Procesamiento del Lenguaje Natural, num. 30 (2003), pp.127-128 recibido 1-03-2003; aceptado 20-05-2003

IR-n un sistema de Recuperación de Información basado en pasajes

Fernando Llopis Pascual
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. Correos, 99 03080 Alicante
llopis@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada por Fernando Llopis Pascual bajo la dirección de los doctores José Luis Vicedo González y Antonio Ferrández Rodríguez de la Universidad de Alicante. El acto de defensa de la tesis tuvo lugar el 6 de mayo de 2003 ante el tribunal formado por los doctores Manuel Palomar Sanz (Univ. de Alicante), Alfonso Ureña López (Univ. de Jaén), Emilio Sanchís Arnal (Univ. Politécnica de Valencia), Miguel Angel Alonso Pardo (Univ. de A Coruña)y Horacio Rodríguez Hontoria (Univ. Politécnica de Barcelona). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: Recuperación de Información, Recuperación por pasajes, Búsqueda de Respuestas

Abstract: PhD Thesis in Computer Science written by Fernando Llopis Pascual under the supervision of Dr. José Luis Vicedo González and Dr. Antonio Ferrández Rodríguez (University of Alicante Spain). The author was examined in May 6^{th}, 2003 by the commitee formed by Dr. Manuel Palomar Sanz (Univ. de Alicante), Alfonso Ureña López (Univ. de Jaén), Emilio Sanchís Arnal (Univ. Politécnica de Valencia), Miguel Angel Alonso Pardo (Univ. de A Coruña) y Horacio Rodríguez Hontoria (Univ. Politécnica de Barcelona). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Information Retrieval, Passage Retrieval, Question Answering.

1 Introducción

Es curioso comprobar el increíble cambio que se ha producido en cuanto a la disponibilidad de información en formato digital en los últimos años. Muchos son los factores que han provocado este cambio, no obstante el que mayor influencia ha tenido, sin duda alguna, ha sido Internet.

Este incremento de la cantidad de información disponible en formato digital ha supuesto un incremento notable del interés en la investigación en sistemas de información textual. El objetivo es claro, se desea automatizar el proceso de búsqueda de todo tipo de información en una serie de documentos.

Este proceso de búsqueda ha evolucionado desde intentar localizar documentos que contienen información relevante (sistemas de recuperación de información, RI) a los ambiciosos proyectos de localizar y generar información estructurada (sistemas de extracción de información, EI) o incluso a localizar respuestas concretas en grandes colecciones de documentos (sistemas de búsqueda de respuestas, BR).

Es evidente que los posibles logros de estos dos últimos tipos de sistemas son mucho más ambiciosos que los alcanzados por los sistemas de RI. No obstante, no hay que olvidar un aspecto muy importante. Tanto los sistemas de EI como los sistemas de BR requieren o utilizan sistemas de RI para filtrar los documentos que no contienen información relevante.

Así, el campo de investigación en RI sigue abierto, no sólo en cuanto a mejora de las técnicas para determinar con mayor precisión que documentos son relevantes, sino también con el objetivo de detectar aquellas partes de un documento qué realmente son relevantes para una pregunta o tema determinado. Esta es la línea que siguen los sistemas de RI basados en pasajes (RP).

ISSN 1135-5948

2 Los sistemas de Recuperación de Información basados en pasajes

Los sistemas de RP, se diferencian de lo que sería la RI, basada en el análisis del documento completo, en la forma de valorar la relevancia de un documento. Estos últimos calculan la relevancia de un documento con respecto a una pregunta mediante la aplicación de una serie de medidas que valoran sobre todo la frecuencia de aparición de los términos de la pregunta en el documento completo. En contraposición, los modelos de RP estudian la aparición de los términos de la pregunta en fragmentos contiguos de texto dentro de cada documento, a los que se denomina pasajes.

Trabajos previos demuestran que la utilización de estos pasajes como unidad básica de información, mejora sensiblemente los resultados. Esta mejora se debe, principalmente, a que ya no sólo se valora el hecho de que los términos de la pregunta aparezcan en el documento, sino que además, se valora la proximidad con que aparecen. Adicionalmente, los sistemas de RP ofrecen la ventaja de indicar no sólo qué documento es relevante, sino que además, permiten localizar, qué parte del documento es realmente relevante.

3 El sistema de Recuperación de Información IR-n

Los modelos de RP se basan en la división de un documento en una serie de pasajes, que permiten definir la relevancia de un documento en función de la relevancia de cada uno de esos pasajes. Sin embargo, no se ha llegado a un consenso acerca de cómo definir pasajes de forma que el sistema alcance un comportamiento óptimo, ni cómo obtener la relevancia de un documento en función de la relevancia de los pasajes que lo forman. En función de cómo se aborda la división del documento en pasajes se diferencian tres enfoques de sistemas de RP: modelos basados en el discurso, modelos semánticos y modelos de ventana. Los modelos del discurso utilizan las propiedades de estructura del documento, tales como frases, marcas de párrafo o marcas HTML, para definir los pasajes. Los modelos semánticos se basan en la aparición de tópicos en el documento para definir los pasajes. Los modelos de ventana dividen los documentos en pasajes de tamaño fijo. Para realizar esta división, estos modelos pueden basarse o no en la estructura del documento.

En general, todos los modelos citados emplean fundamentalmente párrafos y/o palabras como unidad de información básica a partir de la que se definen los pasajes. Los modelos que utilizan el párrafo para definir los pasajes pueden tener problemas en el momento de definir los pasajes si no se dispone de información acerca de la composición de los párrafos en el documento original. Además los párrafos pueden utilizarse en ocasiones, más por motivos visuales que por la propia estructuración del documento. Por otra parte, los modelos basados en el uso de la palabra como unidad para definir los pasajes, son muy dependientes del estilo de escritura utilizado en los documentos. Además, si se utilizan únicamente las palabras como elemento a considerar en la definición del pasaje, puede ocurrir que los pasajes considerados relevantes carezcan de estructura y se dificulte en gran medida la comprensión del texto recuperado. El trabajo que se presenta en esta tesis es una nueva propuesta de sistema de RP, que se encuadraría dentro de los modelos de ventana, pero diferenciándose de las actuales propuestas, fundamentalmente, en la unidad que se utiliza tanto para definir los pasajes como en la medida empleada para calcular la similitud de los mismos. Nuestra línea de investigación ha sido la de utilizar la frase como unidad para la definición y cálculo de similitud de los pasajes. El uso de la frase como unidad en un sistema de RP permite disponer siempre de pasajes con estructura y sentido, así como independizar el estilo utilizado por los diversos autores de los textos donde se realiza la búsqueda de documentos relevantes. Este modelo también contempla la definición de medidas de similitud que permiten optimizar el rendimiento del sistema en función de la pregunta y de las colecciones de texto utilizadas, e incluso, de la tarea en concreto en la que se emplee.

La evaluación del sistema IR-n se ha realizado de forma externa, mediante la participación en las conferencias CLEF (tarea de RI monolingüe español) y en las conferencias TREC junto con el sistema SEMQA (tarea de BR). En ambos casos los resultados han estado muy por encima de la media de los sistemas presentados, y se ha comprobado su gran eficacia a nivel de precisión a los pocos documentos recuperados.