

Desarrollo de un Sistema de Recuperación de Información para Documentos Científicos del Área de Ciencias de la Computación

Emanuel García Pérez

August 29, 2014

1 CONCEPTOS & DEFINICIONES

- SRI para documentos científicos
- Expansión de la consultas en un SRI y ontologías
- Métricas para la evaluación de documentos científicos

2 RECURSOS & DISEÑO

- Estructura & Funcionamiento del SRI
- Expansión de las consultas
- Algoritmo de ranking

3 DESARROLLO & APLICACIÓN

- SRI: Desarrollo del prototipo
- SRI: Validación del prototipo

4 CONCLUSIONES

5 BIBLIOGRAFÍA

Sistema de Recuperación de Información

Un Sistema de Recuperación de Información(SRI) es un proceso que posee capacidad para recuperar, almacenar y mantener información para distintos fines, según el contexto de su aplicación.

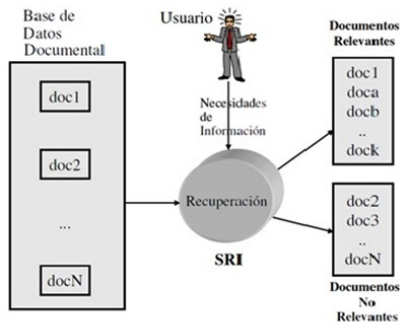


Figura 1. (Proceso de recuperación de información)

Organización de un SRI

- Documentos: Fuente de información sobre la cual se pretende realizar búsquedas.
- Consultas: Generadas por los usuarios del SRI que tienen por objetivo recuperar la información a la cual el sistema provee acceso.
- Representación de los Documentos: Consultas y las relaciones que se definan entre ellas que sean definidas teniendo en cuenta el ámbito de aplicación del SRI.
- Función de Evaluación: Determina la pertinencia de cada documento recuperado para dar solución a la consulta del usuario.

Tipos de SRI

Los principales tipos de SRI que actualmente operan en Internet son: directorios(Yahoo), buscadores(Google), y meta-buscadores(Ixquick). Entre estos destacan los meta-buscadores, debido a que su modularidad permite que los componentes del SRI sean desarrollados para cubrir las necesidades establecidas para una implementación particular.



Implementación particular de SRI

Para desarrollar el SRI requerido por el presente trabajo, se considerarán los siguientes componentes:

- 1 El componente que captura la consulta de usuario y la expande, generando consultas similares para expandir el espectro de búsqueda.
- 2 El componente que accede a las fuentes de datos y recupera de cada una de ellas los documentos resultantes de la ejecución de una consulta.
- 3 El componente que aplica la función de evaluación a los documentos obtenidos de cada fuente para ordenar el listado final para el usuario.

SRI para documentos científicos de Ciencias de la Computación (I)

Existen iniciativas para la generación de SRI de proposito específico en áreas particulares, pero no se encontró evidencia de implementaciones que sean aplicadas a bases de datos de documentos científicos de Ciencias de la Computación. Tampoco se encontró evidencia de productos que implementen soluciones complementarias para aspectos clave, como expansión de consultas considerando el contexto de la búsqueda y la aplicación de métodos de evaluación a los documentos en base a la calidad de los mismos.

SRI para documentos científicos de Ciencias de la Computación (II)

Explotando las capacidades que poseen los meta-buscadores, se consideró factible generar un SRI que utilice bases de datos de otros buscadores que sean específicos para la recuperación de documentos científicos. También se optó por desarrollar componentes complementarios, tanto para el tratamiento de las consultas como para la aplicación de un algoritmo de ranking específico para evaluar el tipo de resultados con el que se desea operar, según distintas métricas, asignando a cada documento una calificación que servirá de referencia para establecer el orden de los resultados de la consulta.

Expansión de consultas

Un SRI posee varias alternativas para optimizar la de búsqueda de información, una de ellas es tomar la consulta del usuario y ampliarla agregando diversos términos, manteniendo coherencia con el dominio de la consulta. Este método es conocido como expansión de consultas(QE); los términos adicionales generan nuevas consultas, denominadas expansiones. De esta forma el SRI puede acceder a una mayor cantidad de documentos relevantes para el usuario, obteniendo listados de resultados individuales por cada expansión, los cuales posteriormente son unificados y ponderados antes de ser devueltos al usuario.

Ontología (I)

Existen diferentes opciones para implementar un proceso de expansión de consultas para un SRI, algunas son: tesauros, diccionarios, sistemas expertos, etc. Para el caso particular del SRI a generar se hace uso de una ontología de dominio específica para una subárea temática de las Ciencias de la Computación.

Una ontología se define como una forma de representar el conocimiento de un ámbito específico, que utiliza los términos y relaciones que conforman su vocabulario base, agregando elementos que permiten extender el vocabulario, como relaciones entre conceptos, permitiendo organizarlos jerárquicamente.

Ontología (II)

Adaptando la definición anterior al área de Ciencias de la Computación, se puede considerar a una ontología como un esquema conceptual correspondiente a un dominio acotado, que permite la comunicación y la transmisión de información entre sistemas. Esto constituye una herramienta de gran utilidad para la recuperación 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.

Características evaluables

Para desarrollar el método particular para evaluar los documentos se optó por considerar las siguientes características:

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

Métricas de evaluación

Para cada una de las características elegidas para valorar un artículo se distinguen diversos indicadores bibliométricos preexistentes que han sido validados por la comunidad científica. A continuación se enuncian y explican aquellos que serán utilizados para evaluar cada una de las características establecidas.

Calidad de la fuente de publicación (I)

- **Publicación en revista científica**
 - **Factor de Impacto(IF) [Web of Knowledge]**

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.
 - **SCImago Journal Rank(SJR) [Grupo SCImago(Scopus)]**

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

Calidad de la fuente de publicación (II)

- Publicación en Congreso o Evento Científico
 - Ranking CORE [Computer Research & Education(Australia)]
Un congreso o conferencia es clasificado, según su importancia, en un determinado nivel preestablecido: A*, A, B y C, respectivamente.

Calidad de los autores (I)

- Índice H [Jorge E. Hirsch]
Se calcula en base a la distribución de las citas que han recibido las publicaciones de un determinado autor. Para hallarlo, solo basta ordenar de forma descendente las publicaciones de un autor por el número de veces que ha sido citada cada publicación, un autor tiene índice h si el h de sus N_p trabajos recibe al menos h citas cada uno, y los otros ($N_p - h$) trabajos tienen como máximo h citas cada uno.

Calidad de los autores (II)

- Índice G [Leo Egghe]

Los artículos de un autor son ordenados de manera descendente de acuerdo con el número de citas recibidas por cada uno de ellos, similar al 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 mayor orden, es considerado como el índice G.

$$g^2 \leq \sum_{i \leq g} c_i$$

Calidad del artículo

- Índice AR [Bihui Jin]

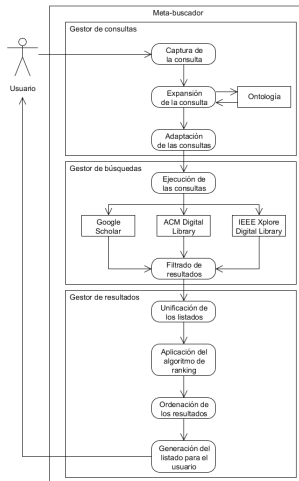
Permite evaluar la calidad de una colección de publicaciones, este indicador combina la cantidad de citas con la antigüedad de cada publicación, para así establecer una valoración de la colección. Es un índice complementario del Índice H.

$$AR = \sqrt{\sum_{p \in H} \frac{cit_p}{ant_p}}$$

- Cantidad de citas

La cantidad de citas recibidas, por sí solas, también es utilizada como métrica para evaluar la calidad de un documento científico en particular.

SRI(meta-buscador)



Construcción de una ontología (I)

El método de expansión de consultas para el meta-buscador es una ontología. Su diseño requirió las siguientes actividades:

- Definición del dominio: dado el contexto de aplicación del SRI se comenzó por seleccionar una subárea temática dentro de las Ciencias de la Computación a partir de la cual realizar la ontología. Se determinó que se comenzaría por la subárea de Inteligencia Artificial (IA).
- Determinar los términos a incluir: se generó una lista 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.

Construcción de una ontología (II)

- 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.

Construcción de una ontología (III)

- Definición de propiedades de las clases: aquellos atributos que describan a los conceptos, considerando las relaciones existentes entre ellos. Se definieron relaciones implícitas a partir de las conexiones entre los conceptos:
 - "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 relaciones explícitas, aquellas definidas para dar mayor valor a la ontología considerando su objetivo de aplicación, siendo representada por la relación de sinonimia, utilizando los términos relevados al inicio del proceso.

Construcción de una ontología (IV)

- Creación de instancias de las clases: utilizando los términos de mayor atomicidad, se definieron los conceptos que conformarían las instancias de las clases de la ontología.

Para la implementación de este diseño de la ontología se utilizo Protégé.

Desarrollo del método de expansión (I)

Este método obtendrá de la ontología aquel concepto que guarde mayor similitud con la consulta ingresada por el usuario (`consulta_original`) y a partir de ese concepto obtener los conceptos relacionados al mismo en forma implícita y explícita, padres, hermanos y sinónimos del término.

Para esto, el algoritmo se divide en dos etapas; primero se busca el concepto de la ontología más similar a la `consulta_original` (`concepto_candidato`) y después se efectúa la expansión al disponer de los conceptos relacionados.

Desarrollo del método de expansión (II)

La búsqueda del concepto_candidato consta de los siguientes pasos:

- 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.

Desarrollo del método de expansión (III)

- En base al contenido de la colección resultante del paso anterior:
 - Si no hay elementos: finaliza la expansión sin resultados validos.
 - Si hay un único elemento: se le considera el concepto_candidato.

Desarrollo del método de expansión (IV)

- Si tiene más de un elemento: se analiza cada concepto cuantificando las coincidencias sintácticas respecto a la consulta_original. Si hay empate, se evalúa al elemento según las relaciones que tenga en la ontología:
 - Si todos tienen el mismo "padre": el concepto es el concepto_candidato.
 - Si no tienen el mismo "padre": se evalúa la cantidad de instancias contenidas por cada "padre", aquel con la mayor cantidad será el concepto_candidato. En caso de otro empate: cada uno de los "padres" será considerado concepto_candidato, conformando una colección nueva.

Desarrollo del método de expansión (V)

Con el o los candidatos determinados, se inicia la etapa de la expansión, que consta de los siguientes pasos:

- Se obtiene el concepto "padre" (concepto_padre) del concepto_candidato.
- Se obtienen los conceptos del mismo nivel que el concepto_candidato, dando lugar a la colección [conceptos_hermanos].
- Se obtienen, en caso de existir, los sinónimos del concepto_candidato, dando lugar a la colección [sinónimos_concepto].

Desarrollo del método de expansión (VI)

- 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:
 - 1 Expansión_1 = consulta_original AND concepto_candidato
 - 2 Expansión_2 = concepto_candidato AND concepto_padre
 - 3 Expansión_3 = concepto_candidato OR [conceptos_hermanos]
 - 4 Expansión_4 = concepto_candidato OR [sinónimos_concepto]

Al finalizar el algoritmo se cuenta con las cuatro expansiones de la consulta_original ingresada por el usuario.

Normalización de las métricas

- Calidad de la fuente de publicación

- Revista Científica: Índice SJR

$$fuentePublicacion = \log_{10}(SJR)$$

- Congreso Científico: Ranking CORE

$$fuentePublicacion = [A* : 1 | A : 0.75 | B : 0.5 | C : 0.25]$$

- Calidad de los autores

- Índice H

$$autores = \log_{10}\left(\sum_i \frac{indiceH_{autor_i}}{i}\right)$$

- Calidad del artículo

- Índice AR

$$calidadPublicacion = \log_{10}\left(\frac{citasRecibidas}{antigüedadPublicacion}\right)$$

Ranking del documento

$$ranking = \alpha(fuentePublicacion) + \beta(autores) + \gamma(calidadPublicacion)$$

α, β, γ : son factores de ajuste entre 0 y 1

Desarrollo

Para el desarrollo del meta-buscador(SRI) se utilizarón las siguientes tecnologías: **Java, JSP, JavaScript, XHTML, y MySQL**. Para implementarse se utilizo como plataforma el servidor web **Tomcat**.

Proceso de implementación:

- Diseño y desarrollo de los métodos para acceder, consultar y extraer los resultados.
- Implementación del algoritmo de ranking.
- Desarrollo de los componentes visuales(interfaz) del meta-buscador.
- Integración de los componentes en un único sistema de software.

Validación

Para la validación del meta-buscador se contó con un grupo de expertos en el desarrollo de métodos de recuperación de información, evaluarán al SRI considerando la relevancia de los documentos retornados a partir de diferentes consultas. Se evalúo el proceso completo de búsqueda: expansión de la consulta, ejecución de las expansiones generadas y ranqueo de los resultados obtenidos.

Tabla de validación

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 retrieval	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%

Conclusiones

El SRI(meta-bsucador) diseñado y desarrollado en el presente trabajo, siendo una implementación particular, permitio recuperar documentos científicos de calidad comprobable del área de Ciencias de la Computación.

Se alcanzaron buenos resultados al trabajar con diferentes componentes, como el acceso a multiples fuentes, expansiones de la conslta original del usuario, ranqueo de cada resultado obtenido utilizando un algoritmo diseñado específicamente para este fin.

Bibliografía

-  Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. Kuna, H., Rey, M., Martini, E., Solonezen, L., Podkowa. Revista Latinoamericana de Ingeniería de Software 2014.
-  THE AR-INDEX: COMPLEMENTING THE H-INDEX. Bihui Jin. National Library of Science CAS, Beijing, 100080, China. 2005-2007.
-  <http://www.core.edu.au/>