ReCIBE, Año 4 No. 1, Febrero 2015

Hacia la utilización de recuperación de información en bibliotecas digitales

Cecilia San Martín Méndez
Universidad Veracruzana,
Facultad de Estadística e Informática
zs13015662@estudiantes.uv.com

Sodel Vázquez Reyes
Universidad Autónoma de Zacatecas,
Ingeniería de Software
vazquezs@uaz.edu.mx

Resumen: A pesar de los avances en los sistemas de Recuperación de Información, el problema sigue siendo el acceso a la información, porque la cantidad de documentos disponibles para ser consultados aumenta diariamente. Por lo tanto, el acceso a la información se ha vuelto complejo, cuando se quiere regresar al usuario resultados relevantes sobre la consulta que él ha realizado. Permitir un acceso más fácil y rápido a la información contenida en los documentos de una biblioteca digital, a través de una herramienta que proporcione al usuario información relevante en base a la consulta que se plantee, es el objetivo general del trabajo. La recuperación de pasajes será implementada en un prototipo para un sistema de búsqueda en biblioteca digital. Contar con un sistema que permita realizar consultas de información en lenguaje natural y revisar el contenido de los documentos sin necesidad de conocer nombres de autor o título, será de gran ayuda a los usuarios.

Palabras clave: Recuperación de pasajes, bibliotecas digitales y procesamiento de lenguaje natural.

Towards using of information retrieval in digital libraries

Abstract: Despite advances in IR systems, the problem remains access to information, because the amount of documents available for consultation increases daily. Therefore, access to information has become complex when the user wants to return relevant results on the inquiry that he has done. Allow easier and faster access to the information contained in the documents in a digital library, through a tool that provides the user with relevant information based on the query that arises, is the general objective of this work. The recovery passages will be implemented in a prototype for a search in digital library. Having a system that allows to process information queries in natural language and to review the content of documents without knowing names of the author or title, will be helpful to users.

Keywords: Passages retrieval, digital libraries and natural language processing.

1. Introducción y Motivación

La recuperación de pasajes relevantes a una consulta dada por el usuario tiene como objetivo, seleccionar aquellos pasajes de texto de algún documento ubicado en la colección de documentos en donde pueda estar la respuesta, a la consulta emitida por el usuario.

Un pasaje es un trozo de texto que pertenece a un documento, el cual puede ser dividido de diferentes maneras, esto dependerá principalmente del criterio que utilice el desarrollador. Algunas posibles divisiones son (Llopis-Pascual, 2001):

- **Pasajes de longitud fija:** consiste en dividir el documento en pedazos de texto con un número fijo de palabras.

- **División de los documentos a partir de las propiedades estructurales:** como lo pueden ser secciones, párrafos o frases.
- Dividir y agrupar el texto de tal forma que queden pasajes cuyo texto contenga una semántica similar y relacionada con la consulta.

Usar pasajes en vez de documentos completos permite encontrar trozos de texto relevantes en un documento. Los pasajes facilitan la búsqueda de información, debido a que es más fácil hacer comparaciones sobre un pedazo de texto a comparar contra todo un documento.

A continuación se describen los algoritmos existentes para la recuperación de pasajes. Posteriormente se representa la propuesta del prototipo encargado de llevar a cabo la recuperación de pasajes en la biblioteca digital. Y finalmente se especifica el trabajo futuro.

2. Recuperación de pasajes

Los algoritmos y técnicas más usadas para la recuperación de pasajes son MITRE, MULTITEXT, sistema de la Universidad de Alicante, S-stemmer, sistemas de recuperación de lógica difusa, técnicas de ponderación de términos y sistema SiteQ.

MITRE está basado en solapamiento de palabras y realiza un conteo del número de términos que un pasaje tiene en común con la consulta. Los pasajes también dependen de su tamaño, y el peso de cada pasaje es multiplicado por el logaritmo del número de caracteres en el pasaje que no sean espacios en blanco (Soriano, 2008). MULTITEXT está basado en la densidad de palabras y favorece los pasajes cortos que contienen muchos términos con alto valor de frecuencia inversa del documento (idf). Para el pesado de los términos se ha usado también la frecuencia de los términos (tf)/idf estándar (Soriano, 2008). El sistema de la Universidad de Alicante reordena los pasajes con base en una medida de

similitud, la cual considera el traslape de los términos de la consulta y el pasaje. Puede ser llevado a otro idioma muy fácilmente, ya que sólo emplea información léxica (Rubio, 2008). S-stemmer reduce las formas plurales al singular. Para el castellano, este algoritmo puede enriquecerse teniendo en cuenta que los plurales de sustantivos y adjetivos terminados en consonante se consiguen con el sufijo –es (M. Vallez, 2007). Los sistemas de recuperación de lógica difusa permite realizar la búsqueda eliminando signos de puntuación, artículos, conjunciones, plurales, tiempos verbales, palabras comunes, dejando sólo aquellas palabras que el sistema considera relevantes. La recuperación se basa en proposiciones lógicas con valores de verdadero y falso, teniendo en cuenta la localización de la palabra en el documento (Molina, 2011). La técnica de ponderación de términos pretende darle un valor adecuado a la búsqueda dependiendo de los intereses del usuario. El valor depende de los términos pertinentes que contenga el documento y la frecuencia con que se repita (Molina, 2011). El sistema SiteQ busca los 1000 pasajes de los documentos recuperados que obtengan el mejor puntaje de acuerdo a una medida de similitud. Esta medida está basada en el emparejamiento de términos de la consulta y la distancia entre ellos. Cada término de la consulta tiene asignado un peso, el cual está determinado por su componente morfológico (Rubio, 2008).

De acuerdo a las descripciones de los algoritmos y de los requerimientos que se tienen en el acceso a la *información de una biblioteca digital*. Los algoritmos que mejor se ajustan son el algoritmo MITRE y el sistema de la Universidad de Alicante. Debido a que son los algoritmos que toman en consideración pocos elementos a evaluar, tales como: a).- realizar conteo de números de términos que un pasaje tiene en común con la consulta, b).- los pasajes dependen de su tamaño, c).- el peso de cada pasaje es multiplicado por el logaritmo del número de caracteres en el pasaje que no sean espacios en blanco, y d).- reordenamiento de los pasajes con base a su medida de similitud. Estos elementos permiten que la recuperación de los pasajes relevantes se pueda llevar a cabo de una manera fácil y práctica. Considerando estos elementos de evaluación los resultados obtenidos pueden ser mucho más acertados a la

búsqueda que el usuario realice, con lo que se puede lograr una mejor eficiencia del prototipo.

3. Propuesta de prototipo

El método que se seguirá para la implementación del sistema de recuperación de pasajes dentro de una biblioteca digital, se encontrará divido en dos fases que a continuación se mencionan (Ver Figura 1). La primera fase estará determinada por la implementación de un subprograma que se ejecutará fuera de línea, el cual estará encargado de generar el índice de cada párrafo contenido en los documentos. La segunda fase inicia cuando el usuario ingresa una consulta en lenguaje natural sobre el sistema de recuperación de pasajes, posteriormente el sistema procesa la consulta para poder seleccionar los párrafos más relevantes a la consulta recibida y pueda regresar al usuario una lista de párrafos relacionados al tema solicitado. La lista de párrafos recuperados ofrecerá al usuario la posibilidad de poder ver o descargar el documento completo. El proceso de comparación entre la consulta y los párrafos contenidos en el índice, para seleccionar los más relevantes a la consulta proporcionada será determinado por el algoritmo que se utilice, actualmente se puede elegir uno de los siguientes cuatro algoritmos: model_distance, model_simple, model_termweight y model_rw. Si el usuario determina que la lista de párrafos proporcionada por el sistema no es relevante a sus necesidades, entonces podrá reformular su consulta para poder recibir resultados más acordes a sus necesidades. La arquitectura que se está planteando contiene procesos que contemplan al usuario, como por ejemplo, el ingreso de la consulta en lenguaje natural, la sugerencia de consultas, el corrector de ortografía y los pasajes que se le regresen al usuario estarán ordenados de acuerdo a su grado de relevancia.

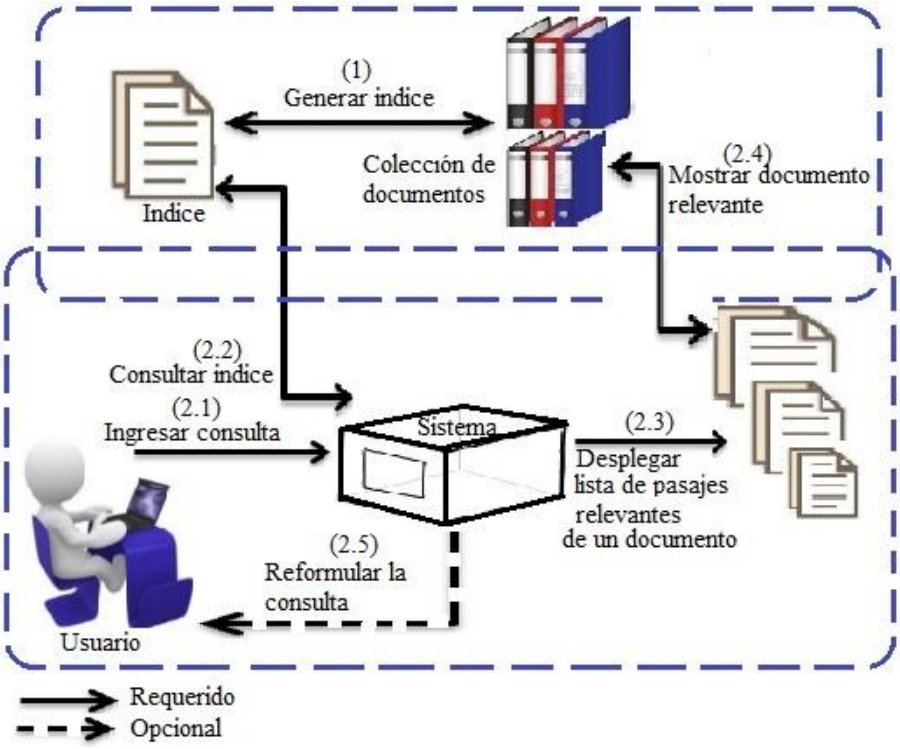

Figura 1. *Método para la implementación del sistema de recuperación de pasajes.*

Para recibir la consulta del usuario (paso 2.1), el prototipo despliega un cuadro de dialogo donde el usuario debe introducir la consulta en lenguaje natural, Figura 2.

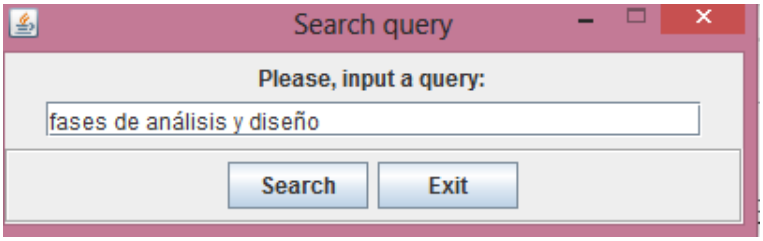

Figura 2. *Ingreso de la consulta a procesar.*

En la Figura 3 podemos ver el despliegue de la lista de pasajes más relevantes a la consulta proporcionada por el usuario (paso 2.3). La lista de pasajes esta ordenada de acuerdo al grado de similaridad entre la consulta y los pasajes de

la colección. Relevancia que el prototipo calcula con el algoritmo seleccionado, donde el valor de 1 indica un 100% de similaridad.

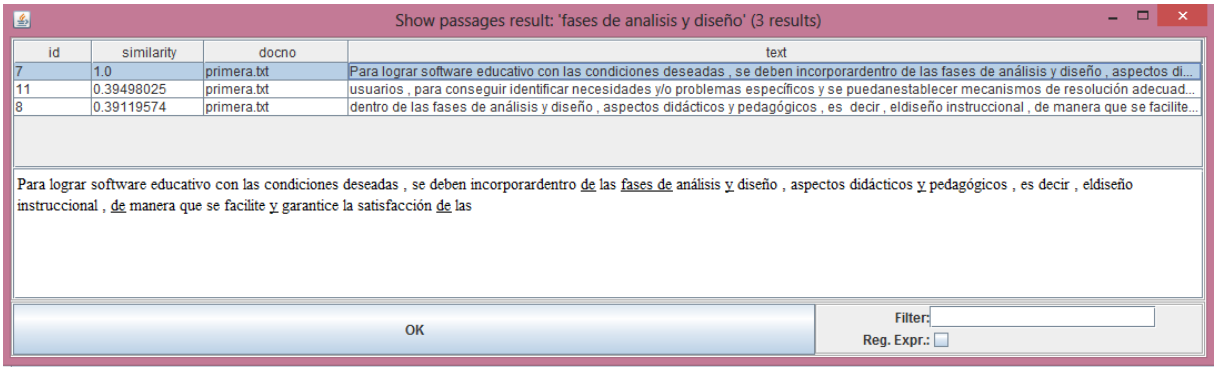

Figura 3. *Lista de pasajes relevantes a la consulta proporcionada por el usuario.*

4. Evaluación del prototipo

Para determinar la viabilidad del prototipo, se han realizado pruebas con los cuatro algoritmos que permiten la recuperación de pasajes, mencionados en la sección anterior. En las pruebas se han utilizado dos tipos de consultas, las consultas textuales (texto como aparece en los documentos) y las consultas parafraseadas. Los documentos procesados están contenidos en una colección Gold-Standard creada con tesis del nivel de licenciatura de la Facultad de Estadística e Informática de la Universidad Veracruzana, son archivos en formato PDF que los alumnos entregan cuando se gradúan a través de la modalidad de tesis.

Con la evaluación analizamos el comportamiento que el prototipo está teniendo, es decir, que tan relevantes son los pasajes que se le están regresando al usuario, en qué posición el usuario podrá encontrar un pasaje que realmente satisfaga su necesidad de información. Y de esta manera conocer si realmente

el acceso a la Información en Bibliotecas Digitales a través de Recuperación de Pasajes es viable.

El sistema no podrá realizar evaluaciones automáticas para indicar que tan relevantes son los pasajes que se le regresan al usuario, puesto que esta es una tarea que el usuario deberá indicar de acuerdo a la necesidad de información que requiera y el nivel de conocimiento que tenga del tema. En la Figura 4 se muestra los resultados obtenidos de las pruebas realizadas, aplicando las métricas de *recuerdo* y *precisión* para los dos tipos de consultas antes mencionadas. Con los resultados obtenidos sobre las pruebas realizadas se puede observar que el modelo que mejor se ajusta a las necesidades planteadas para la recuperación de pasajes en bibliotecas digitales es el *modelo* de distancia. Porque es el algoritmo que mejores resultados obtiene para consultas parafraseadas, las cuales son consideradas las adecuadas para el contexto de bibliotecas digitales, debido a que el usuario no deberá conocer el contenido de las tesis, simplemente con tener la idea de lo que está buscando podrá acceder a la información que requiere.

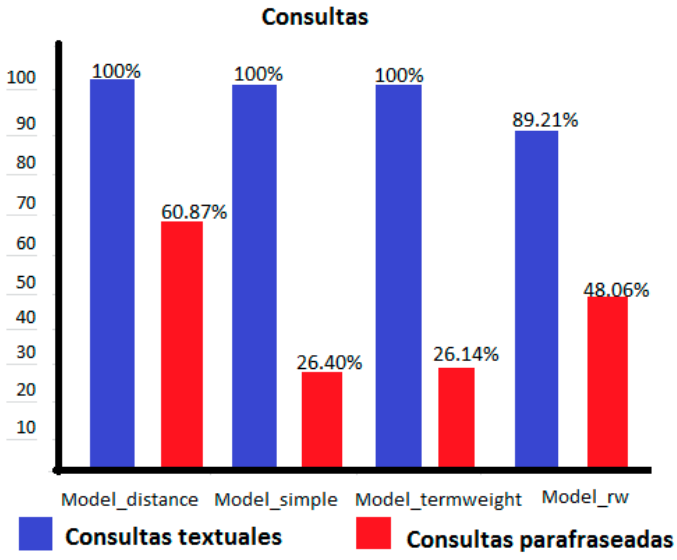

Figura 4. *Comparación de distintos algoritmos para la recuperación de pasajes.*

5. Trabajo Futuro

El procesamiento de la información es una tarea importante, existen ya algunos sistemas que realizan la recuperación de respuestas para una pregunta específica, haciendo uso de la recuperación de pasajes, como por ejemplo el Sistema de Búsqueda de Respuestas Multilingüe S-QUAMUS (García Cumbreras, 2006), el sistema JIRS (Soriano, 2008), y el sistema IR-n (Llopis-Pascual, 2001). Ellos están diseñados para dar respuestas a una pregunta. Sin embargo, en el contexto de bibliotecas digitales, no es la mejor opción porque no se requieren respuestas exactas y cortas, necesitamos contextualizar el tema de búsqueda. Al contar con un sistema que permita realizar consultas de información en lenguaje natural y revisar el contenido de los documentos sin necesidad de conocer nombres de autor o título, será de gran ayuda a los usuarios.

Agradecimientos

Se agradece el apoyo brindado por CONACYT, beca de maestría con número 366077 para CVU 559714.

Referencias

Llopis, P. F. (2001) IR-N: Un Sistema de Recuperación de Información Basado en Pasajes. Tesis de Doctorado, Universidad de Alicante.

Molina, M. P. (2011). Recuperado el 20 de Marzo de 2014, http://www.mariapinto.es/e-coms/recu_infor.htm.

Rubio, G. H. (2008). Recuperación de Pasajes Orientada a la Resolución de Preguntas con Restricción Temporal. Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica.

Soriano, J. M. (2008). Recuperación de pasajes multilingüe para la búsqueda de respuestas. Tesis de Doctorado, Universidad Politécnica de Valencia.

Notas biográficas:

Cecilia San Martín Méndez Licenciada en Informática, egresado de la Universidad Autónoma de Veracruz (UV), actualmente estudia la Maestría en Sistemas Interactivos Centrados en el Usuario (MSICU) Facultad de Estadística e Informática. Sus intereses son el desarrollo de software centrado en el usuario, ingeniería de software y la seguridad informática.

Sodel Vázquez Reyes Profesor-investigador de la Unidad Académica de Ingeniería Eléctrica (UAIE) en la Universidad Autónoma de Zacatecas. Obtuvo el grado de Doctor en Ciencias Computacionales por la “University of Manchester”, Reino Unido, en el 2008 y el grado de Maestría en Ciencias Computacionales por el Instituto Nacional de Astrofísica Óptica y Electrónica en el 2000. Se incorporó a la Universidad Autónoma de Zacatecas (UAZ) en el 2010 y actualmente es profesor investigador con perfil PROMEP, líder del Cuerpo Académico “UAZ-CA-192 Tecnologías de la Información” y Responsable del Programa de Ingeniería de Software.