Una estructura de indexación para la recuperación de documentos con referencias geográficas*

Miguel R. Luaces, Jose R. Paramá, Oscar Pedreira, Diego Seco

Laboratorio de Bases de Datos, Universidad de A Coruña
Campus de Elviña, 15071 A Coruña, España
{*luaces, parama, opedreira, dseco*}*@udc.es*

Jose R. R. Viqueira
Laboratorio de Sistemas, Universidad de Santiago de Compostela
Constantino Candeira s/n, 15782 Santiago de Compostela, España
joserios@usc.es

Resumen

Tanto los *Sistemas de Información Geográfica* como la *Recuperación de Información* han sido campos de investigación muy importantes en las últimas décadas. Recientemente, un nuevo campo de investigación llamado *Recuperación de Información Geográfica* ha surgido fruto de la confluencia de estos dos campos. El objetivo principal de este campo es definir estructuras de indexación y técnicas para almacenar y recuperar documentos de manera eficiente empleando tanto las referencias textuales como las referencias geográficas contenidas en el texto.

En este artículo presentamos la arquitectura de un sistema para recuperación de información geográfica y definimos el flujo de trabajo para la extracción de las referencias geográficas de los documentos. Presentamos además una nueva estructura de indexación que combina un índice invertido, un índice espacial y una ontología. Esta estructura mejora las capacidades de consulta de otras propuestas.

*Este trabajo ha sido parcialmente soportado por el "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, por la "Xunta de Galicia" ref. PGIDIT05SIN10502PR y ref. 2006/4, por el "Ministerio de Educación y Cienciaref. AP-2006-03214 (Programa FPU) para Oscar Pedreira, y por la "Dirección Xeral de Ordenación e Calidade do Sistema Universitario de Galicia, da Consellería de Educación e Ordenación Universitaria-Xunta de Galicia" para Diego Seco.

1. Introducción

Aunque el campo de investigación de Recuperación de Información [2] ha estado activo las últimas décadas, la creciente importancia de Internet y de la World Wide Web ha hecho de él uno de los campos de investigación más importantes hoy en día. Se han propuesto muchas estructuras de indexación, técnicas de compresión y algoritmos de recuperación diferentes en los últimos años. Estas propuestas se han empleado generalmente en la implementación de bases de datos documentales, bibliotecas digitales y motores de búsqueda en el web.

Otro campo que ha recibido mucha atención en los últimos años es el de los Sistemas de Información Geográfica [17]. Las mejoras recientes en el hardware han hecho posible que la implementación de este tipo de sistemas sea abordable por muchas organizaciones. Además, se ha llevado a cabo un esfuerzo colaborativo por dos organismos internacionales (ISO [9] y el Open Geospatial

Consortium [15]) para definir estándares y especificaciones para la interoperabilidad de los sistemas. Este esfuerzo ha hecho posible que muchas organizaciones públicas estén trabajando en la construcción de infraestructuras de datos espaciales [1] que les permitirán compartir su información geográfica.

Muchos de los documentos almacenados en bibliotecas digitales y bases de datos documentales incluyen referencias geográficas en sus textos. Por ejemplo, las noticias de prensa hacen referencia al lugar donde tuvo lugar el evento y, a menudo, al lugar donde ha sido escrito el documento. Las referencias geográficas se pueden extraer también de páginas web usando la información del texto que contienen, la localización del servidor web y muchos otros elementos de información. Sin embargo, las referencias geográficas de los documentos son usadas pocas veces en los sistemas de recuperación de información. Pocas estructuras de indexación o algoritmos de recuperación tienen en cuenta la naturaleza espacial de las referencias geográficas embebidas en los documentos. Las técnicas puramente textuales se centran sólo en aspectos del lenguaje de los documentos y las técnicas puramente espaciales se centran sólo en los aspectos geográficos de los documentos. Ninguna de estas técnicas es adecuada para una aproximación combinada a la recuperación de información porque ignoran completamente el otro tipo de información. Como resultado, hay una falta de arquitecturas de sistemas, estructuras de indexación y lenguajes de consulta que combinen ambos tipos de información.

Algunas propuestas que han aparecido recientemente [3, 13] definen nuevas estructuras de indexación que tienen en cuenta tanto los aspectos textuales como los geográficos de un documento. Sin embargo, las aproximaciones descritas en estos trabajos no tienen en cuenta algunas particularidades específicas del espacio geográfico. En particular, conceptos como la naturaleza jerárquica del espacio geográfico y las relaciones topológicas entre los objetos deben ser consideradas para representar completamente las relaciones entre los documentos y para permitir que se puedan realizar nuevos e interesantes tipos de consulta a estos sistemas.

En este artículo presentamos una arquitectura de un sistema de recuperación de información y una estructura de indexación que tienen en cuenta estas cuestiones. Primero, se describen algunos conceptos básicos y trabajo relacionado en la Sección 2. A continuación, en la Sección 3, presentamos la arquitectura general del sistema y describimos sus componentes. La arquitectura del sistema define un flujo de trabajo para la construcción de una base datos documental en la que tanto las palabras como las referencias geográficas en los documentos son tenidas en cuenta. La estructura del índice se describe en más detalle en la Sección 4. La estructura del índice está localizada en el corazón de la arquitectura del sistema y permite el almacenamiento y acceso eficiente a los documentos empleando tanto referencias textuales como geográficas. A continuación, en la Sección 5, describimos algunos tipos de consulta que pueden ser contestadas con este sistema y esbozamos los algoritmos que se pueden emplear para resolver estas consultas. Finalmente, la Sección 6 presenta algunas conclusiones y futuras líneas de trabajo.

2. Trabajo relacionado

Los índices invertidos son considerados como la técnica de indexación de texto clásica. Un índice invertido asocia a cada palabra en el texto (organizado como un *vocabulario*) la lista de punteros a las posiciones donde la palabra aparece en los documentos. El conjunto de todas las listas se llama *ocurrencias* [2]. El principal inconveniente de esta técnica es que ignora por completo las referencias geográficas. Los nombres de lugar son considerados simplemente como palabras.

A lo largo de los años se han propuesto una gran variedad de estructuras de indexación espacial. En [6] se puede encontrar un buen resumen de esas estructuras. El objetivo principal de las estructuras de indexación espacial es mejorar el tiempo de acceso a las

colecciones de objetos con datos geográficos. Una de las estructuras de indexación espacial más populares y un ejemplo paradigmático es el R-tree [8]. El R-tree es un árbol balanceado derivado del B-tree que divide el espacio en rectángulos (*minimum bounding rectangles*) jerárquicamente anidados y posiblemente solapados. El número de hijos de cada nodo interno varía entre un mínimo y un máximo. El árbol se mantiene balanceado dividiendo los nodos en los que se produce desbordamiento y combinando los nodos que no alcanzan el número mínimo de descendientes. Los rectángulos se asocian con los nodos hoja y cada nodo interno almacena el *minimum bounding rectangle* de todos los rectángulos en su subárbol. La descomposición del espacio proporcionada por un R-tree es adaptativa (dependiente de los rectángulos almacenados) y solapada (los nodos en el árbol pueden representar regiones solapadas). Un inconveniente de estas estructuras es que no tienen en cuenta la jerarquía del espacio. Los nodos internos en la estructura carecen de significado en el mundo real, sólo tienen significado para la estructura de indexación. Por ejemplo, supongamos que queremos construir un índice para una colección de países, provincias y ciudades. Estos objetos están estructurados en una relación topológica de contenido, esto es, una ciudad está contenida en una provincia que a su vez lo está en un país. Si nosotros construimos un R-tree con estos objetos geográficos la jerarquía de contenidos no se mantendrá.

Se han realizado algunos trabajos para tratar de combinar ambos tipos de índices. Los artículos sobre el proyecto SPIRIT (*Spatially-Aware Information Retrieval on the Internet*) [12, 10, 11, 16, 5] son un muy buen punto de partida para comenzar. En [16], los autores concluyen que manteniendo separado el índice espacial del índice textual, en lugar de combinarlos en un único índice, se consigue un menor coste de almacenamiento aunque, por contra, podría implicar mayores tiempos de respuesta. Más recientes son los artículos [13, 3] que resumen este trabajo y proponen mejoras al sistema y a los algoritmos empleados en el mismo. En su trabajo proponen dos algoritmos como base: *Text-First* y *Geo-First*. Ambos algoritmos emplean la misma estrategia, primero se emplea un índice para filtrar los documentos (el índice invertido en el Text-First y el índice espacial en el Geo-First). El conjunto de documentos resultante es ordenado por sus identificadores y posteriormente filtrado usando el otro índice (el índice espacial en Text-First y el índice textual en Geo-First). Sin embargo, ninguna de estas aproximaciones tiene en cuenta las relaciones entre los objetos geográficos que están indexando.

Una estructura que puede describir adecuadamente las características específicas del espacio geográfico es una *ontología*, la cual se define como una especificación explícita y formal de una conceptualización compartida [7]. Una ontología proporciona un vocabulario de clases y relaciones para describir un ámbito determinado. En [4], se propone un método para el mantenimiento efectivo de ontologías con muchos datos espaciales usando un índice espacial para mejorar la eficiencia de las consultas espaciales. Además, en [10, 5] los autores describen cómo se emplean ontologías en tareas de expansión de los términos de las consultas (*query expansion*), en la elaboración de rankings de relevancia y en la anotación de recursos web en el proyecto SPIRIT. Sin embargo, hasta donde nosotros sabemos, nadie ha tratado de combinar ontologías con otros tipos de índices para obtener una estructura híbrida.

3. Arquitectura del sistema

La Figura 1 muestra nuestra propuesta para la arquitectura de un sistema de recuperación de información geográfica. La parte inferior de la figura muestra el flujo de trabajo para el almacenamiento de documentos. El primer paso de este flujo de trabajo es la tarea *Extracción de palabras clave* donde todos los documentos son analizados y se extraen las palabras clave del texto. En esta tarea se pueden emplear técnicas clásicas de Recuperación de Información para reducir el

número de palabras clave, como puede ser la eliminación de *stopwords* y el uso de otras operaciones sobre texto como *stemmers* y reducción a grupos de nombres [2].

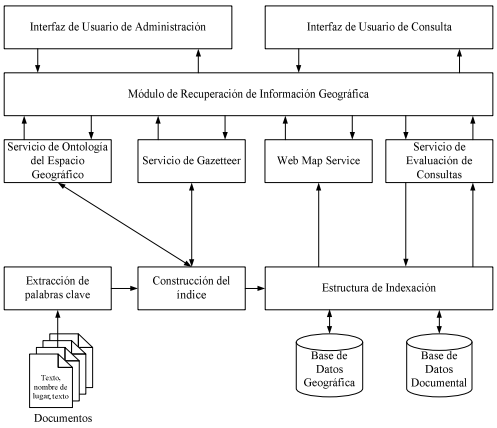

Figura 1: Arquitectura del sistema

Después de esta tarea de extracción de las palabras clave, el sistema está preparado para construir la estructura de indexación. Para esta tarea son necesarios dos servicios. Primero, se emplea un *servicio de gazetteer* en combinación con técnicas de procesamiento del lenguaje natural (NLP) para descubrir los nombres de lugar. Para cada nombre de lugar descubierto se almacenan las referencias geográficas asociadas, obtenidas mediante el *servicio de gazetteer*, junto con la palabra clave. Luego, se emplea una *ontología del espacio geográfico* junto con las palabras clave y las referencias geográficas para construir la estructura de indexación. Este proceso está descrito en más detalle en la Sección 4.

En la parte central de la figura se muestran los servicios de procesado. En la izquierda se pueden ver los ya mencionados *servicio de ontología del espacio geográfico* y *servicio de gazetteer*. En la derecha se pueden ver dos servicios empleados para la resolución de consultas. El situado más a la derecha es el *servicio de resolución de consultas*, que recibe consultas y emplea la estructura de indexación para resolverlas. El otro servicio empleado para crear representaciones cartográficas de los resultados de las consultas es un *Web Map Service* siguiendo la especificación del OGC [14]. Por encima de estos servicios se sitúa un *módulo de recuperación de información geográfica* encargado de coordinar la tarea efectuada por cada servicio en respuesta a las peticiones del usuario.

La capa superior de la arquitectura muestra la interfaz de usuario. El sistema tiene dos interfaces de usuario diferentes: una *interfaz de usuario de administración* que se puede emplear para gestionar la colección de documentos y una *interfaz de usuario de consulta* que puede ser usada para realizar consultas al sistema y navegar sobre los resultados obtenidos.

4. La estructura de indexación

La Figura 2 muestra la estructura de indexación. La base de esta estructura es una ontología espacial. Esta ontología modela tanto el vocabulario como la estructura espacial de las localizaciones geográficas para procesos de recuperación de información. La estructura de una ontología es fija por lo que la estructura de indexación debe ser construida ad-hoc para el dominio en el cual se va a emplear.

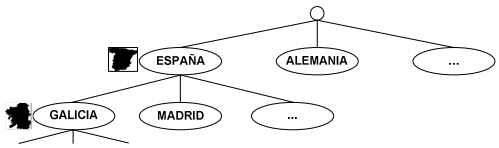

Figura 2: Estructura de indexación

El componente principal de la estructura de indexación es un árbol compuesto por nodos que representan nombres de lugar. Estos nodos están interconectados por medio de relaciones de contenido (por ejemplo, Galicia está contenida en España). En cada nodo almacenamos: (i) la palabra clave (un nombre de lugar), (ii) las referencias geográficas asociadas con el nombre de lugar, (iii) el

minimum bounding rectangle de la geometría que representa ese lugar, (iv) una lista con los identificadores de los documentos que incluyen referencias geográficas a ese lugar y (v) una lista de nodos hijos que están geográficamente contenidos en ese nodo. Si la lista de nodos hijo es muy larga es muy ineficiente acceder a ella de manera secuencial. Por esta razón, si el número de nodos hijo excede un umbral, se empleará un R-tree en lugar de una lista.

En el índice se emplean dos estructuras auxiliares. En primer lugar, una tabla hash almacena para cada nombre de lugar su posición en la estructura de indexación. Esto proporciona un acceso directo a un nodo concreto por medio de una palabra clave que se obtiene mediante el *servicio de gazetteer* si la palabra procesada es un nombre de lugar. La segunda estructura auxiliar es un índice invertido tradicional con todas las palabras de los documentos que se emplea para resolver consultas textuales.

Mantener separados el índice textual del índice espacial tiene muchas ventajas. En primer lugar, todas las consultas textuales pueden ser procesadas de manera eficiente por el índice invertido y todas las consultas espaciales pueden ser procesadas de manera eficiente por el índice espacial. Además, el sistema soporta consultas que combinen aspectos textuales con espaciales. Así mismo se pueden manejar de manera independiente las actualizaciones en cada uno de los índices, esto hace que se puedan añadir o eliminar datos de forma sencilla. Finalmente, se pueden aplicar optimizaciones específicas a cada estructura de indexación de manera individual.

Los principales inconvenientes de esta estructura son: (i) el árbol que soporta la estructura es posiblemente desbalanceado, lo cual penaliza la eficiencia del sistema y (ii) las ontologías tienen un estructura fija y por tanto nuestra estructura es estática y debe ser construida ad-hoc.

5. Tipos de consultas soportadas

La característica más importante de una estructura de indexación es el tipo de las consultas que se pueden resolver con él. Los siguientes tipos de consultas son relevantes en un sistema de recuperación de información geográfica:

- *Consultas puramente textuales*. Estas son consultas del tipo "*recuperar todos los documentos donde aparezcan las palabras* hotel *y* mar".

- *Consultas puramente espaciales*. Un ejemplo de este tipo de consultas es "*recuperar todos los documentos que se refieran a la siguiente área geográfica*". El área geográfica en la consulta puede ser un punto, una ventana de consulta, o incluso un objeto complejo como un polígono.

- *Consultas textuales con nombres de lugar*. En este tipo de consultas, algunas palabras son nombres de lugar. Por ejemplo, "*recuperar todos los documentos con la palabra* hotel *referidos a España*".

- *Consultas textuales sobre un area geográfica*. En este caso se proporciona un área geográfica de interés junto con el conjunto de palabras. Un ejemplo es "*recuperar todos los documentos con la palabra* hotel *que se refieren a la siguiente área geográfica*". Al igual que en las *consultas puramente espaciales* el área geográfica de la consulta puede ser un punto, una ventana de consulta o un objeto complejo.

Los índices invertidos pueden resolver consultas puramente textuales recuperando del índice invertido la lista de los documentos asociados con cada palabra y luego realizando la intersección de las listas. Las consultas puramente espaciales se pueden resolver empleando el índice espacial descendiendo en la estructura teniendo en cuenta sólo aquellos nodos cuyos *bounding box* intersecan con el área geográfica de la consulta. Esta operación devuelve un conjunto de documentos candidatos que tiene que ser refinado con la referencia geográfica actual para decidir si el documento es parte del resultado o no.

Las consultas puramente textuales se pueden resolver en nuestro sistema porque un índice invertido forma parte de la estructura de indexación. De manera similar, las consultas puramente espaciales se pueden resolver porque la estructura de indexación es construida como un índice espacial. Cada nodo en el árbol se asocia con el *bounding box* de los objetos geográficos en cada subárbol. Por tanto, el mismo algoritmo empleado con índices espaciales puede ser empleado con nuestra estructura.

Sin embargo, la estructura de indexación que proponemos puede ser usada para resolver el tercer y el cuarto tipo de consultas los cuales no pueden ser solucionados de manera sencilla empleando un índice invertido y un índice espacial. Para el caso de la consulta con nombres de lugar, nuestro sistema puede descubrir que *España* es una referencia geográfica consultando al servicio de gazetteer y posteriormente emplear la tabla hash de nombres de lugar de la estructura para recuperar el nodo del índice que representa *España*. De este modo se puede ahorrar algún tiempo de acceso suprimiendo parte del recorrido en el árbol.

Con respecto al cuarto tipo de consultas, el índice invertido se emplea para recuperar la lista de documentos que contienen las palabras y la estructura de indexación se emplea para obtener la lista de documentos que hacen referencia al área geográfica. Por tanto, la intersección de ambas listas es el resultado de la consulta. La ventaja de nuestra propuesta en este caso es que las referencias geográficas se pueden proporcionar empleando nombres de lugar.

Otra mejora sobre los índices textuales y espaciales es que nuestra estructura de indexación puede realizar fácilmente expansión de los términos de consulta (*query expansion*) sobre referencias geográficas porque está construida sobre una ontología del espacio geográfico. Consideremos la siguiente consulta "*recuperar todos los documentos que se refieran a España*". El servicio de evaluación de consultas descubrirá que España es una referencia geográfica. El índice de nombres de lugar se empleará para localizar rápidamente el nodo interno que representa el objeto geográfico *España*. Entonces todos los documentos asociados con este nodo forman parte del resultado de la consulta. Sin embargo, todos los hijos de este nodo son objetos geográficos que están contenidos en España (por ejemplo, la ciudad de Madrid). De este modo, todos los documentos referenciados por el subárbol forman también parte del resultado de la consulta. La consecuencia es que la estructura de indexación ha sido empleada para expandir la consulta porque el resultado contiene no sólo aquellos documentos que incluyen el término *España*, sino también aquellos documentos que incluyen el nombre de un objeto geográfico contenido en España (por ejemplo, todas las ciudades y regiones de España).

6. Conclusiones y trabajo futuro

En este artículo se ha presentado una arquitectura de sistema para recuperación de información que tiene en consideración no sólo el texto contenido en los documentos sino también las referencias geográficas incluidas en los documentos y la ontología del espacio geográfico. Esto se logra mediante una nueva estructura de indexación que combina un índice invertido, un índice espacial y una ontología. También se ha presentado cómo las consultas tradicionales pueden ser resueltas usando esta estructura de indexación. Finalmente, se han descrito nuevos tipos de consultas que pueden ser resueltos con la estructura de indexación y se han esbozado los algoritmos que permiten resolver esas consultas.

Actualmente se está finalizando la implementación de un prototipo del sistema y se está trabajando en la evaluación del rendimiento del índice. Son posibles futuras mejoras de la estructura de indexación. En primer lugar, se debe definir un procedimiento para decidir si los hijos de un nodo se deben estructurar como una lista o como un R-tree. Otra línea de trabajo futuro implica explorar el uso de diferentes ontologías y determinar cómo afecta cada

una al índice resultante. Además, está planificado incluir otros tipos de relaciones espaciales en la estructura de indexación complementarias a la de inclusión (por ejemplo, adyacencia). Estas relaciones pueden ser fácilmente representadas en la ontología y la estructura de indexación puede ser extendida para soportarlas. Finalmente, es necesario definir algoritmos para elaborar el ranking de los elementos recuperados por el sistema. Para esta tarea debemos definir una medida de relevancia espacial y combinarla con la relevancia obtenida empleando el índice invertido.

Referencias

[1] Global Spatial Data Infrastructure Association. Retrieved May 2007 from http://www.gsdi.org/.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD Conference*, pages 277–288, 2006.

[4] E. Dellis and G. Paliouras. Management of large spatial ontology bases. In *Proceedings of the Workshop on Ontologies-based techniques for DataBases and Information Systems (ODBIS) of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, September 2006.

[5] Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *Proceedings of In On the Move to Meaningful Internet Systems 2005: ODBASE 2005*, volume 3761 of *Lecture Notes in Computer Science*, pages 1466 – 1482, 2005.

[6] Volker Gaede and Oliver Günther. Multidimensional access methods. *ACM Comput. Surv.*, 30(2):170–231, 1998.

[7] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220, June 1993.

[8] Antonin Guttman. R-trees: A dynamic index structure for spatial searching. In Beatrice Yormark, editor, *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21, 1984*, pages 47–57. ACM Press, 1984.

[9] Geographic information – reference model. International Standard 19101, ISO/IEC, 2002.

[10] Christopher B. Jones, Alia I. Abdelmoty, and Gaihua Fu. Maintaining ontologies for geographical information retrieval on the web. In *Proceedings of On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE Ontologies, Databases and Applications of Semantics, ODBASE'03*, volume 2888 of *Lecture Notes in Computer Science*, 2003.

[11] Christopher B. Jones, Alia I. Abdelmoty, Gaihua Fu, and Subodh Vaid. The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Proceedings of the 3rd Int. Conf. on Geogrphic Information Science*, volume 3234 of *Lecture Notes in Computer Science*, pages 125 – 139, October 2004.

[12] Christopher B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M.J. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies an overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387 – 388, 2002.

[13] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, New York, NY, USA, 2005. ACM Press.

[14] OpenGIS Web Map Service Implementation Specification. OpenGIS Project Do-

cument 01-068r3, Open GIS Consortium, Inc., 2002.

[15] OpenGIS Reference Model. OpenGIS Project Document 03-040, Open GIS Consortium, Inc., 2003.

[16] Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th Int. Symp. on Spatial and Temporal Databases (SSTD)*, volume 3633 of *Lecture Notes in Computer Science*, pages 218 – 235, 2005.

[17] M. F. Worboys. *GIS: A Computing Perspective*. Taylor & Francis, 1995. ISBN: 0-7484-0065-6.