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

Emanuel García Pérez

28 de agosto de 2014

Introducción (I)

Hoy en día podemos tener acceso a casi cualquier información desde cualquier parte del mundo a través de Internet. Administrar toda esta información es una tarea muy compleja, debido a la cantidad de documentos disponibles en la web y la complejidad propia de cada uno de ellos.
La información se busca y recupera de diversas formas, una de ellas es usando motores de busqueda(buscadores). Estos buscadores han mejorado considerablemente, sin embargo los resultados que proveen al usuario no son del todo eficientes en cuanto a calidad y cantidad, ádemas de presentar fallos en correlación con el objetivo de la búsqueda.

Introducción (II)

Debido a las deficiencias que presentan estas formas de recuperación de información se busca desarrollar un sistema de recuperación de información(SRI) de dominio especifico, en particular un meta-buscador que se oriente a la recuperación de documentos científicos de Ciencias de la Computación y establecer un ranking de estos, considerando para ello la calidad de cada uno de los documentos recuperados.

Sistema de Recuperación de Información (I)

Un Sistema de Recuperación de Información(SRI) es un proceso que posee capacidad para gestionar información, es decir, recuperar, almacenar y mantener dicha información para distintos fines, según el contexto de su aplicación.
Existen diversas propuestas sobre la organización interna de un SRI, se eligió utilizar una que se basa en los siguientes elementos:

Sistema de Recuperación de Información (II)

- Documentos: Constituyen la fuente de información sobre la cual se pretende realizar búsquedas.
- Consultas: Son generadas por los usuarios del SRI que tienen por objetivo recuperar la información a la cual el sistema provee acceso.
- Representación de los Documentos: Serán las consultas y las relaciones que se definan entre ellos 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.

Sistema de Recuperación de Información (III)

Los principales tipos de SRI que actualmente operan en Internet son: directorios(Yahoo), buscadores(Google), y meta-buscadores(Ixquick). Esto nos permite asegurar que existen implementaciones de SRI en Internet que utilizan diferentes métodos de búsqueda sobre contextos generales o particulares, incluyendo implementaciones a medida para el incremento de la relevancia de los resultados a presentar al usuario.
De estos se destacan notablemente los meta-buscadores, debido a que su modularidad permite que los componentes del SRI sean desarrollados a medida para cubrir las necesidades establecidas para una implementación particular.

Sistema de Recuperación de Información (IV)

Para desarrollar el SRI particular requerido por el presente trabajo, se considerarón 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 de los documentos obtenidos de cada fuente para ordenar el listado integral a presentar al usuario.

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

Hay diversas iniciativas para la generación de SRI de proposito especifico en áreas particulares, pero no se encontró evidencia de que existan implementaciones de SRI que sean aplicadas a bases de datos de documentos científicos del área de Ciencias de la Computación.

Tampoco se encontró evidencia de productos que implementen soluciones complementarias para aspectos clave, como la expansión de las consultas 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 entregar al usuario.

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

Explotando las capacidades que poseen los meta-buscadores, se considero factible generar un SRI que utilice bases de datos de otros buscadores que sean especificos para la recuperación de documentos científicos de Ciencias de la Computación. También se opto por desarrollar componentes complementarios, tanto para el tratamiento de las consultas como para la aplicación de un algoritmo de ranking especifico para evaluar el tipo de resultados con el que se desea operar, según distintas métricas, ampliamente aceptadas por la comunidad científica, asignando a cada resultado una calificación que servirá de referencia para establecer el orden de los resultados para el usuario.

Expansión de consultas

Un SRI posee varias alternativas para lograr optimizar el proceso de búsqueda de información, una de ellas consiste en tomar la consulta del usuario y ampliarla a partir de agregar diversos términos, obtenidos comúnmente a través de fuentes externas, 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 presentados al usuario.

Ontología (I)

Existen diferentas 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 especifica para una subárea temática de las Ciencias de la Computación.
Una ontología se define como una forma de representar el conocimiento de un ámbito especifico, que utiliza los términos y relaciones que conforman su vocabulario base, agregando elementos que permiten extender el vocabulario, como relaciones entre conceptos, permitiendo organizarlos jerarquícamente.

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, tanto interna como externamente. Esto constituye una herramienta de gran utilidad para la recuperación y análisis del conocimiento a través de una estrcutura 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 opto por considerar las siguientes características de los mismos:

1. 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 características similares.
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 si, considerando la antiguedad del mismo y la cantidad de veces que haya sido citado en otros documentos.

Métricas de evaluación (I)

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 utlizados para evaluar cada una de las características establecidas.

Métricas de evaluación (II)

- Calidad de la fuente de publicación
  1. Publicación en revista científica
     - Factor de Impacto (IF)
     - SCImago Journal Rank (SJR)
  2. Publicación en Congreso o Evento Científico
     - Ranking CORE
- Calidad de los autores
  1. Índice H
  2. Índice G
- Calidad del artículo
  1. Índice AR
  2. Cantidad de citas

Métricas de evaluación (III)

Estructura del SRI.

Funcionamiento del SRI.

Construcción de una ontología

Construcción.

Desarrollo del método de expansión

Desarrollo del método.

Diseño del algoritmo

Disenio algoritmo.

Desarrollo del algoritmo

Desarrollo algoritmo.

Desarrollo del prototipo.

Validación del prototipo.

Conlusiones finales.