El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia de documental en dos pasos

Fernando Martínez Santiago
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, s/n. Edif.. A3
dofer@ujaen.es

Resumen: Tesis doctoral en Informática realizada por Fernando Martínez Santiago bajo la dirección de el doctor L. Alfonso Ureña López (Univ. de Jaén). El acto de defensa de tesis tuvo lugar en septiembre de 2004 ante el tribunal formado por los doctores Felisa Verdejo Maillo (UNED), Manuel Palomar Sanz (Univ. de Alicante), Horacio Rodríguez Hontoria (Univ. politécnica de Cataluña), José Carlos González Cristóbal (Univ. politécnica de Madrid), y Ana García Serrano (Univ. politécnica de Madrid). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Problema de la fusión de documentos, Recuperación de información, Recuperación de Información multilingüe, Recuperación de Información distribuida.

Abstract: PhD Thesis in Computer Science written by Fernando Martínez Santiago under the supervision of Dr. L. Alfonso Ureña López (Univ. of Jaén). The author was examined in September 2004 by the commitee formed by Felisa Verdejo Maillo (UNED), Manuel Palomar Sanz (Univ. of Alicante), Horacio Rodríguez Hontoria (Univ. politécnica of Cataluña), José Carlos Gónzales Cristóbal (Univ. politécnica of Madrid) y Ana García Serrano (Univ. politécnica of Madrid). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Collection fusion problem, Information Retrieval, Cross Language Information Retrieval, Distributed Information Retrieval.

1 Introducción

En este texto se propone un nuevo enfoque, cálculo de la relevancia documental en dos pasos, para afrontar el conocido problema de la fusión de colecciones, o simplemente mezcla de resultados. En breve, la fusión de colecciones está relacionada con la Recuperación de Información la cual, frente a una necesidad de información del usuario, debe responder con una lista de documentos relevantes para la consulta dada. En ocasiones, la obtención de tal lista de documentos debe obtenerse a partir de la fusión o mezcla de varias listas obtenidas con independencia las unas de las otras, y es en ese aspecto en el cual se centra el presente trabajo, ilustrando la bondad del método propuesto en dos escenarios: Recuperación de Información multilingüe y Recuperación de de información de distribuida.

Usualmente, la Recuperación de información tradicional se define como la selección del subconjunto de documentos adecuados a la necesidad de información de un usuario. Entonces, un sistema Recuperación de información multilingüe es aquel sistema de Recuperación de Información que está capacitado para recuperar aquellos documentos relevantes para una determinada necesidad de información con independencia del idioma usado en la consulta y en la colección de documentos consultada. De igual manera, un sistema de Recuperación de información distribuida es aquel que frente a una necesidad de información dada, está capacitado para operar sobre varias bases documentales independientes de manera que el usuario perciba que está consultando una única colección.

Los dos escenarios descritos comparten un problema común: frente a una consulta de usuario, es necesario confeccionar una única lista de documentos relevantes, combinación de varias listas generadas independientemente las unas de las otras a partir de motores de búsqueda independientes. A este problema se le viene llamando el problema de la fusión de colecciones.

Una hipótesis que se defiende en esta tesis es que dada una determinada necesidad de información, tanto la puntuación como la posición alcanzada por dos documentos pertenecientes a dos colecciones distintas no es comparable debido principalmente a que la relevancia asignada a un documento no es un valor absoluto, sino muy al contrario, fuertemente dependiente de la colección a la cual pertenece tal documento. Por otra parte, es posible percibir la unión de todos los documentos devueltos por cada motor de búsqueda como una nueva colección de tamaño reducido y pequeño vocabulario, ya que sólo los términos que aparecen en la consulta formulada por el usuario son de interés en esta nueva colección. En virtud de estas dos simplificaciones, tal colección puede ser reindexada y contrastada con la consulta del usuario, obteniendo así una nueva única lista de documentos puntuados en relación con esta nueva colección creada, indexada y consultada en tiempo de ejecución. Esta es, en esencia, la estrategia que se propone para solucionar el problema de la fusión de colecciones y que hemos denominado cálculo de la relevancia documental en dos pasos.

2 Estructura de la tesis

El primer capítulo define la Recuperación de información, la Recuperación de información multilingüe y la Recuperación de información distribuida. También se introduce el problema de la fusión de colecciones, así como las hipótesis que motivan el trabajo, y qué objetivos se pretenden alcanzar.

El capítulo dos ofrece una introducción a la Recuperación de información: se definen los modelos de Recuperación de información booleano, probabilístico y vectorial, que son los más difundidos en la literatura, así como los recursos y técnicas más apreciados en esta disciplina.

El capítulo tres se dedica a la Recuperación de información multilingüe. Aquí se propone una clasificación de estos sistemas atendiendo a qué traducen (consulta, documento o ambos) y a cómo traducen (máquinas de traducción, tesauros multilingües, etc.).

El capítulo cuatro es una introducción a la Recuperación de información distribuida. De qué elementos constan estos sistemas, qué información requieren y cómo obtener tal información son algunos puntos tratados. El capítulo finaliza con un repaso de los sistemas distribuidos más extendidos en la literatura.

En el capítulo cinco se describe con detalle el algoritmo propuesto para solucionar el problema de la fusión de colecciones: el cálculo de la relevancia documental de dos pasos. Además se proponen diversas extensiones del algoritmo original con la finalidad de poder aplicarlo a la mayor cantidad de escenarios posibles. Finalmente, se propone una novedosa manera de aplicar la conocida técnica de pseudo-realimentación de consultas por relevancia en sistemas distribuidos, a la que denominamos pseudo-realimentación de consultas por relevancia global.

Los capítulos seis y siete son un compendio de los experimentos realizados y resultados obtenidos en el problema de la fusión de colecciones aplicado tanto a Recuperación de información multilingüe como distribuida. Así se realizan diversos experimentos variando los siguientes parámetros:

- En recuperación de información multilingüe:
  - El número de idiomas presentes en la colección documental (cuatro, cinco u ocho).
  - El recurso de traducción: diccionarios electrónicos y máquinas de traducción automáticas.
  - Expansión de la consulta original.
  - Estrategia para la fusión de colecciones.
- En recuperación de información distribuida:
  - El número de colecciones, hasta 80.
  - El algoritmo de selección de colecciones más prometedoras.
  - La expansión de consultas local y globalmente.
  - Estrategia para la fusión de colecciones.

El texto finaliza con las conclusiones que resumen las aportaciones realizadas, así como las líneas de investigación que quedan abiertas.