RECUPERACIÓN DE INFORMACIÓN EN BASES DE DATOS DE TEXTO

Claudia Deco, Cristina Bender, Federico Severino Guimpel
Departamento de Sistemas e Informática
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario
(2000) Rosario, Argentina
Tel (+ 54 341) 4802649 int. 141
{deco, bender, rumpel}@fceia.unr.edu.ar

Nora Reyes
Departamento de Informática,
Universidad Nacional de San Luis
(5700), San Luis, Argentina
Tel (+54 2652) 420822 int. 257
nreyes@unsl.edu.ar

Resumen

La Recuperación de Información de la Web es uno más de los problemas de buscar en un conjunto los elementos más cercanos a una consulta dada bajo un cierto criterio de similitud. Es de interés aprovechar las cualidades de los espacios métricos con el objeto de resolver una consulta de manera efectiva y eficiente.

El objetivo de este proyecto es mejorar la recuperación y extracción de información no estructurada, utilizando recursos lingüísticos para la preparación de una estrategia de búsqueda. Para esto, se consideran aportes desde la lingüística para el refinamiento semántico de los conceptos; y desde la matemática y las ciencias de la computación para la búsqueda por similitud. Además, se pretende lograr un marco unificador para describir y analizar soluciones para el problema de la búsqueda en bases de datos no estructuradas.

Introducción

Con la evolución de las tecnologías de la información y las comunicaciones, han surgido almacenamientos no estructurados de información, tales como texto libre, imágenes, audio y video. Esto requiere modelos más generales que las bases de datos tradicionales. Es decir, nuevos modelos tales como bases de datos métricas, y por lo tanto, se requiere contar con métodos y técnicas que permitan realizar búsquedas eficientes sobre estos tipos de datos.

La Web, por ejemplo, almacena este tipo de información no estructurada. Al convertirse la Web en el mayor repositorio de conocimiento y en un medio de publicación fácilmente accesible para todos, la Recuperación de Información ha dejado de ser un campo exclusivo de los especialistas en Ciencias de la Información y ha pasado a ser un campo relacionado con cualquier persona. Si bien los usuarios no tienen por qué conocer técnicas de recuperación de información, la propuesta de esta investigación es la de mejorar los resultados de su búsqueda por medio de un "especialista" que implementa estas técnicas.

Para esto, por un lado se propone un refinamiento semántico que prepara una estrategia de búsqueda adecuada como lo haría el especialista en ciencias de la información. Este refinamiento utiliza tanto modelos independientes del lenguaje como conocimiento lingüístico específico, para la preparación de una estrategia de búsqueda que represente la necesidad de información del usuario, y así lograr una mejora en la recuperación de información.

Por otro lado, dada una consulta, existirán millones de elementos en la base de datos, y no podemos compararlos uno a uno. Se necesitan métodos de acceso eficientes que permitan recuperar rápidamente los elementos que satisfacen los criterios de la consulta. Entonces, este problema se puede abstraer y presentarlo de la siguiente forma: dados un conjunto de objetos de naturaleza desconocida y una función de distancia definida entre ellos, y dado otro objeto llamado la consulta, encontrar todos los elementos del conjunto similares a la consulta. Por lo tanto, este problema se puede convertir en un problema de búsqueda en espacios métricos. Es decir, se puede realizar una búsqueda aproximada en texto, documentos similares a una consulta, imágenes similares a una imagen de muestra, etc.

Conceptos Básicos

Una base de datos de texto, es un sistema que provee acceso eficiente a amplias masas de datos textuales. Estas bases de datos están organizadas en documentos, cuyo contenido no está estructurado. Un requerimiento importante de estos sistemas es que desarrollen búsquedas rápidas ante la consulta de un usuario.

La Recuperación de Información (*Information Retrieval*) es la representación, almacenamiento, organización y acceso a ítems de información [Baeza et al., 1999]. Tanto la representación y organización de los ítems de información como la caracterización de la necesidad de información del usuario, no son problemas simples de resolver. El objetivo principal de la Recuperación de Información es satisfacer la necesidad de información planteada por un usuario en una consulta en lenguaje natural especificada a través de un conjunto de palabras claves. En general, este proceso hacia la recuperación de documentos textuales relevantes a la consulta presentada, no es un proceso simple debido a la complejidad semántica del vocabulario. Su meta principal es recuperar información que podría ser útil o importante al usuario, y no sólo datos que satisfagan una consulta dada.

Un sistema de recuperación de datos tradicional, tal como una base de datos relacional, trata con datos que tienen una estructura y una semántica bien definidas. Estos sistemas permiten recuperar todos los objetos que satisfacen las condiciones especificadas en una expresión regular o en una expresión del álgebra relacional. Entonces, un sistema de recuperación de datos sólo recupera los datos que coinciden exactamente con el patrón a recuperar. En cambio, un sistema de recuperación de información encuentra datos importantes que hagan la mejor coincidencia parcial con el patrón dado. Esto se debe a que la recuperación de información generalmente trata con texto en lenguaje natural, el cual no está siempre bien estructurado y podría ser semánticamente ambiguo. Por ejemplo, si se realiza una consulta por el término “cáncer”, además de obtener como resultado los documentos que contengan este término, se debería obtener también los documentos en que aparezca neoplasma, ”carcinoma”, “canceroso”, etc.

Una consulta en un sistema de recuperación de información es una solicitud de documentos pertenecientes a algún tema. Dada una colección de documentos y una consulta del usuario, el objetivo de una estrategia de búsqueda es obtener todos y sólo los documentos relevantes a la consulta. El problema central se reduce a establecer una correspondencia entre el lenguaje de la consulta y el lenguaje del documento.

La Recuperación de Información es una tarea compleja porque se enfrenta con varios problemas. Por un lado, los autores y los usuarios frecuentemente utilizan diferentes palabras o expresiones cuando se refieren a un mismo concepto. Si en un documento, en lugar del término "cáncer" apareciera la palabra "neoplasma", este documento no se recuperaría. Este problema se puede resolver haciendo uso de sinónimos. Por otro lado, algunos términos pueden tener significados diferentes. Por ejemplo, la palabra "cáncer" puede referirse a una enfermedad en medicina, a un signo zodiacal en astrología o a una constelación de estrellas en astronomía. Esto se soluciona desambiguando el término o agregando otros términos específicos relacionados con la acepción de interés. En ambos problemas se pueden utilizar recursos lingüísticos, tales como diccionarios, diccionarios multilingües, tesauros y ontologías. Un recurso lingüístico puede incluir sinónimos, variantes de escritura, ampliación de siglas, variaciones de deletreo, términos equivalentes en otros idiomas, hiperónimos, hipónimos, y/o merónimos, entre otros.

Se han desarrollado técnicas para mejorar la Recuperación de Información. Una de ellas es el *stemming*, que consiste en obtener la raíz de las palabras, de forma que el proceso de búsqueda se realice sobre las raíces y no sobre las palabras originales.

Por otro lado, en el entorno de búsqueda tradicional, el usuario debe dividir su interés de búsqueda en distintos conceptos. No siempre un término representa en forma adecuada un concepto. Encontrar otros términos equivalentes o más adecuados para expresar un concepto es realizar una expansión de consulta [Efthimiadis, 1996]. Esta situación requiere un cambio en el pensamiento del proceso para elegir los términos de búsqueda. Podría ser necesario consultar recursos lingüísticos, tales como un tesauro o un diccionario, para incorporar nuevos términos. La expansión de consultas es el proceso de suplementar la consulta original con términos adicionales, y es un método para mejorar el desempeño de la recuperación. En [Deco et al., 2005] se propone la expansión semántica de la consulta utilizando recursos lingüísticos para mejorar la Recuperación de Información. En dicho trabajo, se experimenta con el recurso WordNet [Miller, 1995], y se muestra que tanto la cobertura como la precisión de los resultados mejora con este recurso. La mejora de la precisión de una búsqueda se logra al presentarle al usuario una estructura jerárquica de conceptos que le permita hacer un recorrido conceptual de su consulta. Es decir, moverse por jerarquías conceptuales, subiendo o bajando de nivel conceptual, y para seleccionar un término más preciso a su necesidad de información.

Objetivos

La web es un repositorio de tamaño continuamente creciente. Como se pretende discriminar las páginas por su contenido, se plantea representarlas con vectores de sus palabras representativas, lo que presupone el manejo de vectores con miles de componentes. Esto no permite asegurar que la dimensión intrínseca del espacio métrico será alta o no, pero sí que soluciones para espacios vectoriales no son aplicables. Considerando la incertidumbre sobre la dimensión intrínseca del espacio, los algoritmos de tipo Voronoi aparecen como más aptos. Además, el ambiente de la web es muy dinámico, continuamente se están agregando, modificando y borrando páginas.

El objetivo de este proyecto es mejorar la recuperación y extracción de información no estructurada, utilizando recursos lingüísticos para la preparación de una estrategia de búsqueda. Para esto, se consideran aportes desde la lingüística para el refinamiento semántico de los conceptos; y desde la matemática y las ciencias de la computación para la búsqueda por similitud. Además, se pretende lograr un marco unificador para describir y analizar soluciones para el problema de la búsqueda en bases de datos no estructuradas.

Los objetivos específicos de este proyecto son:

- Desarrollar nuevas metodologías para ampliar las capacidades de recuperación de información que utilicen recursos lingüísticos, tales como diccionarios, tesauros y ontologías.
- Proponer nuevos algoritmos que permitan buscar eficientemente en bases de datos no convencionales, como ser algoritmos para búsqueda en espacios métricos.
- Utilizar las propiedades de los índices sobre espacios métricos para mejorar la calidad de los resultados de una búsqueda de información.
- Diseñar nuevas estructuras de datos para espacios métricos que, aprovechando las características del tipo de recuperación que se necesita resolver, permitan responder eficientemente las consultas. Además se requiere que sean dinámicas; es decir capaces de actualizarse sin necesidad de reconstruir completamente la estructura.
- Una clase de algoritmos para búsqueda en espacios métricos son los basados en pivotes. Por lo tanto, nos proponemos en particular trabajar sobre ellos y proponer algún nuevo algoritmo basado en pivotes y criterios para la selección de pivotes.

Velocidad en la recuperación y calidad de los resultados son dos propiedades necesarias en cualquier sistema de Recuperación de Información. Los espacios métricos cuentan con índices que permiten la recuperación de objetos cercanos a uno dado de una forma rápida y adecuada, por lo que resultan estructuras prometedoras sobre las cuales se pueden construir motores de búsqueda.

El desempeño de un motor de búsqueda se puede evaluar a través de los indicadores *Precisión* y *Recall*. La Precisión es el número de documentos relevantes a la consulta dada dividido el total de documentos recuperados. El Recall es el cociente entre la cantidad de documentos recuperados y el total de documentos relevantes de la colección. Un motor alcanza un buen rendimiento cuando maximiza ambos valores; es decir, recupera la mayoría de los documentos relevantes disponibles en la colección con la menor cantidad de documentos irrelevantes.

En trabajos ya iniciados se analizaron algoritmos de indexado sobre espacios métricos lo que permitió vislumbrar la aptitud del M-tree [Ciaccia *et al.*, 1997] para el entorno Web. Se optó entonces por extender el M-tree ya que al realizar una búsqueda, elige los recorridos analizando la consulta ingresada con información que contiene su estructura. Además, logra velocidad porque en este proceso va descartando los subárboles que no tienen datos próximos a la consulta. Los resultados obtenidos son correctos porque en el proceso de poda los subárboles rechazados contienen siempre datos irrelevantes, por definición del M-tree. Para alcanzar ambos objetivos, velocidad y calidad de resultados, se elige una adecuada representación de las páginas Web y un conveniente criterio de similitud entre las mismas que posibiliten el empleo de un M-tree. La propuesta del índice XM-tree [Deco et al., 2007] logra un alto rendimiento en cuanto a calidad de resultados en un proceso de Recuperación de Información en la Web, alcanzando buenos valores de Precisión y Recall. Además, la eficiencia de las búsquedas ofrece importantes mejoras sobre los índices para espacios vectoriales e índices invertidos.

Una extensión futura del XM-tree es el borrado real de datos, ya que actualmente las páginas fuera de línea siguen presentes en el índice y figuran como accesibles desde un caché, tal cual lo hace el M-tree; éste es un aspecto importante ya que entonces el XM-tree se volvería un índice realmente apto para el entorno Web, porque sería completamente dinámico.

BIBLIOGRAFIA

[Baeza et al., 1999] Baeza-Yates, R., Ribeiro-Neto, B. (eds.), Modern Information Retrieval. 1999, New York. ACM Press.

[Chávez *et al*., 2001] Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J. Searching in metric spaces. ACM Computing Surveys, 33(3):273-321, September 2001.

[Ciaccia *et al.*, 1997] P. Ciaccia, M. Patella y P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proc. of the 23rd Conference on Very Large Databases (VLDB'97)*, 426–435, 1997. URL http://www-db.deis.unibo.it/research/papers/VLDB97.pdf

[Deco *et al*., 2005] Deco, C. Bender, J. Saer, M. Chiari, Motz, R. Semantic Refinement for Web Information Retrieval. C. Proceedings of the 3rd Latin American Web Congress. La Web 2005. IEEE Press. Pp 106-110. Argentina.

[Deco *et al*., 2007] C. Deco, G. Pierángeli, C. Bender, N. Reyes. XM-tree, un nuevo índice para Recuperación de Información en la Web. En Proceedings del IV Workshop de Ingeniería de Software y Bases de Datos en el marco del XIII Congreso Argentino de Ciencias de la Computación, CACIC 2007. pp. 656-667, ISBN 978-950-656-109-3. Corrientes, Argentina, 2007.

[Efthimiadis, 1996] Efthimiadis E.N. Query Expansion. In Annual Review of Information Systems and Technology (ARIST), v31, pp 121-187, 1996.

[Miller, 1990] Miller, G. WordNet: An on-line lexical database. International Journal of Lexicography 3(4). 1990.