RECUPERACIÓN DE INFORMACIÓN EN BASES DE DATOS NO ESTRUCTURADAS

Claudia Deco[1], Nora Reyes[2], Cristina Bender[1]

[1] Facultad de Ciencias Exactas, Ingeniería y Agrimensura,
Universidad Nacional de Rosario, Rosario, Argentina.
{ deco, bender }@fceia.unr.edu.ar

[2] Departamento de Informática, Universidad Nacional de San Luis, San Luis, Argentina
nreyes@unsl.edu.ar

Resumen

En la búsqueda de información en la Web hay tres problemas a enfrentar. Uno es lograr una estrategia de consulta adecuada que represente la necesidad de información del usuario. Un segundo problema es recuperar rápidamente los elementos que satisfacen los criterios de esta consulta. El tercero es personalizar los resultados obtenidos al perfil de cada usuario. Respecto al segundo problema, si la búsqueda de información se realiza en bases de datos no estructuradas, como es el caso de la web, existirán millones de elementos en la base de datos y compararlos uno a uno con la consulta no es eficiente. Por lo tanto, se necesitan métodos de acceso eficientes que permitan recuperar rápidamente los elementos que satisfacen los criterios de la consulta. Esto se puede tratar como un problema de búsqueda en bases de datos métricas realizando una búsqueda por similitud aproximada. Esta línea de investigación aborda este problema.

Palabras Claves: Bases de datos, Recuperación de Información, Espacios Métricos.

Contexto

Esta línea de I+D se está llevando a cabo a través de proyectos de la Universidad Nacional de Rosario (UNR), y de la Universidad Nacional de San Luis (UNSL).

El PID de la UNR involucrado es:

– [ING348] *Búsqueda personalizada de información en Bases de Datos*, dirigido por Claudia Deco. (2011-2014).

El cual es continuación del PID [ING201] *Recuperación de información en bases de datos de texto* (2007-2010).

El PID de la UNSL involucrado es:

– [22/F034] *Nuevas Tecnologías para el Tratamiento Integral de Datos Multimedia*, dirigido por Marcela Printista. (2010-2013), en particular la línea *Recuperación de Datos e Información Multimedia*, dirigida por Nora Reyes.

Además, los integrantes de esta línea, trabajan en conjunto con otros grupos de investigación:

- Grupo Concepción del Sistemas de Información, de la Universidad de la República (UdelaR), cuya directora es la Dra. Regina Motz.
- Grupo INFOSUR, de la Facultad de Humanidades y Artes (UNR), dirigido por la Dra. Zulema Solana.
- Grupo Sistemas de Información Inteligentes, de la Universidad Nacional de Rosario.
- Proyecto 22/F014 *Tecnologías Avanzadas de Bases de Datos* de la UNSL.
- Investigadores de la Universidad Michoacana de San Nicolás de Hidalgo, de la Universidad de Chile y de la Universidade da Coruña, colaboraciones que se iniciaron en el marco del Proyecto RIBIDI VII.19 de CYTED.
- Otros grupos de investigación de universidades latinoamericanas en el marco

de la Comunidad Latinoamericana de Objetos de Aprendizaje (LACLO).

Introducción

La Recuperación de Información (o Information Retrieval) es la representación, almacenamiento, organización y acceso a ítems de información [1]. 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. Un sistema de recuperación de información encuentra datos importantes que hagan la mejor coincidencia parcial con el patrón dado. 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. 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. 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 esto un problema es establecer una correspondencia entre el lenguaje de la consulta y el lenguaje del documento. Por otro lado, para escribir su consulta 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. Para esta expansión, que puede ser desarrollada manual, automática o interactivamente, se pueden utilizar recursos lingüísticos (diccionarios, 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, merónimos, entre otros. Entonces, la expansión de consultas es el proceso de suplementar la consulta original con términos adicionales, y es un método para mejorar el desempeño en la recuperación de información. Algunos resultados sobre este tema se obtuvieron el proyecto anterior ING201 (2007-2010) ([2], [3], [4], [5]).

Con la evolución de las tecnologías de la información y las comunicaciones, otro problema que surge es la posibilidad de acceso a grandes volúmenes de información no estructurada, 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 las bases de datos métricas [6]. Y, por lo tanto, se requiere contar con metodologías que permitan realizar búsquedas eficientes sobre estos tipos de datos. Un requerimiento importante es que se desarrollen búsquedas rápidas ante la consulta de un usuario. Este problema se comenzó a abordar dentro del proyecto ING201 del cual éste es continuación, y algunos resultados se pueden ver en [7], [8], [9], y [10].

En estos escenarios es de interés realizar búsquedas por similitud. Esta similitud es modelizada usando una función distancia, provista por un experto del dominio. Esta función distancia es costosa de calcular, por lo que han surgido técnicas basadas en índices para intentar reducir el número de evaluaciones. Estas técnicas se apoyan en índices basados en pivotes e índices basados en particiones compactas, que tienen como objetivo dividir el espacio en clases de equivalencia y utilizar el índice para filtrar algunas clases en tiempo de consulta [6]. La mayoría de las técnicas fueron desarrolladas asumiendo que la topología de la colección de objetos es razonablemente regular, pero en experimentaciones hechas sobre espacios donde las colecciones de objetos puede agruparse en subespacios o clusters han demostrado que no son tan eficientes. A partir de [11], donde se propone una estructura de dos niveles, Sparse Spatial Selection for Nested Metric Spaces (SSSNMS), para trabajar con este tipo de espacios, en la línea de I+D presentada aquí, analizamos el uso de SSS y Listas de Clusters. Algunos resultados preliminares pueden verse en [12].

Otro problema que se presenta en la búsqueda, es que dos usuarios obtienen los mismos resultados frente a la misma consulta, aunque existan diferencias en sus características y preferencias. Sin embargo, si se consideran estas características individuales se pueden lograr resultados más afines a cada persona. Entonces, además del interés temático, también las características personales y las preferencias de cada usuario, deben ser consideradas en el proceso de recuperación de información para lograr resultados más afines a cada persona. Esto es lo que se llama personalización. Las características y preferencias de los usuarios son almacenadas como metadatos. Para los documentos resultantes de la búsqueda, los metadatos son un conjunto de

atributos necesarios para describir sus principales características. Ambos conjuntos de metadatos permitirán que un sistema recomendador sugiera a un usuario una lista ordenada de los recursos que sean más afines a su perfil. En este sentido, algunos resultados obtenidos se pueden ver en [13] y [14].

Las integrantes de esta línea de investigación trabajan desde hace varios años en proyectos de I+D relacionados con estos temas.

Líneas de investigación y desarrollo

Para alcanzar los objetivos que proporcionen una ayuda a los problemas mencionados en la introducción, se han planteado distintas líneas de investigación las cuales se interrelacionan.

Respecto a la personalización de resultados, se plantea utilizar sistemas multiagentes diseñando a sus componentes con arquitecturas de agentes que los capaciten para actuar de forma flexible y eficiente, en entornos multiagentes. Se ha utilizado este modelo de agente para diseñar e implementar agentes recomendadores en el dominio de la educación ([13], [14]). Estos trabajos son realizados en conjunto con los grupos de Sistemas de Información Inteligentes (UNR, Concepción de Sistemas de Información (UdelaR), y Universidades pertenecientes a LACLO.

Paralelamente, se trabaja en la expansión semántica de la búsqueda en el contexto del proyecto [ING348], en conjunto con el grupo INFOSUR (UNR). El objetivo es producir la estrategia de búsqueda temática utilizando recursos lingüísticos. Cuando el usuario hace una consulta, ingresa un conjunto de términos que describen el tema de su interés. Luego es necesario un proceso que desambigüe estos términos y los expanda semánticamente incorporando sinónimos y conceptos relacionados.

Respecto al requerimiento de desarrollar búsquedas rápidas ante la consulta de un usuario, se trabaja sobre las búsquedas por similitud. En particular, se está trabajando sobre el problema de colecciones donde los objetos pueden agruparse en clusters: Espacios Métricos Anidados (Nested Metric Spaces). Este trabajo surge de trabajos previos desarrollados en colaboración con investigadores del Laboratorio de Bases de Datos de la Facultade de Informatica de la Universidade da Coruña.

Resultados Obtenidos / Esperados

Entre los resultados obtenidos en esta línea de investigación se encuentran:

- Diseño de la arquitectura del sistema recomendador como un sistema multiagente.
- 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, en particular en espacios anidados.

Entre los resultados esperados en esta línea de investigación se encuentran:

- 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.

Formación de Recursos Humanos

El equipo de trabajo está integrado por la Dra. Claudia Deco (investigadora de la Universidad Nacional de Rosario y de la Universidad Católica Argentina), la M. Sc. Cristina Bender (investigadora de la Universidad Nacional de Rosario y de la Universidad Católica Argentina), y la M. Cs. Nora Reyes (investigadora de la Universidad Nacional de San Luis).

Dentro del marco de esta línea de investigación, se desarrolla actualmente la tesis de Doctorado en Ciencias de la Computación “Estructuras de Datos para Memorias Jerárquicas” de Nora Reyes, bajo

la dirección del Dr. Gonzalo Navarro de la Universidad de Chile, en la UNSL.

Además, se ha desarrollado la siguiente tesina de grado de la Licenciatura en Ciencias de la Computación, de la Universidad Nacional de Rosario: Combinando Métodos para Búsquedas en Espacios Métricos Anidados. Alumno Hugo Gercek, Directora: Nora Reyes y Claudia Deco.

Asimismo, se mantiene abierta la propuesta de tesis de grado y de posgrado, como así la realización de pasantías en el marco de InterU.

Referencias

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

[2] Deco C., Bender C., Saer J., Chiari M., Motz R. Semantic refinement for Web Information Retrieval. Proc. 3rd Latin American Web Congress. La Web 05. IEEE Press. pp 106-110. ISBN 0-7695-2471-0. 2005

[3] Deco C., Bender C., Chiari M. Problemas de la traducción de la consulta en la búsqueda de información multilingüe. Revista INFOSUR. Año 1 Nro 1. UNR. ISSN 1851 1996. pp 39-50. 2007.

[4] Deco C., Bender C., Severino Guimpel F., Reyes N. Recuperación de información en bases de datos de texto. Proc. Workshop Investigadores en Ciencias de la Computación WICC 2008. Gral Pico, La Pampa, Argentina. ISBN 978-950-863-101-5, 2008.

[5] Ponce A., Deco C., Bender, C. Proposal of an ontology based Web search engine Proc. Workshop de Bases de Datos, XIV CACIC. Chilecito, Argentina. ISBN 978-987-24611-0-2, 2008.

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

[7] Deco C., Pierángeli G., Bender C. Reyes N. XM-Tree: A new index for web information retrieval. Journal of Computer Science and Technology (JCS&T). Vol. 8, Nro 2, pp 78-84. July 2008.

[8] Deco C., Salvetti M., Reyes N., Bender C. Dynamic selection of suitable pivots for similarity search in metric spaces. Anales Workshop de Bases de Datos y Minería de Datos, XV CACIC. Universidad Nacional de Jujuy. Jujuy, Argentina. pp 991-1000. ISBN: 978-897-24068-4-1, 2009.

[9] Mariano Salvetti, Claudia Deco, Nora Reyes, Cristina Bender. Adaptive and dynamic pivot selection for similarity search. Journal of Information and Data Management, Vol. 2, No. 1, February 2011, Pages 27–34. An official publication of the Brazilian Computer Society Special Interest Group on Databases.

[10] Claudia Deco, Mariano Salvetti, Nora Reyes, Cristina Bender. Book Chapter: Dynamic selection of suitable pivots for similarity search in metric spaces. En: Computer Science & Technology Series. Simari, G., Pesado, P., Paganini J. (Eds) Editorial de la Universidad de La Plata. pp 213-226 ISBN 978-950-34-0684-7. (260 p). La Plata 2010.

[11] Pedreira O., Brisaboa N.R. Spatial Selection of Sparse Pivots for similarity search in metric Spaces. In: 33nd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'07), LNCS vol: 4362, pp. 434-445. Springer (2007)

[12] Hugo Gercek, Nora Reyes, Claudia Deco, Cristina Bender, Mariano Salvetti. Combining methods for searches in nested metric spaces. En Anales del Workshop de Bases de Datos y Minería de Datos en el marco del XVII Congreso Argentino de Ciencias de la Computación. Universidad Nacional de La Plata. La Plata. Argentina, octubre de 2011.

[13] Casali A., Gerling V., Deco C. y Bender C. Sistema inteligente para la recomendación de objetos de aprendizaje **LACLO 2010 Best Papers**. Revista Generación Digital Vol 9, No 1. pp. 88-95. Colombia. 2011.

[14] Ana Casali, Claudia Deco, Cristina Bender and Valeria Gerling. Book Chapter: Recommender system for personalized retrieval of Learning Objects. Book of Educational Recommender Systems and Technologies: Practices and Challenges. ERSAT Edited by Olga C. Santos and Jesus G. Boticario. aDeNu Research Group. UNED, Spain. ERSAT.