# Recuperación de Información Geográfica basada en una Descripción Semántica del Espacio*

Nieves R. Brisaboa, Miguel R. Luaces, Diego Seco

Laboratorio de Bases de Datos, Universidade da Coruña
Campus de Elviña, 15071, A Coruña, España
{brisaboa,luaces,dseco}@udc.es

**Resumen** La recuperación de información geográfica constituye un novedoso campo de investigación surgido unos pocos años atrás para atender una incipiente demanda de los usuarios consistente en recuperar información relevante no sólo en cuanto a su contenido textual sino también en cuanto a su referente geográfico. La naturaleza espacial de los referentes geográficos motiva el tener en cuenta las características propias de este tipo de información que han sido muy estudiadas en el campo de los sistemas de información geográfica.
En este artículo presentamos las diferentes alternativas existentes en la literatura para realizar el proceso de recuperación de información geográfica. Además, realizamos un análisis en el que mostramos que existen distintos tipos de consultas que no pueden ser resueltas por estas alternativas. Finalmente, describimos una alternativa que incorpora información semántica al proceso y que permite recuperar documentos relevantes no sólo en cuanto a su contenido textual (empleando para ello un índice invertido clásico) sino también en cuanto a su referente geográfico (mediante una descripción semántica del espacio indexado). Esta alternativa abre un nuevo camino de investigación que incrementa de forma cualitativa las consultas que se pueden realizar.

**Key words:** Recuperación de información geográfica, índice espacio-textual, relevancia espacial.

## 1. Introducción

La enorme cantidad de repositorios de información digital disponibles en los últimos años (incluyendo la Web, bases de datos documentales, bibliotecas digitales, etc.) han convertido a la recuperación de información (IR) [1] en una de las áreas de investigación más activas dentro de la informática. Una de las demandas de los usuarios de estos repositorios que no se resuelve de manera adecuada dentro de la IR es la explotación de la información geográfica contenida en los mismos. A pesar de que la información más común en la mayoría de

* Este trabajo ha sido financiado parcialmente por el Ministerio de Educación y Ciencia (PGE y FEDER) ref. TIN2009-14560-C03-02 y por la Xunta de Galicia ref. 08SIN009CT.

los repositorios es de tipo textual, es posible encontrar con bastante frecuencia referencias geográficas en el texto de los documentos que permiten asignar a dichos documentos referentes geográficos (i.e. zonas del espacio en las cuales son relevantes). La naturaleza espacial de estos referentes geográficos les proporciona unas características especiales bien conocidas en el campo de los sistemas de información geográfica (GIS) [2]. El explotar estas características es el objetivo fundamental de un campo de investigación, bastante joven todavía, que se ha denominado recuperación de información geográfica (GIR) [3].

La recuperación de información geográfica consiste por tanto en la explotación de repositorios de información digital mediante consultas que pueden tener naturaleza textual, geográfica o una combinación de ambas (por ejemplo, *recuperar de la web documentos sobre conferencias de computación que se celebren en España*). El enfoque tradicional de los sistemas de recuperación de información sólo resuelve este problema de manera parcial ya que se basan en la aparición de los términos buscados en el documento (siguiendo con el ejemplo, si un documento contiene los términos *conferencia* y *Madrid* pero no *España* será sólo parcialmente relevante para la consulta). Otras técnicas como la expansión de consultas mejoran estos resultados (ya que pueden buscar términos relacionados con los indicados en la consulta) pero siguen sin ofrecer una solución elegante ya que no pueden resolver otras consultas donde la parte espacial se indique mediante una ventana o región de consulta (por ejemplo, *conferencias celebradas en un radio de 100 Km. centrado en Madrid*). Además de esta limitación para resolver determinados tipos de consulta, la información geográfica también requiere de otras metáforas de consulta y presentación de la información que han sido muy estudiadas en el campo de los GIS (por ejemplo, las interfaces de usuario en este tipo de sistemas proporcionan herramientas muy amigables para el dibujo de ventanas o regiones de consulta y el posterior marcado de los lugares relevantes).

En este artículo revisamos el estado del arte de este novedoso campo de investigación centrándonos en la parte de indexación, resolución de consultas y medidas de evaluación. Además, describimos nuestra principal aportación en el área, una estructura de indexación que refleja (y por ende permite explotar) tanto la naturaleza textual como la naturaleza geográfica de los documentos. Esta estructura de indexación está basada en el clásico índice invertido para la parte textual y en una descripción semántica del espacio donde se distribuyen los referentes geográficos de los documentos indexados para la parte geográfica. Esta alternativa permite solucionar de manera elegante la mayoría de los tipos de consulta de interés en sistemas GIR. Por último, describimos nuevas líneas de trabajo que abre esta estructura.

## 2. Trabajo relacionado

El proceso de recuperación de información geográfica, de manera similar al de la recuperación de información, se puede descomponer en tres etapas fundamentales: el análisis de los documentos, la indexación y la resolución de consultas.

Sin embargo dentro de cada una de estas etapas existen diferencias muy significativas. Por ejemplo, en GIR la etapa de análisis de documentos añade, además de todos los procesos en la IR clásica como lematización o borrado de *stopwords*, la localización de nombres de lugar (*geo-parsing*) y su traducción a un modelo geográfico del universo como pueden ser sus coordenadas en latitud/longitud (*geo-referenciation*). En esta revisión del estado del arte nos centraremos más en las etapas de indexación y resolución de consultas.

En la sección anterior introducimos las carencias que tiene un enfoque de IR clásico para los nuevos objetivos de la recuperación de información geográfica. Fundamentalmente estas carencias se deben a que los nombres de lugar (o los referentes geográficos) se tratan como términos de la misma naturaleza que cualquier otro término contenido en el documento. Sin embargo, estos nombres de lugar tienen una naturaleza geográfica implícita que les otorga unas características especiales bien estudiadas en el área de los GIS. De este modo, *A Coruña* y *Pontevedra* más allá de ser términos contenidos en un documento representan localizaciones reales que se pueden representar en un mapa y que además tienen ciertas características como las de ser provincias adyacentes, estar contenidas en una misma comunidad autónoma (y consecuentemente en el mismo país) o ser adyacentes a otra provincia como *Lugo*. Todas estas características son imposibles de capturar empleando únicamente el clásico índice invertido e incluso el empleo de técnicas más complicadas como la expansión de consultas no proporciona una solución elegante al problema.

Por otra parte, se podría pensar en crear sistemas GIR que empleen un índice invertido [4] (o cualquier otra estructura de indexación textual) para resolver la parte textual de las consultas y un R-tree [5] (o alguno de los muchos otros métodos de acceso espacial [6] existentes) para resolver la parte geográfica. Bajo esta idea surgieron las primeras aproximaciones para realizar recuperación de información geográfica.

Los primeros trabajos surgieron del proyecto SPIRIT [7] y se basan en la combinación de una estructura de *grid* [8] con un índice invertido. La estructura *grid* es uno de los índices espaciales más sencillos. Esta estructura divide el espacio en celdas y cada punto se asocia a la celda en la que está contenido. Los autores de este trabajo realizaron pruebas combinando el *grid* y el índice invertido en una única estructura y también manteniéndolos por separado. Su conclusión más importante es que manteniendo ambos índices separados se consigue un menor coste de almacenamiento aunque, por contra, puede implicar mayores tiempos de respuesta. Además, tener separados los índices tiene ventajas en cuanto a la modularidad, facilidad de implementación y facilidad de mantenimiento. Sus resultados también muestran que los métodos propuestos son capaces de competir en términos de velocidad y coste de almacenamiento con estructuras de indexación textual clásicas. Aunque la estructura propuesta es muy sencilla, y ya se han propuesto otras que la superan tanto en velocidad como en coste de almacenamiento, este trabajo es muy relevante ya que ha establecido una de las características distintivas de todas las propuestas posteriores. Dicha característica establece la distinción entre estructuras híbridas, que combinan los índices

textual y espacial en una única estructura, y estructuras de doble índice, que los mantienen por separado.

Trabajos más recientes, como [9,10], describen las dos estrategias base de la indexación en sistemas GIR teniendo en cuenta las propuestas del proyecto SPIRIT. Estas dos propuestas se nombran como *Text-First* y *Geo-First*. A nivel general, ambos algoritmos asumen la existencia de un índice espacial y de un índice textual, y emplean la misma estrategia para resolver las consultas: primero se emplea un índice para filtrar los documentos (el índice textual en el caso de *Text-First* y el índice espacial en el caso de *Geo-First*); el conjunto de documentos resultante se ordena por sus identificadores y posteriormente se filtra usando el otro índice (el índice espacial en el caso de *Text-First* y el índice textual en el caso de *Geo-First*). Estos nombres se pueden emplear también para estructuras híbridas. De tal modo que si la estructura emplea primero el índice textual es de tipo *Text-First* y si la estructura emplea primero el índice espacial es de tipo *Geo-First*. En la figura 1 mostramos las tres propuestas básicas de la indexación en recuperación de información geográfica. Las tres estructuras emplean un índice textual (sombreado en la figura) y un índice espacial (sin sombrear). La estructura de la izquierda es de doble índice, ya que mantiene por separado un índice textual y un índice espacial, mientras que las otras dos son estructuras híbridas. La estructura del centro pertenece a la categoría de *Geo-First*, ya que se accede primero al índice espacial, y la estructura de la derecha a la de *Text-First*, ya que se accede primero al índice textual.

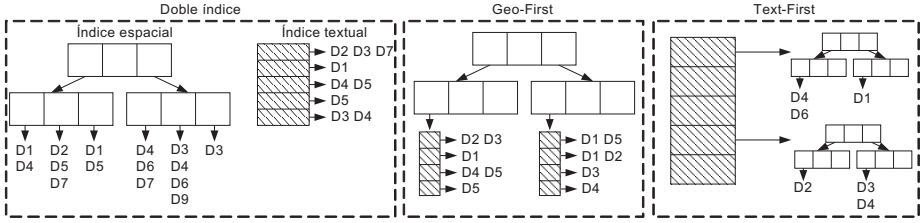


**Figura 1.** Estructuras de indexación básicas en sistemas GIR.

En [11] los autores proponen emplear un índice invertido y un R-tree (el ejemplo paradigmático y sin duda el método de acceso espacial más empleado). Realizan pruebas combinándolos de las tres formas que acabamos de describir y en sus experimentos concluyen que mantener por separado los índices es menos eficiente (esta misma conclusión había sido obtenida en [7]) y que sus estructuras son más eficientes que las que emplean la estructura de *grid*.

En [10], los autores comparan tres estructuras que combinan un índice invertido con la estructura *grid*, con el R-tree y con curvas de llenado del espacio [12,13]. Las curvas de llenado del espacio se basan en el almacenamiento de los objetos espaciales en un orden determinado por la forma de la curva de llenado. La conclusión de los autores es que la estructura que emplea estas curvas de llenado del espacio mejora el rendimiento de las otras aproximaciones.

Finalmente, en el proyecto STEWARD [14], los autores proponen emplear una estructura que mantenga por separado un índice invertido y un *Quad-tree*

[15]. El Quad-tree es una estructura similar al *grid* (ambas son dirigidas por el espacio y no por los objetos a indexar), que va dividiendo el espacio en cuadrantes hasta que los objetos se pueden almacenar en una página de disco. Además, en este trabajo se propone el empleo de un optimizador de consultas que decida si emplear primero el índice textual o el índice espacial en función de la previsión de cuál va a devolver menos resultados. Para poder emplear este optimizador de consultas el sistema debe almacenar estadísticas que permitan realizar la estimación del número de documentos resultante de una búsqueda de términos clave particular o de una ventana de consulta espacial determinada.

En cuanto al trabajo de otros grupos españoles, en [16] se describen los sistemas desarrollados por tres de los principales grupos del país y un mecanismo para fusionar sus resultados mejorando la eficacia individual de cada uno.

Un inconveniente de la mayoría de estas aproximaciones es que las estructuras empleadas en la parte espacial no tienen en consideración la jerarquía del espacio geográfico. Es decir, los nodos internos en las estructuras de indexación espacial no tienen significado en el mundo real (son significativos únicamente para la propia estructura). Por ejemplo, supongamos que queremos construir un índice para una colección de países, regiones y ciudades. Estos objetos se encuentran estructurados en una relación topológica de contenido; es decir, una ciudad se encuentra contenida en una región que a su vez está contenida en un país. Si construimos un R-tree (o cualquier otro método de acceso espacial) los nodos internos de la estructura no representan regiones ni países y, por tanto, la jerarquía del espacio no se mantiene en el índice. No es posible asociar ningún tipo de información con el nodo de una región y que las ciudades que pertenecen a esa región hereden automáticamente esa información ya que no hay relación de ningún tipo entre una región y sus ciudades en la estructura de indexación.

Una estructura que puede describir adecuadamente las características específicas del espacio geográfico es una ontología [17] (i.e. una especificación explícita y formal de una conceptualización compartida). Una ontología proporciona un vocabulario de clases y relaciones para describir un ámbito determinado. En [18], 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 [19,20] los autores describen cómo se emplean ontologías en tareas de expansión de los términos de las consultas, en la elaboración de rankings de relevancia y en la anotación de recursos web en el proyecto SPIRIT. Nuestra principal aportación al campo consiste en una estructura de indexación que para la parte espacial se basa en una descripción ontológica del espacio geográfico [21]. En la siguiente sección describimos en más detalle la ontología y la estructura de indexación basada en ella.

## 3. Estructura de indexación ontológica

Antes de desarrollar nuestra estructura hemos definido una ontología espacial, accesible en la URL http://lbd.udc.es/ontologies/spatialrelations, empleando la especie OWL-DL de OWL [22]. Las clases OWL se pueden interpre-

tar como conjuntos que contienen individuos (también conocidos como instancias). A su vez, estos individuos se pueden considerar como instancias de clases. Nuestra ontología describe ocho clases de interés: *SpatialThing*, *GeographicalThing*, *GeographicalRegion*, *GeopoliticalEntity*, *PopulatedPlace*, *Region*, *Country* y *Continent*. Además, existen relaciones jerárquicas entre *SpatialThing*, *GeographicalThing*, *GeographicalRegion* y *GeopoliticalEntity* ya que *GeopoliticalEntity* es subclase de *GeographicalRegion*, *GeographicalRegion* es subclase de *GeographicalThing* y *GeographicalThing* es subclase de *SpatialThing*.

Es decir, estas cuatro clases están organizadas en una jerarquía de especialización superclase – subclase, también conocida como *taxonomía*. Las subclases especializan (están *subsumidas por*) sus superclases *GeopoliticalEntity* tiene cuatro subclases: *PopulatedPlace*, *Country*, *Continent* y *Region*, y todos los individuos son miembros de esas subclases. Estas cuatro subclases tienen necesariamente una condición de aserción relativa a sus relaciones con cada una de las otras. Están conectadas por la propiedad *spatiallyContainedBy* que describe la existencia de una relación de contenido espacial entre ellas. Por ejemplo, todos los individuos de la clase *PoulatedPlace* están espacialmente contenidos (*spatiallyContainedBy*) en individuos de la clase *Region* (esto se describe en OWL como *PopulatedPlace spatiallyContainedBy only (AllValuesFrom) Region*).

La formalización de la ontología nos permite la definición de una estructura de indexación basada en ella. Por tanto, la estructura que proponemos es un árbol con cuatro niveles, cada uno de ellos correspondiente a una de las subclases de *GeopoliticalEntity*. El nivel superior del árbol contiene un nodo por cada una de las instancias de la clase *Continent*. A su vez, cada nodo en ese nivel referencia las instancias de la clase *Country* que están conectadas por medio de la relación de contenido espacial (*spatiallyContainedBy*). Los niveles correspondientes a *Region* y a *PopulatedPlace* los construimos empleando la misma estrategia. Es decir, la estructura del árbol sigue la taxonomía de la ontología. La figura 2 muestra la estructura de indexación espacial construida con las instancias de nuestra ontología.

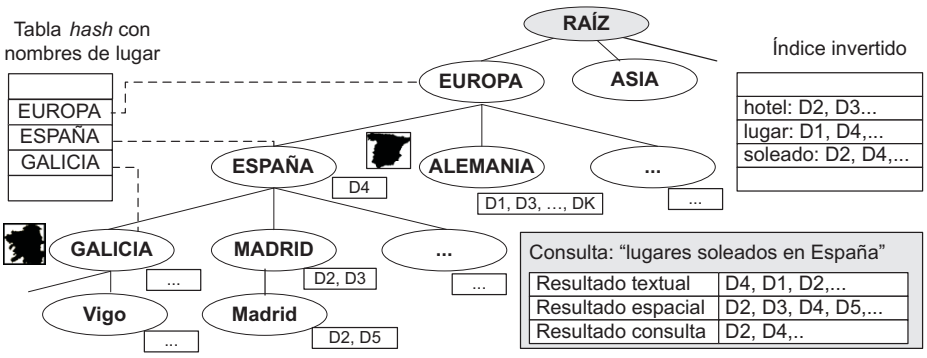


**Figura 2.** Ejemplo de la estructura de indexación.

Cada nodo contiene la siguiente información, además de la lista de nodos hijo que están contenidos espacialmente por él (*spatiallyContainedBy*): (i) la palabra clave (un nombre de lugar), (ii) el MBR de la geometría (i.e., el rectángulo más pequeño que la contiene) que representa ese lugar y (iii) una lista con los identificadores de los documentos que incluyen referencias geográficas a ese lugar. Además, cada nodo contiene un R-tree que mejora el rendimiento del acceso desde un nodo a sus hijos que satisfacen una consulta.

Al ser una estructura en forma de árbol y almacenar los MBRs en cada nodo, podemos usar la estructura para resolver consultas espaciales empleando el mismo algoritmo que se utiliza con las estructuras de indexación espacial clásicas. Dicho algoritmo consiste en descender a través de la estructura descartando aquellas ramas del árbol que no intersecan con la ventana de consulta. Los nodos del árbol que continúan tras el descenso constituyen el resultado de la consulta.

La ventaja principal de esta estructura de indexación espacial sobre otras alternativas es que los nodos intermedios de la estructura tienen significado en el espacio geográfico y pueden tener información adicional asociada. Por ejemplo, podemos asociar una lista de documentos que referencien a un país determinado y usar esa lista de documentos para resolver consultas que combinen una parte textual y una parte espacial. Además, dado que existe una relación superclase – subclase entre niveles, los niveles inferiores pueden heredar las propiedades asociadas con sus niveles superiores. En particular, los documentos asociados con un nodo de la estructura también se refieren a todos los nodos en el subárbol que parte de él. Esto permite que nuestra estructura de indexación puede realizar fácilmente expansión de los términos de consulta sobre referencias geográficas. Consideremos la consulta "*recuperar todos los documentos que se refieran a* España". El índice de nombres de lugar se emplea para localizar el nodo interno que representa el objeto geográfico correspondiente a *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 se ha empleado 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*).

Otra ventaja es que la estructura es general en el sentido de que la ontología del espacio geográfico se puede adaptar para cada aplicación en particular. Por ejemplo, si una aplicación en concreto usa un área restringida del espacio geográfico donde las clases *Continent* y *Country* no son necesarias pero, en cambio, son necesarias las clases *Province*, *Municipality*, *City* y *Suburb*, podemos definir una ontología del espacio diferente y basar la estructura de indexación en ella ya que la relación *spatiallyContainedBy* continúa siendo válida entre las clases. Finalmente, podemos definir relaciones espaciales adicionales en la ontología, como

por ejemplo la de adyacencia (*spatiallyAdjacent*), y mantener esas relaciones en la estructura de indexación para mejorar las capacidades de consulta del sistema.

## 4. Consultas y relevancia

Una de las características más importantes de toda estructura de indexación es el tipo de consultas que permite resolver. En nuestra opinión los tipos de consulta más relevantes en un sistema de recuperación de información geográfica son los siguientes:

- *Consultas textuales puras.* Estas son consultas del tipo "*recuperar todos los documentos donde aparezcan las palabras* hotel *y* mar". Son las consultas típicas del campo de la recuperación de información. En nuestra estructura se resuelven empleando el índice invertido que forma parte de la misma y la relevancia se puede calcular empleando alguna de las múltiples variantes de la conocida fórmula $tf \times idf$.
- *Consultas espaciales puras.* 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 un objeto complejo como un polígono. Estas consultas, procedentes del campo de los GIS, adquieren un nuevo enfoque en los sistemas GIR al definir el concepto de relevancia espacial. En estos sistemas, los objetos en el resultado pertenecen al mismo con un valor de relevancia (es decir, no se hace simplemente la distinción entre objetos que pertenecen al resultado y objetos que no lo hacen). Dado que la estructura espacial que proponemos almacena en cada nodo el MBR del objeto geográfico correspondiente es posible resolver estas consultas empleando el clásico algoritmo de estructuras de indexación espacial descrito en la sección anterior. Además, podemos calcular la relevancia espacial de un documento teniendo en cuenta (mediante una función de agregación como el máximo) las relevancias individuales de cada uno de los referentes geográficos citados en el documento con respecto a la consulta. La ecuación 1 permite calcular la relevancia de un documento $d$ para la consulta $q$ debida al lugar $l$. Intuitivamente esta fórmula combina (ponderando mediante pesos) la relevancia debida al área de solape entre el MBR del lugar y la ventana de consulta, y la debida a la distancia al centro del foco de atención de la consulta. Además, toda la relevancia está ponderada por la importancia intrínseca del lugar (un parámetro precalculado que permite definir por ejemplo que *a priori* un documento que mencione *Londres* se refiere a la ciudad del *Reino Unido* y no a la de *Canadá*).

$$relevancia_{q,d,l} = \frac{w_{dc} * dc_{q,l} + w_{as} * as_{q,l}}{importancia} \quad (1)$$

- *Consultas textuales sobre un área geográfica.* En este caso se proporciona un área geográfica de interés junto con el conjunto de palabras. Un ejemplo de este tipo de consultas es "*recuperar todos los documentos con la palabra* hotel

*que se refieren a la siguiente área geográfica*". Al igual que en las consultas espaciales puras el área geográfica de la consulta puede ser un punto, una ventana de consulta o un objeto complejo. Dado que este tipo de consultas combina los dos anteriores el algoritmo más sencillo para resolverlas consiste en resolver la parte textual por un lado, la parte espacial por el otro y combinar los resultados de ambas subconsultas. En la bibliografía se pueden encontrar distintas fórmulas para combinar la relevancia textual de un documento con su relevancia espacial pero sin duda la suma ponderada de ambas es el método más sencillo y también el más empleado para dicho propósito.

- *Consultas textuales con nombres de lugar.* En este tipo de consultas algunos de los términos son nombres de lugar. Por ejemplo, "*recuperar todos los documentos con la palabra* hotel *referidos a España*". Estas consultas, que se realizan frecuentemente en sistemas de recuperación de información clásicos, alcanzan una nueva dimensión es sistemas GIR al poder aprovechar las características de la parte espacial de la consulta. El algoritmo para resolver estas consultas consiste en detectar los nombres de lugar mencionados en la consulta y traducirlos, mediante la tabla con nombres de lugar, a nodos internos de la estructura. Todos los documentos referenciados en dichos nodos son relevantes para la consulta pero también lo son todos los documentos referenciados en los subárboles que comienzan en ellos. Para calcular la relevancia espacial en este caso se debe tener en cuenta la profundidad a la que se encuentra cada nodo dentro del subárbol correspondiente ponderando siempre el valor por la importancia intrínseca del nodo.

Aunque estos son los tipos de consulta básicos y proporcionan una buena aproximación a cualquier otro tipo de consulta en sistemas GIR, es posible tener en cuenta otros tipos de consulta más específicos en la parte espacial. A continuación describimos algunos de ellos y esbozamos los algoritmos o las modificaciones que se deben realizar sobre la estructura para resolverlas.

- *Documentos geo-referenciados en los $k$-lugares más cercanos al de consulta.* Esta consulta, conocida como $k$-NN en métodos de acceso espacial (los $k$ vecinos más próximos), se puede resolver en nuestra estructura mediante un procedimiento de refinamiento empezando a buscar con una ventana de consulta pequeña que se va expandiendo hasta llegar al cupo de los $k$ vecinos. Sin embargo, dada la naturaleza semántica de nuestra estructura podemos reflejar más relaciones topológicas como puede ser la de adyacencia que nos ayuden en la resolución de la consulta. Otras relaciones como la de proximidad podrían funcionar mejor para este caso determinado aunque su definición es más subjetiva y aportan menos claridad al modelo topológico que define nuestra estructura (la relación topológica de adyacencia es mucho más común). Además, la relación de adyacencia nos permite resolver también el siguiente tipo de consultas bastante habitual en GIS cuando se consideran modelos topológicos.
- *Documentos geo-referenciados en lugares adyacentes al de consulta (o en el de consulta).* Un claro ejemplo de este tipo de consultas consiste en recuperar

todos los documentos sobre puntos de interés turístico que se encuentren en el ayuntamiento donde se aloja un turista o en los ayuntamientos adyacentes. Este tipo de consulta se pude resolver de manera sencilla en nuestra estructura si contemplamos la relación topológica de adyacencia. Es importante tener en cuenta que al añadir nuevas relaciones el árbol original se convierte en un grafo. En la siguiente sección veremos algunas consideraciones a tener en cuenta para la implementación eficiente de la estructura en ambos casos.

- *Documentos geo-referenciados en lugares que se encuentran en una cierta orientación (al norte, sur, noreste, etc.) respecto a la consulta.* Por ejemplo, cuando el mismo turista del ejemplo anterior quiere visitar a un familiar que vive al noreste de su hotel puede estar interesado en consultar todos los puntos de interés turístico que se encuentren al noreste de su hotel. Una aproximación para resolver este tipo de consultas consiste en emplear una matriz de transformación para las orientaciones más habituales y resolver una consulta de tipo región en la zona correspondiente a la orientación consultada. En el ejemplo, la matriz de transformación determinará que la esquina inferior izquierda de la región se debe posicionar en el lugar de consulta.

## 5. Implementación compacta

Realizar una implementación eficiente y compacta de esta estructura se puede descomponer en realizar una implementación eficiente y compacta de las dos partes fundamentales: el índice invertido y la estructura de indexación espacial. La compresión del índice invertido ha sido muy estudiada en los últimos años y es bien conocido que puede llegar a necesitar sólo un 20 % de espacio adicional sobre el texto [1]. Además, existen técnicas más novedosas de compresión que permiten representar de manera conjunta el texto y el índice invertido ocupando tan solo un 35 % del espacio necesario para el texto original [23,24].

Por otra parte, la estructura espacial consiste en un árbol donde en cada nodo sólo se necesitan las operaciones de acceder a los hijos y al padre. La implementación tradicional de este tipo de estructuras de árbol desperdicia mucho espacio al emplear punteros para permitir el acceso a los hijos (en un árbol de $n$ nodos esto supone que cada puntero necesita $\log n$ bits). Sin embargo, existen técnicas más novedosas para representar estas estructuras de árbol empleando estructuras de datos compactas que necesitan tan sólo $2n + o(n)$ bits y permiten resolver las operaciones que necesitamos en tiempo constante. Las estructuras más conocidas son las de paréntesis balanceados (BP), LOUDS y DFUDS (en [25] puede encontrarse un buen resumen y una comparación de su rendimiento en la práctica). Por tanto, empleando dichas estructuras y almacenando la información satélite (el MBR del objeto geográfico, su topónimo, etc.) en estructuras auxiliares indexables por el identificador del nodo podemos representar la estructura completa de forma compacta. Además, podemos emplear técnicas similares a las de la compresión de las listas de ocurrencias para comprimir las listas de documentos geo-referenciados en cada nodo.

En la sección anterior describimos ciertas operaciones que se pueden resolver de manera más eficiente contemplando otras relaciones topológicas además de

la de contenido espacial. Contemplar todas estas relaciones de manera conjunta resulta en que el árbol original se convierte en un grafo y, por tanto, las representaciones que acabamos de nombrar no nos sirven. Sin embargo, las representaciones compactas de grafos han sido también muy estudiadas en los últimos años debido sobre todo a la importancia del grafo de la web o de los grafos que representan las redes sociales. Por ejemplo, el K2-tree [26], que permite representar la matriz de adyacencia de un grafo de forma comprimida, puede ser extendido fácilmente para representar todas la relaciones topológicas en nuestra estructura incluyendo además atributos en cada relación (por ejemplo una determinada orientación como *norte* o *sureste* en la relación de adyacencia).

## 6. Conclusiones y trabajo futuro

La recuperación de información geográfica se está consolidando como una prometedora extensión al campo de la recuperación de información debido fundamentalmente a la demanda de servicios donde poder consultar y representar información en mapas. En este artículo revisamos el estado del arte del área presentando nuestras aportaciones al mismo y describiendo los avances y las nuevas líneas de investigación que se han ido abriendo.

La principal carencia de la mayoría de las estructuras de indexación para sistemas GIR es el no tener en cuenta las características propias de la naturaleza espacial de los referentes geográficos. Este problema está presente en los índices invertidos, donde los referentes geográficos se consideran términos de la misma naturaleza que el resto de las palabras en el documento, pero también en las más recientes estructuras para GIR que contemplan sólo de manera parcial dichas características. Por ejemplo, no tienen en cuenta la naturaleza jerárquica del espacio ni otras relaciones topológicas existentes entre los referentes geográficos indexados. Nuestra principal aportación al área consiste en una estructura espacial que solventa estas carencias al estar basada en una ontología del espacio geográfico que permite representar todas las relaciones topológicas.

Esta estructura nos permite resolver nuevos tipos de consulta que no serían posibles (o las soluciones serían complicadas y poco elegantes) empleando métodos de acceso espacial clásicos que no contemplan las relaciones topológicas existentes entre los elementos indexados. Una línea de trabajo interesante consiste en el estudio de las posibles relaciones topológicas que se pueden representar en nuestra estructura y cómo afectan a los tipos de consulta existentes (o si permiten resolver nuevos tipos de consultas). Además, la implementación eficiente tanto en espacio como en tiempo de esta estructura constituye un interesante reto algorítmico. Finalmente, su integración en sistemas de recuperación de información existentes constituye el principal objetivo a más largo plazo.

## Referencias

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)

2. Worboys, M.F.: GIS: A Computing Perspective. CRC (2004) ISBN: 0415283752.
3. Jones, C.B., Purves, R.S.: Geographical information retrieval. In: Encyclopedia of Database Systems. (2009) 1227–1231
4. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comput. Surv. **38**(2) (2006) 6
5. Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A.N., Theodoridis, Y.: R-Trees: Theory and Applications. Springer-Verlag New York, Inc. (2005)
6. Gaede, V., Günther, O.: Multidimensional access methods. ACM Comput. Surv. **30**(2) (1998) 170–231
7. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-Textual Indexing for Geographical Search on the Web. In: Proc. of SSTD. (2005) 218 – 235
8. Nievergelt, J., Hinterberger, H., Sevcik, K.C.: The grid file: An adaptable, symmetric multi-key file structure. In: Proc. of the ECI Conference. (1981) 236–251
9. Martins, B., Silva, M.J., Andrade, L.: Indexing and ranking in Geo-IR systems. In: Proc. of GIR, ACM Press (2005) 31–34
10. Chen, Y.Y., Suel, T., Markowetz, A.: Efficient query processing in geographic web search engines. In: Proc. of SIGMOD. (2006) 277–288
11. Zhou, Y., Xie, X., Wang, C., Gong, Y., Ma, W.Y.: Hybrid index structures for location-based web search. In: Proc. of CIKM, ACM (2005) 155–162
12. Morton, G.M.: A computer oriented geodetic data base and a new technique in file sequencing. Technical report, IBM Ltd. (1966)
13. Böhm, C., Klump, G., Kriegel, H.P.: Xz-ordering: A space-filling curve for objects with spatial extension. In: Proc. of the SSD Conference. (1999) 75–90
14. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEWARD: Architecture of a Spatio-Textual Search Engine. In: ACMGIS. (2007) 186 – 193
15. Nelson, R.C., Samet, H.: A consistent hierarchical representation for vector data. In: Proc. of the SIGGRAPH Conference. (1986) 197–206
16. Buscaldi, D., Perea-Ortega, J.M., Rosso, P., López, L.A.U., Ferrés, D., Rodríguez, H.: Geotextmess: Result fusion with fuzzy borda ranking in geographical information retrieval. In: Proc. CLEF. (2008) 867–874
17. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition **5**(2) (June 1993) 199 – 220
18. Dellis, E., Paliouras, G.: Management of Large Spatial Ontology Bases. In: Proc. of ODBIS. (September 2006)
19. Jones, C.B., Abdelmoty, A.I., Fu, G.: Maintaining ontologies for geographical information retrieval on the web. In: Proc. of ODBASE. (2003)
20. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-Based Spatial Query Expansion in Information Retrieval. In: Proc. of ODBASE. (2005) 1466 – 1482
21. Brisaboa, N.R., Luaces, M.R., Places, A.S., Seco, D.: Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. GeoInformatica **14**(3) (2010) 307–331
22. World Wide Consortium: Owl web ontology language reference. (2008) Fecha de consulta: Marzo de 2008. Disponible en: http://www.w3.org/TR/owl-ref/.
23. Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes: Compressing and Indexing Documents and Images. Academic Press (1999)
24. Ziviani, N., de Moura, E.S., Navarro, G., Baeza-Yates, R.A.: Compression: A key for next-generation text retrieval systems. IEEE Computer **33**(11) (2000) 37–44
25. Arroyuelo, D., Cánovas, R., Navarro, G., Sadakane, K.: Succinct trees in practice. In: Proc. 11th ALENEX, SIAM Press (2010) 84–97
26. Brisaboa, N.R., Ladra, S., Navarro, G.: K2-trees for compact web graph representation. In: Proc. 16th SPIRE - LNCS 5721. Volume 5721. (2009) 18–30