando un mayor número de métricas que pudieran ser obtenidas desde más de una base de datos y utilizadas, en caso de ser posible, para la evaluación de más de una característica de un documento particular.

De esta manera se plantea una solución para los problemas mencionados previamente, ya que el modelo generado, ver figura 3, a través de un enfoque integrador permite la evaluación de las propiedades de un artículo científico sin depender de una métrica en especial y utilizando datos provenientes de diversas fuentes. Además, se ha mejorado la escalabilidad del modelo para su expansión ya que la integración de nuevas métricas podrá realizarse sin necesidad de modificar el modelo; por otra parte el modelo también es más robusto ya que incorpora más métricas y fuentes desde donde son calculadas las mismas, generando una evaluación de mayor calidad para cumplir con los objetivos del SRI.

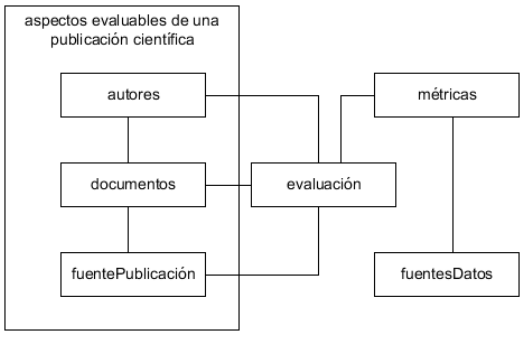

Fig. 3. Modelo conceptual para la evaluación de documentos científicos

D. Evolución del algoritmo de ranking

Con el modelo de evaluación en su nueva versión, se debieron realizar ajustes en los cálculos que definen la calificación de una determinada publicación *Q*. El valor final a asignar al resultado se compone de tres factores: *FP* (fuente de publicación), *A* (autores) y *D* (documento); donde cada parámetro es la representación de la evaluación de cada una de las propiedades inicialmente seleccionadas. La fórmula principal se completa con la multiplicación de cada uno de los factores por un valor de ajuste α, β y γ, ver ecuación 2, que son

utilizados para variar el peso de los factores de evaluación en base a la importancia que se desee establecer para cada uno de ellos.

$$Q = \alpha * FP + \beta * A + \gamma * D \quad (2)$$

El cálculo de cada parámetro *P*, en forma genérica, es realizado internamente a partir del número de métricas que se encuentren disponibles al momento de la evaluación de la publicación en cuestión. De esta manera, cada parámetro se calcula a partir de la sumatoria de todas las métricas *m* que se encuentren para la publicación, ver ecuación 3, y este valor se divide por la cantidad de métricas que hayan sido empleadas en el cálculo.

$$P = (\Sigma_{1\text{-}n}\, m) / n \quad (3)$$

Debido a que los rangos de valores de las métricas varían de una a otra, se debieron incorporar métodos de normalización para poder integrar todas las métricas dentro de una misma fórmula y que la misma resulte equilibrada. Considerando las características de las métricas relevadas se propusieron dos tipos de normalización, dependiente justamente del tipo de métrica. En primer lugar se utiliza como normalizador el mayor valor registrado en la base de datos de la que se haya obtenido la métrica, a través del cociente entre el valor de la misma y ese máximo valor. Este tipo de normalización es utilizado en las métricas implicadas en el cálculo de los factores *FP* y *A*. En las métricas utilizadas para la obtención del valor del parámetro *D* se utiliza como método de normalización el logaritmo en base 10 del valor de la métrica. Ambos métodos de normalización pueden observarse en las ecuaciones 4 y 5 respectivamente.

$$m = m / max(m, BDm) \quad (4)$$

$$m = log_{10}(m) \quad (5)$$

En todos los casos *m* hace referencia al valor de la métrica a normalizar y el resultante de la normalización, *max(m, BDm)* hace referencia al valor máximo para la métrica *m* obtenido desde la misma base de datos de su origen *BDm*.

E. Implementación del algoritmo de ranking

Una vez finalizado el diseño del modelo de evaluación y modificadas las fórmulas a utilizar para el cálculo del algoritmo se comenzó la implementación del mismo para ser incorporado al meta-buscador, concretamente al módulo para la gestión de resultados. En el proceso de implementación se comenzó por seleccionar aquellas métricas a integrar al modelo y posteriormente se capturaron sus valores desde diversas fuentes.

En base al relevamiento descripto en la sección II.C, se han detectado métricas aplicables para la evaluación de las diferentes propiedades de una publicación científica. Sin embargo, se ha podido observar que la cantidad de datos a recolectar para el cálculo de las métricas hace muy difícil su implementación en forma interna en el SRI en desarrollo. Por lo tanto, se ha optado por utilizar los valores ya calculados de distintos indicadores que fueran basados en fuentes de datos reconocidas como son Scopus, GS (Google Scholar), ISI (Institute for Scientific Information) y DBLP, entre otras.

Identificadas las fuentes se procedió a elegir aquellas implementaciones de métricas a incorporar al meta-buscador, siendo seleccionadas: diferentes versiones de una misma métrica, siempre que fueran calculadas a partir de un origen de datos diferente y métricas cuyo planteo teórico fue reformulado para ser aplicadas a la evaluación de otra propiedad de un documento, variando los datos de origen, por ejemplo: el índice H para revistas.

Posteriormente se debió evaluar la factibilidad de desarrollar un método que permitiera cargar los valores de tales métricas a una base de datos interna del meta-buscador que sea utilizada para el cálculo del algoritmo de ranking. Como resultado se desarrollaron diversos componentes de software que automatizan la extracción, transformación a un formato homogéneo basado en el modelo de evaluación previamente descripto y carga en la base de datos. En aquellos casos en los cuales los valores para el cálculo de la métrica a utilizar se pueden extraer a partir de los meta-datos que obtiene el SRI se ha optado por calcularlas a medida que se procesan internamente los resultados. El conjunto de indicadores y sus respectivas fuentes de datos integradas a la presente versión del algoritmo de ranking puede observarse en la tabla II.

TABLA II. MÉTRICAS INCORPORADAS AL ALGORITMO DE RANKING

Propiedad a evaluar		**Métrica**	**Origen de los datos**
Calidad de la fuente de publicación	Publicación en revista científica	SJR	Scopus
		RIP	Scopus
		SNIP	Scopus
		Índice H	Scopus
		AI	ISI
		EI	ISI
		EI	MAS
		Índice H	MAS
	Publicación en Congreso o Evento Científico	CORE	CORE
		ERA	ERA
		IF	CiteSeerX + DBLP
		Índice H	MAS
Calidad de los autores		Índice H	ArnetMiner
		Índice G	ArnetMiner
		Índice H	GS
Calidad del artículo		Índice AR	(*)
		Cantidad de citas	

(*). Se utilizan los meta-datos que obtiene el SRI en las búsquedas

Las herramientas software utilizadas en las tareas de este proceso fueron: los lenguajes Java, HTML, XML y JSON para la extracción de contenido desde la web y el módulo de integración de datos de Pentaho (Pentaho Data Integration) junto al motor de bases de datos PostgreSQL para la transformación y carga de los datos a la base interna del meta-buscador.

IV. EXPERIMENTACIÓN

A. Validación del método de expansión de consultas

Una vez concluido el desarrollo de las modificaciones introducidas en el método para la expansión de las consultas del SRI, se debió proceder con la validación del funcionamiento del mismo. Este proceso se basó en la colaboración de un grupo de expertos en el área de IA, quienes debieron determinar si la expansión generada a través de la ontología permitiría obtener un conjunto de resultados de mayor relevancia para el usuario del meta-buscador.

La experimentación, concretamente ha consistido en la ejecución del proceso de expansión sobre un grupo de consultas, cuyos resultados tanto finales como intermedios fueron presentados a los expertos con el objetivo de que se pudiera evaluar tanto la construcción de las expansiones como el resultado final del proceso. En ese sentido, cada experto ha determinado un valor entre 1 y 10 como medida de calidad de la expansión. Los resultados de este proceso pueden observarse en la tabla III.

En general los resultados de la validación han sido positivos, con la salvedad de que en algunos casos los expertos han advertido que la inclusión en la consulta de términos que pertenezcan al primer nivel de la ontología podrían generar expansiones de una amplitud excesiva, lo que podría generar la recuperación de documentos muy generales y de baja relevancia para el usuario.

TABLA III. RESULTADOS DE LA VALIDACIÓN DEL MÉTODO DE EXPANSIÓN DE CONSULTAS REALIZADA POR LOS EXPERTOS

Consulta realizada	Efectividad promedio
agentes inteligentes AND recuperación de información	6.2
search methods AND deep first search	7.4
unsupervised learning AND backpropagation networks	6.8
algoritmos genéticos OR algoritmos evolutivos	7
fuzzy sets AND expert systems	8.2

Esta situación se presentaría ante la posibilidad de que alguno de los conceptos registre como padre a la raíz de la ontología, motivo por el cual se han diseñado reglas para el proceso de expansión que contemplan esta situación y cuya aplicación evitaría los mencionados inconvenientes. En una futura versión del método se espera contar con estas mejoras integradas al SRI.

B. Validación del algoritmo desarrollado

Con los cambios al algoritmo de ranking ya implementados, se debió comenzar el proceso de su validación. La variación con las evaluaciones realizadas previamente [4], [10] consistió en que el planteo esta vez fue en dos etapas, una inicialmente realizada por expertos y otra en la que se evaluara estadísticamente a los resultados considerando a la correlación entre las diferentes métricas como una medida de desempeño.

Para el caso de la instancia de validación con expertos se ha abordado un esquema similar al utilizado en ocasiones anteriores. Se han ejecutado consultas utilizando el meta-buscador y se han exportado los resultados junto al detalle de los cálculos que fueron aplicados para determinar el ranking de cada uno de los resultados presentados. Estos datos han sido valorados por los expertos en forma conjunta con el listado de documentos recuperados, presentando el mismo orden que sería presentado al usuario final, otorgando como resultado un porcentaje de la efectividad en la clasificación que se otorgó a cada publicación científica del listado. Con el objetivo de comparar esta evolución con la versión inicial del algoritmo de ranking, se han realizado las mismas consultas y se obtuvo la diferencia en la medida de efectividad determinada por los expertos. El resultado de esta etapa de validación puede observarse en la tabla IV.

TABLA IV. (A) RESULTADOS DE LA VALIDACIÓN DEL ALGORITMO DE RANKING POR PARTE DE LOS EXPERTOS

Consulta realizada	Cantidad de resultados procesados	Efectividad evaluada por los expertos	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%

En lo que respecta a la segunda fase de la evaluación, se ha planteado que el foco debería estar en uno de los problemas encontrados al momento del planteo del modelo de evaluación para documentos científicos, que es la correlación que puede haber entre las diferentes métricas que se utilizan. Un ejemplo de esta situación podría darse entre la cantidad de citas que registra un artículo en particular y el índice SJR de la revista en la cual el mismo ha sido publicado. En esta segunda instancia de validación, el planteo busca determinar si la integración en el algoritmo de un conjunto más grande de métricas heterogéneas ha resultado en que la evaluación generada sobre los resultados de las búsquedas no se encuentra sesgada por un indicador en particular.

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

V. CONCLUSIONES

La generación de herramientas para la recuperación de información desde la web constituye una tarea en la que ha cobrado gran importancia la eficiencia en las búsquedas a fin de obtener los resultados de mayor relevancia posible para el usuario. En este trabajo se han presentado avances en el desarrollo de los componentes de un SRI, en particular un meta-buscador, cuyo aplicación se ámbito de aplicación son las ciencias de la computación y que opera para la recuperación de un tipo particular de documentos como son las publicaciones científicas.

Se han presentado modificaciones en el método que realiza la expansión de la consulta del usuario, concretamente incorporando al mismo la capacidad para detectar el idioma de la consulta ingresada, modificar el proceso de expansión y generar expansiones tanto en castellano como en inglés. Con esta evolución se permite al SRI obtener un conjunto de resultados de mayor relevancia en base a los requisitos del usuario. Por otra parte, se ha mejorado el algoritmo de ranking utilizado para la calificación y ordenación de los resultados a presentar al usuario. En este caso se ha ampliado el número de métricas, conjuntamente a la cantidad de bases de datos a las que se accede para obtener los valores de tales métricas y se han modificado los métodos para el cálculo del impacto de cada documento recuperado. Las mejoras introducidas en ambos componentes ha permitido presentar soluciones a los problemas detectados en trabajos anteriores, incrementando la calidad del meta-buscador en desarrollo.

Como trabajos a futuro se pueden mencionar: avanzar en el desarrollo del método de expansión de consultas en lo que respecta a las recomendaciones de los expertos surgidas en la actividad de validación; completar la siguiente instancia de evaluación del comportamiento del algoritmo de ranking; determinar la necesidad y, de corresponder, la factibilidad de incorporar métodos que permitan la adaptación de la evaluación de los documentos procesados por el algoritmo de ranking en base al sub-área temática en la que hayan sido presentados, incorporar elementos que permitan la generación de una evaluación más completa del perfil de un autor; entre otros.

REFERENCIAS

[1] [1] G. Salton y M. Mcgill, Introduction to Modern Information Retrieval. McGraw-Hill, Inc., 1983.

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

[3] [3] 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.

[4] [4] H. Kuna, M. Rey, E. Martini, L. Solonezen, y L. Podkowa, «Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación», Rev. Latinoam. Ing. Softw., vol. 2, n.o 2, pp. 107-114, 2013.

[5] [5] 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.

[6] [6] H. Kuna, M. Rey, L. Podkowa, E. Martini, y L. Solonezen, «Expansión de Consultas Basada en Ontologías para un Sistema de Recuperación de Información», presentado en XVI Workshop de Investigadores en Ciencias de la Computación, 2014.

[7] [7] H. Kuna, M. Rey, J. Cortes, E. Martini, y L. Solonezen, «Generating a Ranking Algorithm for Scientific Documents in the Computing Science Area», en XIX Argentine Congress of Computer Science Selected Papers, La Plata, Buenos Aires, Argentina: EDULP, 2014, pp. 185-195.

[8] [8] 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.

[9] [9] 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.

[10] [10] 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.

[11] [11] 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.

[12] [12] 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.

[13] [13] L. Ballesteros y W. B. Croft, «Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval», en Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1997, pp. 84–91.

[14] [14] D. W. Oard, «A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval», en Machine Translation and the Information Soup, D. Farwell, L. Gerber, y E. Hovy, Eds. Springer Berlin Heidelberg, 1998, pp. 472-483.

[15] [15] D. W. Oard, «Alternative approaches for cross-language text retrieval», en AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, 1997, vol. 16.

[16] [16] Garfield E, «The history and meaning of the journal impact factor», JAMA, vol. 295, n.o 1, pp. 90-93, ene. 2006.

[17] [17] 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.

[18] [18] H. F. Moed, «Measuring contextual citation impact of scientific journals», J. Informetr., vol. 4, n.o 3, pp. 265–277, 2010.

[19] [19] C. Bergstrom, «Measuring the value and prestige of scholarly journals», Coll Res Libr News, vol. 68, n.o 5, p. 3146, 2007.

[20] [20] Microsoft Academic Search, Help Center. 2014.

[21] [21] CORE, CORE Conference Ranking. Computer Research & Education of Australia, 2008.

[22] [22] The Australian Research Council, ERA 2012 Journal and Conference Lists. 2012.

[23] [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] [24] CiteSeerX, Venue Impact Factors. 2014.

[25] [25] R. Van Noorden, «Metrics: A profusion of measures.», Nature, vol. 465, n.o 7300, pp. 864–866, 2010.

[26] [26] L. Egghe, «Theory and practise of the g-index», Scientometrics, vol. 69, n.o 1, pp. 131-152, abr. 2006.

[27] [27] Q. Wu, «The w-index: A significant improvement of the h-index», J. Am. Soc. Inf. Sci. Technol., p. n/a-n/a, 2009.

[28] [28] C.-T. Zhang, «The e-Index, Complementing the h-Index for Excess Citations», PLoS ONE, vol. 4, n.o 5, p. e5429, may 2009.

[29] [29] B. Jin, «The AR-index: complementing the h-index», ISSI Newsl., vol. 3, n.o 1, p. 6, 2007.

[30] [30] S. Bastrup y C. Pöpper, «Language detection based on unigram analysis and decision trees», Proj. 2003, p. 27, 2003.

[31] [31] G. Russell, G. Lapalme, y P. Plamondon, «Automatic Identification of Language and Encoding», Rapp. Sci. Lab. Rech. Appliquée En Linguist. -Form. RALI Univ. Montr. Can., pp. 7–2003, 2003.

[32] [32] C. D. Manning, P. Raghavan, y H. Schütze, Introduction to information retrieval, vol. 1. Cambridge university press Cambridge, 2008.

[33] [33] D. Torres-Salinas y E. Jiménez-Contreras, «Introducción y estudio comparativo de los nuevos indicadores de citación sobre revistas científicas en Journal Citation Reports y Scopus», El Prof. Inf., vol. 19, n.o 2, pp. 201–208, 2010.

[34] [34] J. Ewing, «Measuring journals», Not.-Am. Math. Soc., vol. 53, n.o 9, p. 1049, 2006.

[35] [35] J. Bar-Ilan, «Which h-index? — A comparison of WoS, Scopus and Google Scholar», Scientometrics, vol. 74, n.o 2, pp. 257-271, nov. 2007.

[36] [36] M.-A. Sicilia, S. Sánchez-Alonso, y E. García-Barriocanal, «Comparing impact factors from two different citation databases: the case of Computer Science», J. Informetr., vol. 5, n.o 4, pp. 698–704, 2011.

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 y Doctor de Ingeniería en Sistemas y Computación por la Universidad de Málaga, España. Profesor Titular, Co-Director del Departamento de Informática y Director del Programa de Investigación en Computación de la Fac. De Cs.Exactas Químicas y Nat. De la Universidad Nacional de Misiones. 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.

Alice Rambo es Ingeniera en Informática y Profesora de Informática, Concejera departamental del Departamento de Informática, Co-Directora de las carreras Licenciatura en Sistemas de Información, Analista en Sistemas de Computación y Profesorado Universitario en Computación. Profesor Titular y Personal de Investigación interviniente en Proyectos de Investigación en la Fac. De Cs. Exactas Químicas y Nat. De la Universidad Nacional de Misiones. Realizó Estancia de Postgrado e Investigación en la Universidad de Málaga, Departamento de Lenguajes y Ciencias de la Computación.

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.