# An Index Structure To Retrieve Documents With Geographic Information *

Miguel R. Luaces[1], Jose R. Paramá[1], Oscar Pedreira[1], Diego Seco[1], Jose R. R. Viqueira[2]

[1] *Database Laboratory, University of A Coruña*
*Campus de Elviña, 15071 A Coruña, Spain*
*{luaces, parama, opedreira, dseco}@udc.es*

[2] *Systems Laboratory, University of Santiago de Compostela*
*Constantino Candeira s/n, 15782 Santiago de Compostela, Spain*
*joserios@usc.es*

## Abstract

*Both* Geographic Information Systems *and* Information Retrieval *have been very active research fields in the last decades. Lately, a new research field called* Geographic Information Retrieval *has appeared from the intersection of these two fields. The main goal of this field is to define index structures and techniques to efficiently store and retrieve documents using both the text and the geographic references contained within the text.*

*We present in this paper the architecture of a system for geographic information retrieval. It defines a workflow for the extraction of the geographic references in the document. In addition, a new index structure is defined that combines an inverted index, a spatial index, and an ontology. This structure improves the query capabilities of other proposals.*

## 1. Introduction

Although the research field of Information Retrieval [2] has been active for the last decades, the growing importance of Internet and the World Wide Web have made it one of the most important research fields nowadays. Many different index structures, compression techniques and retrieval algorithms have been proposed in the last few years. More importantly, these proposals have been widely used in the implementation of document databases, digital libraries and web search engines.

Another field that has received much attention during the last years is the field of Geographic Information Systems [17]. Recent improvements in hardware have made the implementation of this type of systems affordable for many organizations. Furthermore, a cooperative efort has been undertaken by two international organizations (ISO [9] and the Open Geospatial Consortium [15]) to define standards and specifications for interoperable systems. This effort is making possible that many public organizations are working on the construction of spatial data infrastructures [1] that will enable them to share their geographic information.

Many of the documents stored in digital libraries and document database include geographic references within their texts. For example, news documents reference the place where the event happened and often the place where the document has been written. Geographic references can also be attached to web pages by using information from the text, the location of the web server, and many other information elements. However, the geographic references of documents are rarely used in information retrieval systems. Few index structures or retrieval algorithms take into account the spatial nature of geographic references embedded within documents. Pure textual techniques focus only on the language aspects of the documents and pure spatial techniques focus only on the geographic aspects of

*This work has been partially supported by "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, by "Xunta de Galicia" ref. PGIDIT05SIN10502PR and ref. 2006/4, by "Ministerio de Educación y Ciencia" ref. AP-2006-03214 (FPU Program) for Oscar Pedreira, and by "Dirección Xeral de Ordenación e Calidade do Sistema Universitario de Galicia, da Consellería de Educación e Ordenación Universitaria-Xunta de Galicia" for Diego Seco.

the documents. None of them are suitable for a combined approach to information retrieval because they completely neglect the other type of information. As a result, there is a lack of system architectures, index structures and query languages that combine both types of information.

Some proposals have appeared recently [3, 13] that define new index structures that take into acount both the textual and the geographic aspects of a document. However, there are some specific particularities of geographic space that are not taken into account by these approaches. Particularly, concepts such as the hierarchical nature of geographic space and the topological relationships between the geographic objects must be considered in order to fully represent the relationships between the documents and to allow new and interesting types of queries to be posed to the system.

In this paper, we present a system architecture and an index structure that takes these issues into account. First, some basic concepts and related work are described in Section 2. Then, in Section 3, we present the general architecture of the system and describe its components. The system architecture defines a workflow for constructing a document database where both the words and the geographic references in the documents are considered. Section 4 describes the index architecture in more detail. The index structure is located at the core of the system architecture and enables the system to store and access efficiently the documents using both their textual references and their geographic ones. Then, in Section 5 we describe some types of queries that can be answered with this system and we sketch the algorithms that can be used to solve this queries. Finally, Section 6 presents some conclusions and future lines of work.

## 2. Related Work

Inverted indexes are considered the classical text indexing technique. An inverted index associates to each word in the text (organized as a *vocabulary*) the list of pointers to the positions where the word appears in the documents. The set of all those list is called the *occurrences* [2]. The main drawback of these indexes is that geographic references are completely ignored. Place names are just considered words.

Many different spatial index structures have been proposed along the years. A good survey of these structures can be found in [6]. The main goal of spatial index structures is improving access time to collections of geographic data objects. One of the most popular spatial index structure and a paradigmatic example is the R-tree [8]. The R-tree is a balanced tree derived from the B-tree which splits space in hierarchically nested, possibly overlapping, minimum bounding rectangles. The number of children of each internal node varies between a minimum and a maximum. The tree is kept in balance by splitting overflowing nodes and merging underflowing nodes. Rectangles are associated with the leaf nodes, and each internal node stores the bounding box of all the rectangles in its subtree. The decomposition of space provided by an R-tree is adaptive (dependent on the rectangles stored) and overlapping (nodes in the tree may represent overlapping regions). A drawback of these structures is that they do not take into consideration the hierarchy of space. Internal nodes in the structure are meaningless in the real world, they are just meaningful for the index structure. For example, imagine that we want to build an index for a collection of countries, provinces, and cities. These objects are structured in a topological relationship of containment, that is, a city is contained within a province that is itself contained within a country. If we build an R-Tree with these geographic objects the containment hierarchy will not be maintained.

Some work has been done to combine both types of indexes. The papers about the SPIRIT project (Spatially-Aware Information Retrieval on the Internet) [12, 10, 11, 16, 5] are a very good starting point to begin with. In [16], the authors conclude that keeping separate text and spatial indexes, instead of combining both in one, results in less storage costs but it could lead to higher response times. More recently, [13, 3] survey this work and propose improvements to the system and the algorithms defined. In their work they propose two naive algorithms: *Text-First* and *Geo-First*. Both algorithms use the same strategy, one index is first used to filter the documents (inverted index in Text-First and spatial index in Geo-First). The resulting documents are sorted by their identifiers and then filtered using the other index (spatial index in Geo-First and inverted index in Text-First). Nevertheless, none of these approach take into account the relationships between the geographic objects that they are indexing.

A structure that can properly describe the specific characteristic of geographic space is an *ontology*, which is a formal explicit specification of a shared conceptualization [7]. An ontology provides a vocabulary of classes and relations to describe a given scope. In [4], a method is proposed for the efficient management of large spatial ontologies using a spatial index to improve the efficiency of the spatial queries. Furthermore, in [10, 5] the authors describe how ontologies are used in query term expansion, relevance ranking, and web resource annotation in the SPIRIT project. However,

as far as we know, nobody has ever tried to combine ontologies with other types of indexes to have a hybrid structure.

## 3. System Architecture

Figure 1 shows our proposal for the system architecture of a geographic information retrieval system. The bottom part of the figure shows the document storage workflow. The first step of this workflow is the *keyword extraction* task where all documents are parsed and relevant keywords are extracted. Classic information retrieval techniques can be employed in this task to reduce the number of keywords such as removing stopwords, and using other text operations such as stemmers and reduction to noun groups [2].

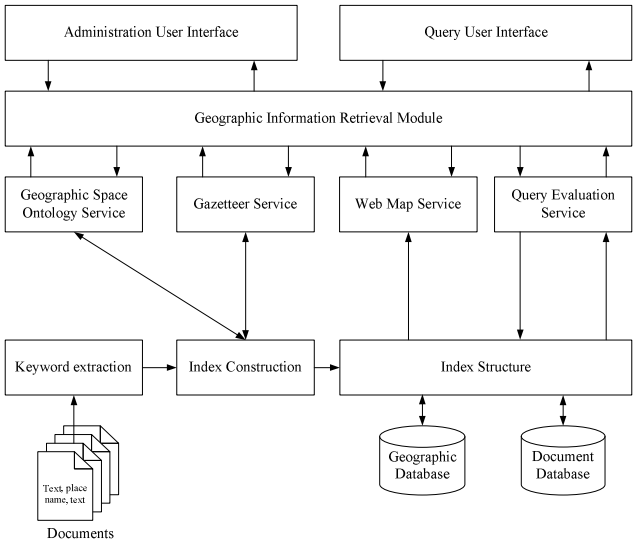


**Figure 1. System Architecture**

After extracting the keywords, the system is ready to build the index structure. For this task, two services are needed. First, all keywords of the document have to be processed by a *gazetteer service* in order to discover whether the keyword is a place name. In that case, the geographic references associated to the place name are stored together with the keyword. Then, an *ontology of the geographic space* is used together with the keywords and the geographic references to build the index structure. This process is described in more detail in Section 4.

The processing services are shown in the middle of the figure. On the left, the aforementioned *geographic space ontology service* and the *gazetteer service* can be seen. On the right, one can see the two services that are used to solve queries. The rightmost one is the *query solving service*, which receives queries and uses the index structure to solve them. The other service is a *web map service* following the OGC specification [14] that is used to create cartographic representations of the query results. On top of this services a *geographic information retrieval module* is in charge of coordinating the task performed by each service to response the user requests.

The topmost layer of the architecture shows the user interface. The system has two different user interfaces: an administration user interface that can be used to manage the document collection, and a query user interface that can be used to pose queries to the system and browse the results.

## 4. The Index Structure

Figure 2 shows the index structure. The base of this structure is a spatial ontology. This ontology models both the vocabulary and the spatial structure of places for purposes of information retrieval. The structure of an ontology is fixed so our index structure must be constructed ad-hoc for the concrete domain which it will be used.

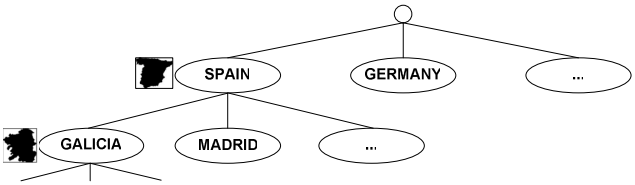


**Figure 2. Index Structure**

The main component of the index structure is a tree composed by nodes that represent place names. These nodes are interconnected by means of inclusion relationships (for instance, Galicia is included in Spain). In each node we store: (i) the keyword (a place name), (ii) the geographic references associated to the place name, (iii) the bounding box of the geometry representing this place, (iv) a list with the document identifiers of the documents that include geographic references to this place, and (v) a list of children nodes that are geographically within this node. If the list of children nodes is very long, using sequential access is very unefficient. For this reason, if the number of children nodes exceeds a threshold, an R-Tree is used instead of a list.

Two auxiliary structures are used in the index. First, a *place name hash table* stores for each place name its position in the index structure. This provides direct access to a single node by means of a keyword

that is returned by the Gazetteer Service if the word processed is a place name. The second auxiliary structure is a traditional inverted index with all the words in the documents that is used to solve textual queries.

Keeping separate indexes for text and geographical scopes has many advantages. First of all, all textual queries can be efficiently processed by the inverted index, and all spatial queries can be efficiently processed by the index structure. Moreover, queries combining textual and spatial aspects are supported. Furthermore updates in each index are handled independently, which makes the addition and removal of data easier. Finally, specific optimizations can be applied to each individual indexing structure.

The main drawbacks of this structure are: (i) the tree that supports the structure is possibly unbalanced penalizing the efficiency of the system, and (ii) ontologies have a fixed structure and thus our structure is static and it must be constructed *ad-hoc*.

## 5. Supported Query Types

The most important characteristic of an index structure is the type of queries that can be solved with it. The following types of queries are relevant in a geographic information retrieval system:

- *Pure textual queries.* These are queries such as "*retrieve all documents where the words* hotel *and* sea *appear*".
- *Pure spatial queries.* An example of this type of queries is "*retrieve all documents that refer to the following geographic area*". The geographic area in the query can be a point, a query window, or even a complex object such as a polygon.
- *Textual queries with place names.* In this type of queries, some of the words are place names. For instance, "*retrieve all documents with the word* hotel *that refer to Spain*".
- *Textual queries over a geographic area.* In this case, a geographic area of interest is given in addition to the set of words. An example is "*retrieve all documents with the word* hotel *that refer to the following geographic area*".

Inverted indexes can solve pure textual queries by retrieving from the inverted index the lists of documents associated to each word and then performing the intersection of the lists. Pure spatial queries can be solved by spatial indexes by descending the structure and taking into consideration only those nodes whose bounding box intersects with the geographic area of the query. This operation returns a set of candidate documents that has to be refined with the actual geographic reference in order to decide whether the document is part of the result or not.

Pure textual queries can be solved by our system because an inverted index is part of the index structure. Similarly, pure spatial queries can also be solved because the index structure is built like a spatial index. Each node in the tree is associated with the bounding box of the geographic objects in its subtree. Therefore, the same algorithm that is used with spatial indexes can be used with our structure.

However, the index structure that we propose can be used to solve the third and fourth types of queries, which cannot be easily solved using an inverted index and a spatial index. For the case of the query with place names, our system can discover that *Spain* is a geographic reference by querying the Gazetteer service and then we can use the place name hash table in the structure to retrieve the index node that represents *Spain*. Thus, we save some time by avoiding a tree traversal.

Regarding the fourth type of query, the inverted index is used to retrieve the list of documents that contain the words, and the index structure is used to compute the list of documents that reference the geographic area. Then, the intersection of both lists is the result to the query. The advantage of our proposal in this case is that geographic references can be given using place names.

Another improvement over text and spatial indexes is that our index structure can easily perform query expansion on geographic references because the index structure is built from an ontology of the geographic space. Consider the following query "*retrieve all documents that refer to Spain*". The query evaluation service will discover that Spain is a geographic reference and the place name index will be used to quickly locate the internal node that represents the geographic object *Spain*. Then all the documents associated to this node are part of the result to the query. However, all the children of this node are geographic objects that are contained within Spain (for instance, the city of Madrid). Therefore, all the documents referenced by the subtree are also part of the result of the query. The consequence is that the index structure has been used to expand the query because the result contains not only those document that include the term *Spain*, but also all the documents that contain the name of a geographic object included in Spain (e.g., all the cities and regions of Spain).

## 6. Conclusions and Future Work

We have presented in this paper a system architecture for an information retrieval system that takes into account not only the text in the documents but also the geographic references included in the documents and the ontology of the geographic space. This is achieved by a new index structure that combines an inverted index, a spatial index and an ontology. We have also presented how traditional queries can be solved using the index structure. Finally, new types of queries that can be solved with the index structure are described and the algorithms that solve these queries are sketched.

We are currently finishing the implementation of a prototype of the system, and we are currently working on the evaluation of the performance of the index. Future improvements of the index structure are possible. First, a procedure must be defined to decide whether the children of a node must be structured as a list or as an R-Tree. Another line of future work involves exploring the use of different ontologies and determining how each ontology affects the resulting index. Furthermore, we plan on including other types of spatial relationships in the index structure in addition to inclusion (e.g., adjacency). These relationships can be easily represented in the ontology, and the index structure can be extended to support them. Finally, it is necessary to define algorithms to rank the documents retrieved by the system. For this task, we must define a measure of spatial relevance and combine it with the relevance computed using the inverted index.

## References


[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] Y.-Y. Chen, T. Suel, and A. 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] G. Fu, C. B. Jones, and A. 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] V. Gaede and O. Günther. Multidimensional access methods. *ACM Comput. Surv.*, 30(2):170–231, 1998.

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

[8] A. Guttman. R-trees: A dynamic index structure for spatial searching. In B. 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] C. B. Jones, A. I. Abdelmoty, and G. 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, ODBASEŠ03*, volume 2888 of *Lecture Notes in Computer Science*, 2003.

[11] C. B. Jones, A. I. Abdelmoty, G. Fu, and S. 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] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. 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] B. Martins, M. J. Silva, and L. 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 Document 01-068r3, Open GIS Consortium, Inc., 2002.

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

[16] S. Vaid, C. B. Jones, H. Joho, and M. 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.