New Discovery Methodologies in GIS: Improving the Information Retrieval Process

Nieves R. Brisaboa, Miguel R. Luaces, Diego Seco
Database Laboratory, University of A Coruña
Campus de Elviña, 15071 A Coruña, Spain
{brisaboa, luaces, dseco}@udc.es

ABSTRACT

In the last decade the availability of on-line resources, and also the number of users accessing those resources, have grown exponentially. The information retrieval process, which aims at the improvement of the access to such resources, has been the focus of interest of many researchers. The presence of geographic data in these repositories of information is surprisingly high (for example, note that most of the web pages about business contain information about the locations of their offices). In order to properly manage this geographic data the information retrieval process has been extended using architectures, data structures, and other techniques developed by the GIS community. This has meant the beginning of a new research field called Geographic Information Retrieval. In this chapter we present a study of the state-of-the-art of this new field and we also highlight the main open problems that will concentrate efforts during the next years.

INTRODUCTION

The need to manage information has been one of the key factors behind the consolidation of information technology as an essential driving force for the development of our society. Over the years, many system architectures, index structures, and other components have been proposed with the fundamental goal of allowing efficient access to information stored in huge document databases. The research field that focuses on this goal is called *Information Retrieval* (IR) (Baeza-Yates & Ribeiro-Neto, 1999) and it started with the work of Salton (1963). This research field has recently undergone a spectacular development motivated by the growth of the Internet and the need to search the Web. A very important feature of IR is that it deals with the problem of retrieving information by its content rather than its metadata. Thus, there are a number of techniques for retrieving documents of various types: texts, images, sound and video files, etc.

Textual information often includes *geographic references* in the text (for example, press releases usually mention the place where the events happened). Taking these geographic references into account provides added value to classic information retrieval systems. The research on *Geographic Information Systems* (GIS) (Worboys, 2004) has dedicated much effort to study the special features of geographic information and to develop systems able to use and take advantage of them. This field has received much attention in recent years due to recent improvements in hardware that have made possible the development of such systems by many organizations. In addition, two international organizations ISO (ISO/IEC, 2002) and the Open Geospatial Consortium (OGC, 2003) are undertaking a major collaborative effort to define standards and specifications to develop interoperable systems. At the European level, the INSPIRE (Infrastructure for Spatial Information in Europe) directive (European Commission, 2011) has enabled a breakthrough in the field of corporate GIS and it remarks the future importance of geographic information. Thanks to these initiatives, many public organizations are working in the development of spatial data infrastructures (GSDI, 2011) that enable them to share their spatial information.

These two research areas have progressed independently over the years. On the one hand, the index structures and techniques from the IR field do not take into account the spatial nature of geographic references that appear in text documents. On the other hand, spatial index structures are not directly applicable in information retrieval systems. However, users increasingly demand services that allow them to locate the information in its spatial context and even to access this information using queries that take into account the spatial information. These demands have caused that researchers in each area have began to pay attention to the other one resulting in a new research field called *Geographic Information Retrieval* (GIR). The aim of this field is to propose new system architectures, index structures, and other components in order to develop systems to *retrieve documents both thematically and geographically relevant in response to queries of the form <subject, place>*. An example of the type of queries studied in this new field is the following: "*Ph.D. dissertations regarding geographic information systems published in Spain*". The reader familiar with classic information retrieval systems knows that the relevance of the documents in a textual search engine is based on the frequency of the words that appear in the text of the documents. Therefore, if the word *Spain* does not appear explicitly in a document its relevance will be low with respect to this query. This happens even if the word *Madrid* appears in the document (or any other autonomous region, province or city of Spain) because traditional IR systems are not prepared to take into account the special characteristics of the geographic information space (e.g., the *contained by* spatial relationship between Madrid and Spain). Query expansion techniques in classical IR systems reformulate queries by adding new terms to the original query in an attempt to provide a better context (Baeza-Yates & Ribeiro-Neto, 1999) (for instance, in the previous example we could expand the query using the term *thesis* that is related with the term *dissertation*). Some examples of these query expansion techniques are: *term reweighting*, *local clustering*, *local context analysis*, or those based on *thesaurus*. However, all the particularities of the geographic references are not properly represented by any of them.

Among the topics of interest in the area of geographic information retrieval are the definition of system architectures, index structures, and other components that allow to model, capture, store, manipulate, access, retrieve, analyze, and display information efficiently. In addition, these tasks involve additional difficulties over the same tasks in the area of IR because of the special features and requirements of geographic information.

Despite the common geographical nature of the information, there are two fundamental differences on the requirements between GIS and GIR systems that must be taken into account. First, the spatial component of the queries in GIR system is much simpler than the queries that are usually posed to GIS because the latter involve complex spatial relationships. For example, a typical query in GIS can be *monuments located in municipalities adjacent to the location of a particular hotel*. In this case, the municipality where the hotel is located must be found, and then the monuments in adjacent municipalities must be retrieved. In contrast, a typical query to a GIR system can be *monuments in London* where the only spatial relationship to check is whether the spatial scope of the document lies within the geometry associated with London.

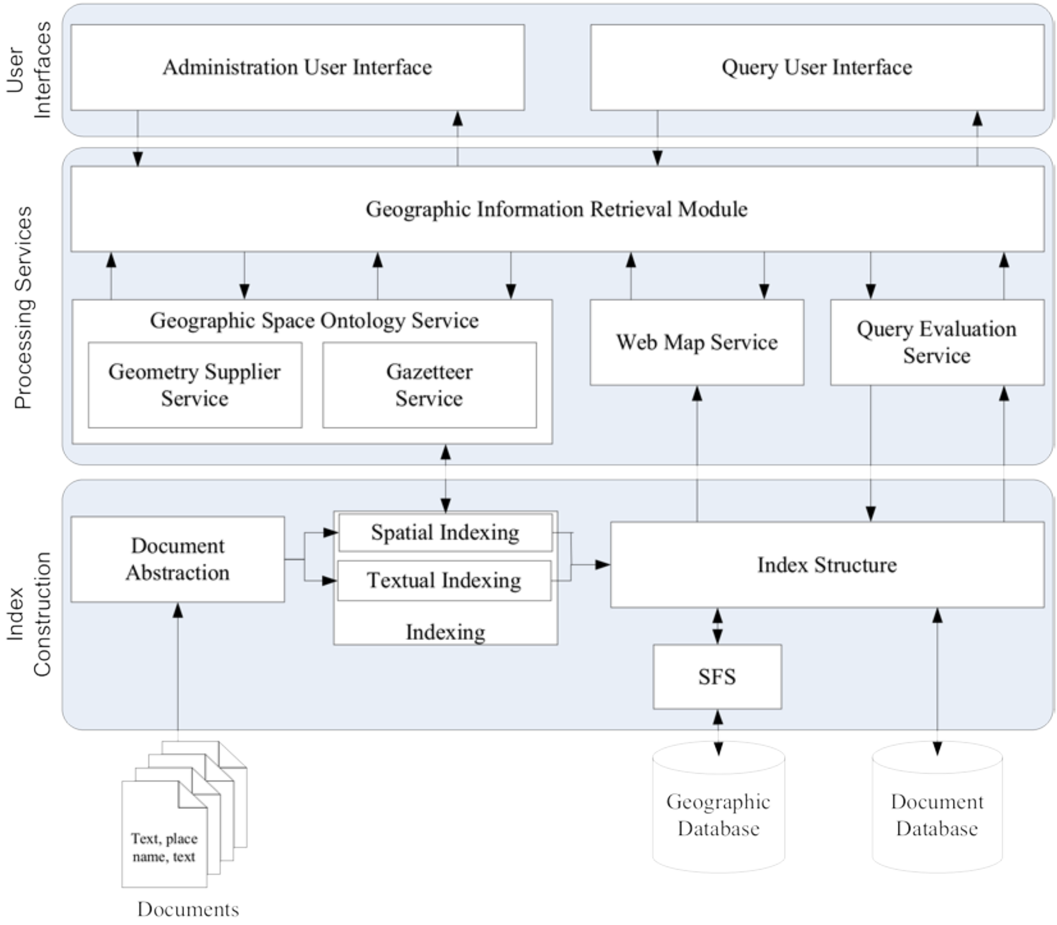
The second difference is related to the type of results expected for queries to both systems. In GIS queries, the expected results are similar to those of traditional databases, that is, objects in the database either belong to the result or not. However, in GIR systems, the expected results are similar to those of IR systems, that is, objects belong to the result with a certain probability. Following the example queries, in the GIS query only monuments in the municipalities adjacent are part of the results. On the other hand, in a GIR system query, monuments located in a city very close to London may be relevant to the user.

In this chapter, we review the most important contributions made in recent years to the field of geographic information retrieval as part of the description of our system architecture and index structure (Brisaboa et al., 2010).

SYSTEM ARCHITECTURE FOR GEOGRAPHIC INFORMATION RETRIEVAL

Figure 1 shows our proposal for the system architecture of a geographic information retrieval system. The architecture can be divided into three independent layers: the index construction workflow, the processing services, and the user interfaces. The influence of GIS architectures and spatial data infrastructures can be clearly identified. This influence is also reflected in the use of the standards WMS (OGC, 2002) (map generation) and SFS (OGC, 2006) (geographic data storage).

Figure 1. System architecture.



The bottom part of the figure shows the index construction workflow, which in turn consists of three modules: the document abstraction module, the index structure, and the index construction module.

The processing services are shown in the middle of the figure. The *Geographic Space Ontology Service* used in the spatial index construction is shown on the left side. This service is used extensively in the index construction module. On the right side, one can see the two services that are used to solve queries. The rightmost one is the *query evaluation service*, which receives queries and uses the index structure to solve them. The other service is a *Web Map Service* following the OGC specification (OGC, 2002) that is used to create cartographic representations of the query results. On top of these services a *Geographic Information Retrieval Module* is in charge of coordinating the task performed by each service to respond to the user requests.

The topmost layer of the architecture shows the two user interfaces that exist in the architecture: the *Administration User Interface* and the *Query User Interface*.

INDEX CONSTRUCTION WORKFLOW

Document Abstraction

Given that the system must be generic, it must support indexing several kinds of documents. These documents will be different not only because they may be stored using different file formats (plain text, XML, etc.), but also because their content schema may be different. The set of attributes that have to be stored in the index may be different in each document collection. For instance, a document collection may have a set of attributes (such as *document id*, *author*, and *document text*), whereas other document collection may have a different set (such as *document id*, *summary*, *text*, *author*, and *source*).

To solve this problem, we have defined an abstraction for documents similar to the one used in the Lucene text search engine (Apache, 2011). We have extended this idea adding the spatial indexing possibility. In our abstraction, a *document* consists of a set of *fields*, each one with a value that is extracted from the document text. Each field can either be *stored*, *indexed*, or both. If a field is stored, its contents are stored in the index structure and they can be retrieved by a query. If a field is indexed, then this field is used to build the index structure. Furthermore, a field can be indexed textually, spatially, or in both indexes.

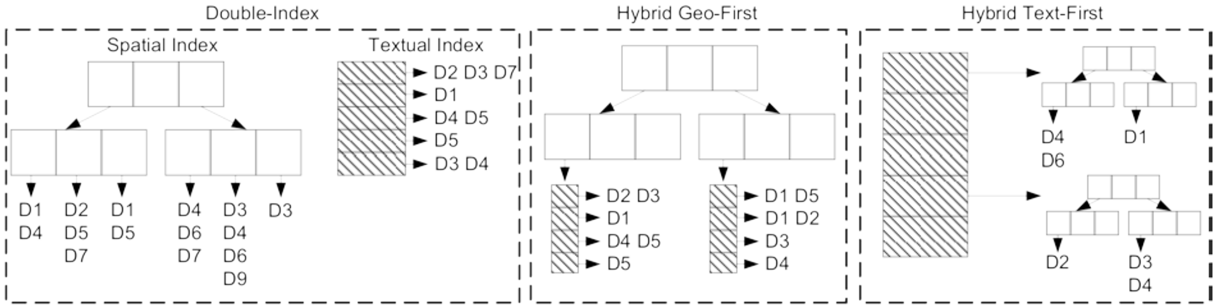
Indexing

Some work has been done to combine textual indexes and spatial indexes in structures able to solve the queries of interest in GIR systems. These structures can be broadly classified into *hybrid structures* (i.e., textual and spatial indexes are kept separate) and *double-index structures* (i.e., both indexes are merged in one single structure). The index structure proposed in the SPIRIT project (Vaid et al., 2005) is based on the combination of a grid (Nievergelt et al., 1981) and an inverted index. In Vaid et al. (2005), the authors conclude that keeping separate text and spatial indexes, instead of merging both in one, results in less storage costs but it could lead to higher response times. Keeping both indexes separated has many advantages (Martins et al., 2005). First of all, all textual queries can be efficiently processed by the textual index and all spatial queries can be efficiently processed by the spatial one. Moreover, queries combining textual and spatial aspects are supported. Updates in each index are handled independently, which makes easier the addition and removal of data. Finally, specific optimizations can be applied to each individual indexing structure.

In more recent works (Martins05 et al., 2005; Chen et al., 2006), the authors survey the work in the SPIRIT project and propose improvements to the system and the algorithms defined. In their work, two naive algorithms are proposed: *Text-First* and *Geo-First*. Both algorithms use the same strategy: one index is first used to filter the documents (textual 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 Text-First and textual index in Geo-First). These two naive algorithms provide a broad classification of GIR index structures. Note that it can also be used to classify hybrid structures. Figure 2 shows the three basic structures according to this classification. The left most structure belongs to the *double-index* class and it can be both *Text-First* and *Geo-First* depending on the algorithm used to solve the queries. The other two are *hybrid* structures. The first one belongs to the *Geo-First* class (the spatial index is always accessed first) and the second one belongs to the *Text-First* class (the textual index is always accessed first).

In Zhou et al. (2005), the use of an inverted index and an R-tree is proposed. Authors combine both structures in the three ways described above and they conclude that keeping both indexes separated is less efficient than combining them (a similar conclusion had been presented in Vaid et al. (2005)) and that the use of the R-tree outperforms the efficiency of the grid based structures.

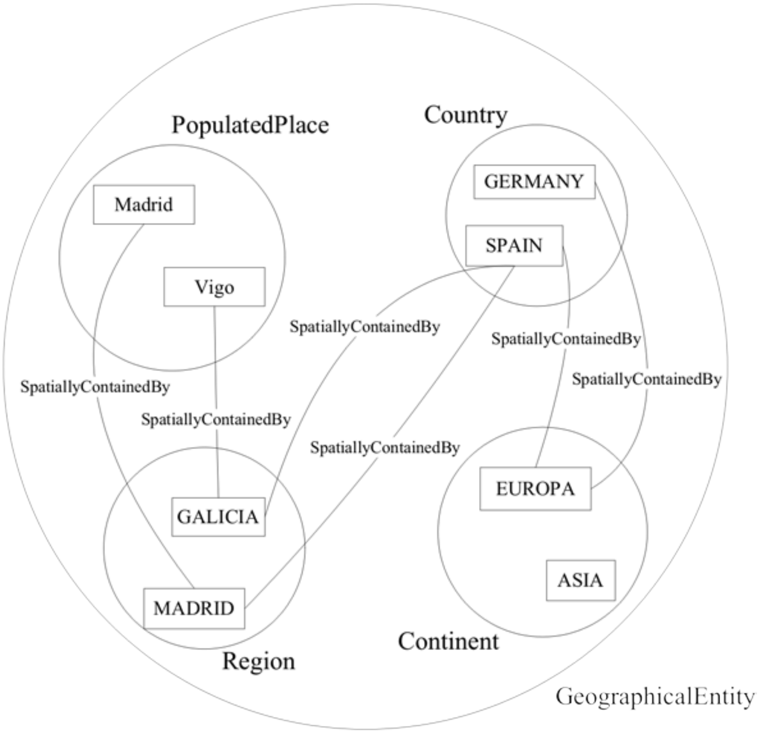
Figure 2. GIR index structures.



In Chen et al. (2006), authors propose the use of space-filling curves (Morton, 1966; Böhm et al., 1999) and compare the resulting structure both with the grid based and the R-tree based structures. Space-filling curves are based on the storage of the spatial objects according to the order determined by a filling curve. According to their experiments, the use of the space-filling curves outperforms both the grid and the R-tree based approaches.

Finally, in the STEWARD project (Lieberman et al., 2007), a double-index structure based on an inverted index and a Quad-tree (Nelson & Samet, 1986) is presented. In addition, the authors propose the use of a query scheduler in charge of choosing a *Text-First* or a *Geo-First* algorithm according to which index may return fewer results (this decision is based on statistics collected by the system).

Figure 3. Ontology instances.



Nevertheless, none of these approaches take into account the relationships between the geographic objects that they are indexing. A structure that can properly describe the specific characteristics of geographic space is an *ontology*, which is a formal explicit specification of a shared conceptualization (Gruber, 1993). An ontology provides a vocabulary of classes and relations to describe a given scope. In Brisaboa et al. (2010), we present an index structure based on an ontology of the geographic space that describes the concepts in our domain and the relationships that hold between them. Our spatial ontology is described in OWL-DL (W3C, 2011) and it can be downloaded from the following URL: http://lbd.udc.es/ontologies/spatialrelations. OWL classes can be interpreted as sets that contain individuals (also known as instances). Individuals can be considered *instances of classes*. Our ontology describes eight classes of interest: *SpatialThing*, *GeographicalThing*, *GeographicalRegion*, *GeopoliticalEntity*, *PopulatedPlace*, *Region*, *Country*, and *Continent*. In our ontology there are hierarchical relations among *SpatialThing*, *GeographicalThing*, *GeographicalRegion*, *GeopoliticalEntity* because:

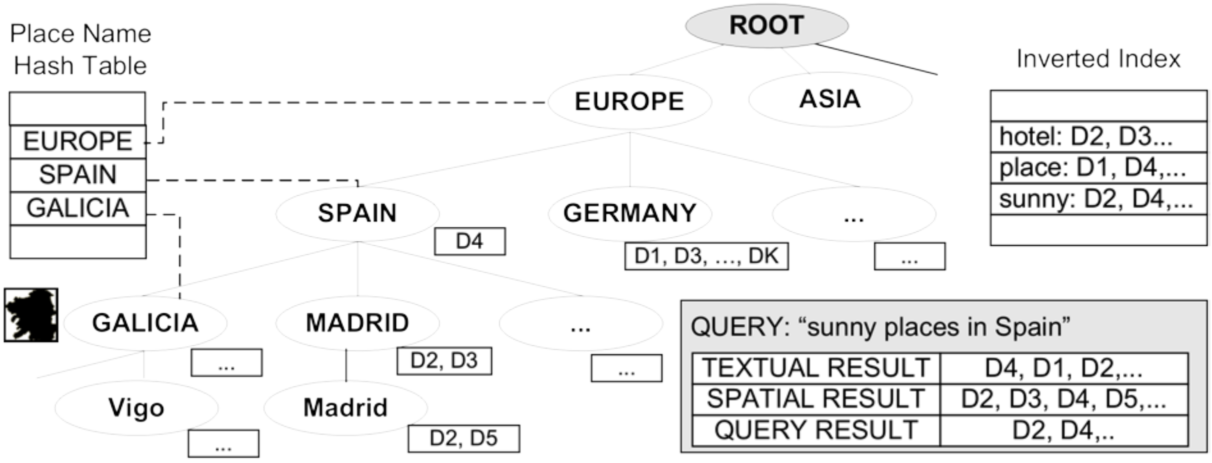
- *GeopoliticalEntity* is subclass of *GeographicalRegion*
- *GeographicalRegion* is subclass of *GeographicalThing* and
- *GeographicalThing* is subclass of *SpatialThing*.

That is, these four classes are organized into a superclass-subclass hierarchy, which is also known as *taxonomy*. Subclasses specialize (are subsumed by) their superclasses. *GeopoliticalEntity* has four subclasses: *PopulatedPlace*, *Country*, *Continent*, and *Region*. All the individuals are members of these subclasses. These four subclasses have an additional necessarily asserted condition regarding their relations with each other. They are connected by the property *spatiallyContainedBy* that describes the existence of a spatial relationship among them. For instance, all the individuals of class *PopulatedPlace* are *spatiallyContainedBy* individuals of class *Region* (described in OWL as *PopulatedPlace spatiallyContainedBy only (AllValuesFrom) Region*). Figure 3 shows an example of these relationships. Ontology classes are represented as circles, individuals as rectangles, and the relationships as labeled lines.

After having defined this ontology, we can define an spatial index structure based on it. This structure is a tree with four levels, one for each of the subclasses of *GeopoliticalEntity*. The top-most level contains a node for each of the instances of the class *Continent*. Each node in this level references the instances of the class *Country* that are connected by the *spatiallyContainedBy* relationship. The levels of *Region* and *PopulatedPlace* are built using the same strategy. That is, the structure of the tree follows the taxonomy of the ontology. Figure 4 shows the spatial index structure built from the instances shown in Figure 3.

The main advantage of this spatial index structure over other alternatives is that intermediate nodes in the structure have a meaning in the geographic space and they can have additional information associated. For instance, we can associate a list of documents that reference a given *Country* and use this list of documents to solve combined textual and spatial queries. Moreover, given that there is a superclass-subclass relationship between the levels, the bottom levels can inherit the properties of the top levels. Particularly, the documents associated to a node in the structure also refer to all nodes in its subtree. Furthermore, the index structure is general in the sense that the ontology of geographic space can be adapted to each particular application. For example, if a particular application uses a restricted area of the geographic space where the classes *Continent* and *Country* are not necessary and, on the other hand, the classes *Province*, *Municipality*, *City*, and *Suburb* are needed, we could define a different ontology of space and base the index structure on it as long as the relationship *spatiallyContainedBy* still holds between the classes. Finally, we could define additional spatial relationships in the ontology such as *spatiallyAdjacent* and maintain these relationships in the index structure to improve the query capabilities of the system.

Figure 4. Example of the index structure.



QUERY: "sunny places in Spain"

TEXTUAL RESULT	D4, D1, D2,...
SPATIAL RESULT	D2, D3, D4, D5,...
QUERY RESULT	D2, D4,..

This structure automatically performs query expansion over the geographic component of a query. In classical information retrieval systems, the use of ontologies to perform query expansion is well-known as an ontology can be regarded as a generalization of a thesaurus. However, our structure does not use an additional geographic ontology but it is defined using one. Thus, the query expansion process is intrinsic to the nature of the index and not an additional process. Furthermore, our whole structure combines this geographic ontology-based index with a textual index (an inverted index), and classical query expansion techniques can be used over the textual component achieving *full query expansion* (in the sense that both the textual and geographic components of a query are expanded). Considering again the example *Ph.D. dissertations in Spain*, using this full query expansion technique we could retrieve documents containing the terms *thesis* and *Madrid*.

Geo-referencing of Documents

This is probably the most complex stage of the workflow. In this stage, a geographic footprint is assigned to each document allowing its spatial indexing. A geographic footprint shows the geographic scope of the document and it can be set as a list of geographic coordinates, a *bounding box* grouping that coordinates, etc. For example, if the cities of London and Liverpool are cited in a document, the geographic coordinates of these cities or the minimum bounding box containing them can be used as geographic footprint of the document. This stage comprises two steps. First, the system analyses the document fields that are spatially indexable and extracts candidate location names from the text (i.e., *discovery of location names*). In a second step, these candidate locations are processed in order to determine whether the candidates are real location names, and, in this case, to compute their geographic locations (i.e., *translation of location names* to a geographic model).

The main problem that can happen at this point is the ambiguity of the geographic references. A recent research by Garbin & Mani (2005) claims that more than the 67% of the geographic references cited in texts are ambiguous. Furthermore, in Tjong et al. (2003) two kinds of ambiguity are presented. First, a location name can be ambiguous (*polysemy*). For instance, "*London*" is the capital of the United Kingdom and it is a city in Ontario, Canada too. Second, there can be multiple names for the same geographic location, such as "*Los Angeles*" and "*LA*". A third class of ambiguity could be considered. This is due to the use of the same word to refer both a place name and a organization, company, or person (e.g., Santiago).

Discovery of Location Names

Unlike geographic information systems, information in GIR systems is not structured. It is not possible to know *a priori* where geographic references are stored, nor their categories (e.g., city, state, country, etc.).

In this kind of systems, geographic references are contained in the text of the indexed documents. Therefore, these texts have to be analyzed in order to discover the geographic references.

In this analysis, all the spatially indexable fields are processed in order to discover the place names contained within. There are two *Linguistic Analysis* techniques that are widely used for this: *Part-Of-Speech* tagging (Brill, 1992) and *Named-Entity Recognition* (Chinchor & Robinson, 1997; Pustejovsky et al., 2005). On the one hand, Part-Of-Speech tagging is a process whereby tokens are sequentially tagged with syntactic labels, such as *verb* or *gerund*. On the other hand, Named-Entity Recognition is the process of finding mentions of predefined categories such as the names of persons, organizations, locations, etc. Combine both techniques is a good solution to discover possible place names contained in the text of documents. In our prototype, we use the *Natural Language Tool LingPipe* (Alias-i, 2011) to find locations. It is a suite of Java libraries for the linguistic analysis of human language free for research purposes that provides both Part-Of-Speech tagging and Named-Entity Recognition. LingPipe involves the supervised training of a statistical model to recognize entities. The training data must be labeled with all of the entities of interest and their types.

In spite of the good performance of these linguistic analysis techniques, when the discovered location names are translated to a geographic model many problems related with the ambiguity of the location names arise. Although these problems mainly affect the next step (i.e., *translation of location names*), some issues are related with this step. First, the system must determine if the discovered location names are true place names. Gazetteers have been widely used for this purpose. A Gazetteer is a geographical dictionary that contains, in addition to location names, alternative names, populations, location of places, and other information related to the location. Although some years ago the availability of these gazetteers was very scarce (Petasis et al., 2000), nowadays there are many resources that provide this kind of information. A more complex problem is the disambiguation of a location name once it has produced a true positive in the gazetteer. Many clues within the whole text of the documents are used by human beings to disambiguate each location name cited in the text. For example, if the location name *Santiago* is cited in a document near other location names, such as *A Coruña*, we will assume that the cited place is *Santiago de Compostela*. However, if it is cited near *Atacama*, we will assume that the cited place is *Santiago de Chile*. Perform this process in an automatic, and even semi-automatic, way becomes a challenge for the GIR community. Bruno Martins (2008) describes in his Ph.D. some basic principles to guide this automatic disambiguation process:

- *One referent per discourse*. A location name cited several times in the same text is likely to mean the same place. For example, if *Santiago* is cited several times in a document it should refer always either *Santiago de Compostela* or *Santiago de Chile*.
- *Related referents per discourse*. Geographic references appearing in the same document tend to refer to related locations. In our previous example, we use the place name *A Coruña* or *Atacama* to disambiguate the place name *Santiago*.
- *Default senses*. Important places are more likely to be referenced. Therefore, if no clues are available to disambiguate a location name, the most important place should be assigned. For example, countries are most important than cities, a capital is more important than each other city, a city is more important than a street, etc.

In the SPIRIT project (Jones et al., 2001; Jones et al., 2003; Jones et al., 2004; Fu et al., 2005), a spatial ontology is used instead of a gazetteer. Therefore, searches in the ontology both check if the location name is a true place and provide a disambiguation based on the height of the node in the ontology. This follows the default senses principle (the height of the regions is less than the height of the cities). Besides this project, the most important works in the area are: Web-a-where (Amitay et al., 2004), which uses *spatial containers* in order to identify locations in documents; MetaCarta (the commercial system described by Rauch et al. (2003)), which uses a natural language processing method; and STEWARD (Lieberman et al., 2007), which uses an hybrid approach.

Translation of Location Names

Once the location names have been located and disambiguated, they have to be translated to a geographic model, i.e., a geographic footprint has to be assigned to each document in order to make it indexable by a spatial index. There are some differences in the methods proposed in the bibliography. The first one is the type of geographic object used to represent those footprints. Some of the common options chosen for this purpose are: geographic points (the geographic coordinates of all the places cited in document), minimum bounding boxes (the boxes of minimum extent that cover the geographic positions of all the places cited in the document), and the centroid of such bounding boxes. A second feature that makes proposed methods different is the uniqueness of the footprint. Although documents are better described when several footprints are allowed, because many distant places can be cited in the same document (consider for example a document about the evolution of the world economy), many of the approaches in the bibliography suggest the use of a single footprint. The usage of one footprint simplifies the process and improves the performance of both the indexing and querying process.

One of the pioneering projects, previous to the SPIRIT project, aiming at the geo-referencing of contents in digital libraries is GIPSY (*Georeferenced Information Processing SYstem*) (Larson, 1995). In this project, each location name is translated to a geographic representation (for example, a polygon) and a weighting value is assigned to it. These values depend on features intrinsic to the content of the documents (e.g., frequency of the location name in the text of the document). Then, all these geographic representations are combined in three-dimensional topographic representations that consider the weighting values. Finally, a threshold determines the minimum elevation of the topographic representation that makes the area relevant.

In the SPIRIT project, each location name cited in the text of a document is translated to a bounding box and the footprint of a document consists of several bounding boxes (one for each location name cited in its text). This schema was also used by Zhou et al. (2005). In Smith & Crane (2001), authors propose the use of a set o points as the footprint of a document. Thus, the footprint of a document consists of the geographic coordinates of all the locations cited in its text. These coordinates are weighted by the frequency of the location name in the document. Then, the centroid of this set of points and its standard deviation are computed according to the weights of the points, and all the points that are more than twice from the centroid are prune (remaining points make up the footprint of the document).

Finally, as we mentioned in the previous section, in the Web-a-where (Amitay et al., 2004) project the disambiguation process is based on *spatial containers*. These containers are defined according to the topological relationships that exist in a gazetteer. Most of these relationships belong to the class *part-of* (for example, Galicia is *part-of* Spain). Once place names have been disambiguated, all the related places are merged in a taxonomy. The levels in this taxonomy are ordered according to their relevance and those levels which height is less than a threshold make up the footprint of the document.

In our prototype, we have developed a service based on an ontology of the geographic space that is built using a *Gazetteer* (Geonames, 2011) and a *Geometry Supplier* (NIMA, 2011). This service uses information available in the gazetteer (such as, place type, population, capital, etc.). All these data are combined to compute the intrinsic *importance* of each place.

QUERY EVALUATION SERVICE

The *Query Evaluation Service* is the component in charge of using the index structure to answer the queries posed by the users. Moreover, in order to return a useful result, this service must also provide a relevance ranking of the results. In this section, we describe the types of queries in GIR systems, the algorithms to solve them and the equations used to compute the relevance of the result for each of such types: *pure textual queries*, *pure spatial queries*, and *hybrid queries*.

Note that the concept of relevance of a document, although well-known in the field of IR, had not been introduced in the GIS before of the arisen of the GIR systems. In Godoy & Rodríguez (2004) some qualitative measures for the spatial relevance of a document are introduced based on concepts such as bounding boxes, distances, overlapping, and relative sizes. In Jones et al. (2001), an hybrid approach combining the distance in an ontology of the geographic space and the Euclidean distance in the geographic space is presented. Furthermore, in the context of the Tumba project (Martins et al., 2005; Andrade & Silva, 2006), authors use some well-known semantic concepts (e.g., adjacency, connectivity, etc.) to calculate this relevance when an ontology is available in the system. A different approach, which does not assume the use of an ontology, but based on similar concepts is presented in Zhou et al. (2005).

Pure Textual Queries

These are queries such as "*retrieve all documents where the words* hotel *and* sea *appear*". The textual index that is part of the index structure is used to solve them. In our prototype we use Lucene to implement this textual index, and thus, the relevance ranking depends on its scoring. Lucene scoring uses a combination of the vector space model and the boolean model of information retrieval (Baeza-Yates & Ribeiro-Neto, 1999). All scores are guaranteed to be 1.0 or less. More information about the Lucene scoring can be found in Gospodnetic & Hatcher (2005), and Apache Lucene (2011).

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. The spatial index that is part of the index structure is used to solve them. Given that a document in the result set of a query can include geographic references to one or more location names relevant to the query, the relevance of the document *d* with respect to the query *q* due to each location name *l* has to be computed. We denote this relevance as $toponymRelevance_{q,d,l}$. We guarantee that this value is 1.0 or less in order to make the integration of both spatial and textual relevancies easier. In van Kreveld et al. (2005) both spatial and textual relevancies are also normalized to values between 0 and 1. Finally, we compute the relevance of the document *d* with respect to the query *q* as the maximum relevance due to any location name (Equation 1).

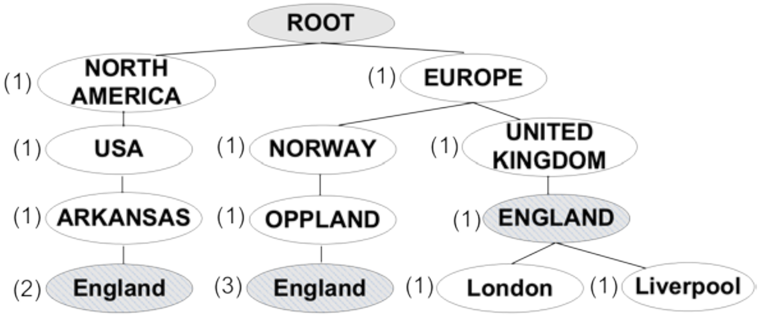
$$spatialRelevance_{q,d} = \max\{toponymRelevance_{q,d,l}\} \tag{1}$$

The computation of $toponymRelevance_{q,d,l}$ for queries specified selecting a node in the spatial index is a simplification of the previous one because in this case we have the certainty that the query refers to a specific node in the tree. Therefore, the documents associated to this node have relevance 1.0. The relevance of a document associated to the nodes in the subtree is computed using the previous equation. This is reflected in Equation 2.

$$toponymRelevance_{q,d,l} = \begin{cases} 1 & \text{if } l \text{ is specified in the query} \\ \dfrac{0.5^{distance}}{importance} & \text{otherwise} \end{cases} \tag{2}$$

The sketch of the index structure shown in Figure 5 is useful to understand the difference between both types of queries. Each node in the figure is annotated with its importance between parentheses. On the one hand, when the user specifies a query using the location name *England*, the relevance of a document due to *England* (an important city of Arkansas) will be higher than the relevance due to *England* (a small city of Oppland Fylke), and lower than the relevance due to *England* (a part of the United Kingdom). Concrete values of relevance are 0.5 for England in Arkansas, 0.33 for England in Oppland, and 1.0 for England in the United Kingdom. Moreover, the relevance of the document due to important cities of England (UK) like London or Liverpool is 0.5. This value is high enough to be taken into consideration. On the other hand, when the query is specified selecting the node for England in Arkansas the relevance of a document due to this node is 1.0 because the user explicitly indicates the interest about documents with geographic references to that location.

Figure 5. Queries specified using a location name vs. queries specified selecting a node.



Finally, in the case of queries specified using a query window the nodes are selected using the classical algorithm of spatial indexes. Therefore, the computation of *toponymRelevance*$_{q,d,l}$ must be performed using the distance ($dc_{q,l}$) and the overlap area ($oa_{q,l}$) between the query window and the location name. Equation 3 defines this computation. We use parameters w_{dc} and w_{oa} to weight the relevance of each factor and we use the importance of the location name to assign more relevance to the most important nodes that reference the location name.

$$toponymRelevance_{q,d,l} = \frac{w_{dc} \times dc_{q,l} + w_{oa} \times oa_{q,l}}{importance} \tag{3}$$

Equation 4 defines how to calculate the relevance due to the distance to the query window. *centerDistance*$_{q,l}$ represents the Euclidean distance between the location name *l* and the query window *q*. Similarly, *cornerDistance*$_{q}$ is a weight factor that represents the maximum distance to the center of the window.

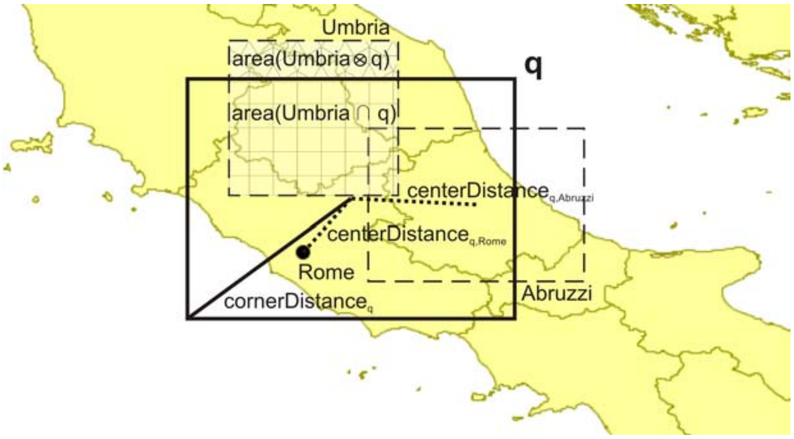
$$dc_{q,l} = 1 - \frac{centerDistance_{q,l}}{cornerDistance_{q}} \tag{4}$$

The relevance due to the overlap area with the query window is calculated according to Equation 5. When the geometry stored in the node is a point (leaf node), the overlap area is not significant. Thus, we use $1/[\text{area}(q)+1]^{0.15}$. This value depends only on the query window and is inversely proportional to its area. The concrete equation has been constructed based on the average area of the nodes in each level of the ontology of geographic space.

$$oa_{q,l} = \begin{cases} \dfrac{1}{\left[area(q)+1\right]^{0.15}} & \text{if } l \text{ is a point} \\ \max\left\{0, \dfrac{area(l \cap q)}{area(q)} - \dfrac{area(l \otimes q)}{area(l)}\right\} & \text{otherwise} \end{cases} \tag{5}$$

Figure 6 uses an example query window in central Italy to clarify the aforementioned equations. The bounding boxes of two regions, Umbria and Abruzzi, and a populated place, Rome, are shown in this figure. These bounding boxes as well as the query window *q* are used to compute the area of their respective entities (i.e., *area*(Umbria), *area*(Abruzzi), and *area*(q)). The region of Umbria is used to illustrate the relevance due to the overlap area (Equation 5). This relevance is computed using the area of the intersection of the region with the query and the area in the part of the region that does not intersect with the query. Moreover, three distances used to compute the relevance due the distance to the query window (Equation 4) are shown. The weight factor *corner distance* is depicted as a solid line, and the distances from Rome and Abruzzi to the center of the query window are depicted as dotted lines.

Figure 6. Queries specified using a query window.



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*". As in the previous case, the geographic area in the query can be a point, a query window, or a complex object. Both the textual and spatial indexes are used to solve them. Hence, we use the previous equations to compute the spatial and textual relevance. Equation 6 defines how we combine both relevance rankings. The weighted sum of the spatial and textual ranking values is one of the simplest methods and is commonly used (Martins et al., 2005; Zhou et al., 2005; Andrade & Silva, 2006). Furthermore, it is the base of more complex ranking methods (Yu & Cai, 2007). We assume w_t = 1-w_s and calculate w_t to normalize the differences between textual and spatial rankings.

$$relevance_{q,d} = w_t \times textualRelevance_{q,d} + w_s \times spatialRelevance_{q,d} \tag{6}$$

Textual Queries with Place Names

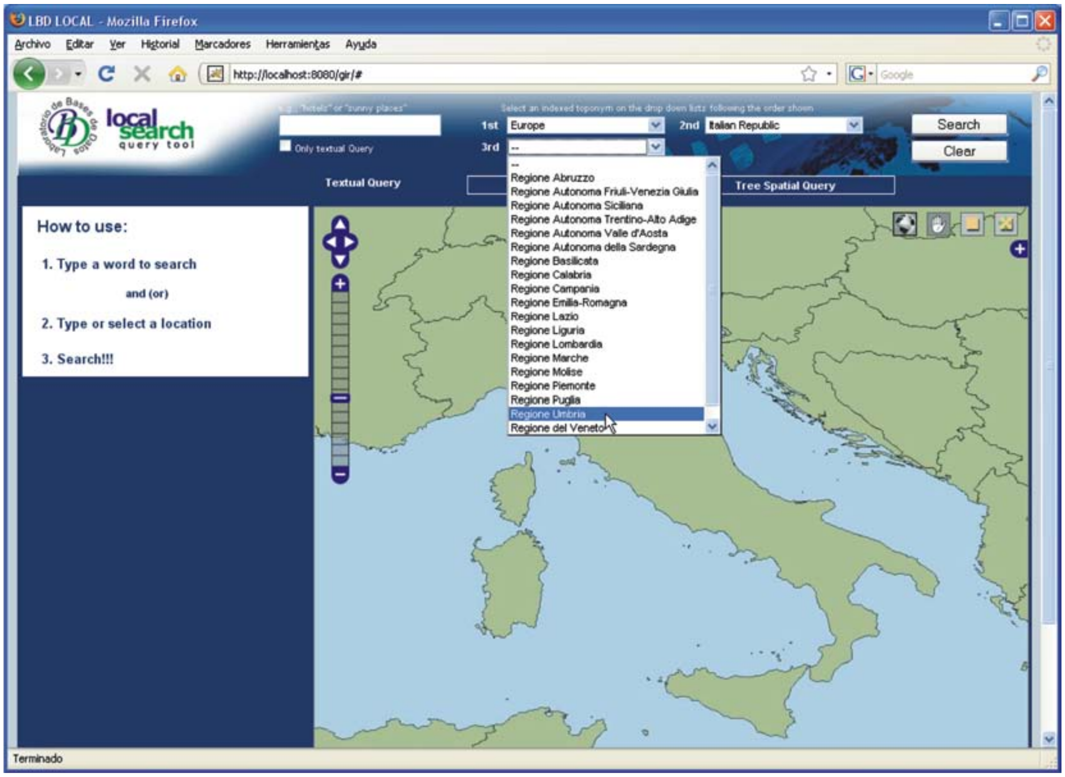
In this type of queries, some of the words are place names. For instance, "*retrieve all documents with the word* hotel *that refer to* Spain". Both the spatial and textual indexes are used to solve them and, as in the previous case, the weighted sum can be used to combine the relevancies obtained in each index.

Our index structure presents an improvement over the rest of proposals: it 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 then the internal node that represents the geographic object *Spain* will be quickly located. Then, all the documents associated to this node are part of the query result. Moreover, 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 documents 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). This geographic query expansion is complementary to other classical techniques expanding the textual component of the query.

USER INTERFACES

The system has two different user interfaces: an administration user interface and a query user interface. The administration user interface can be used to manage the document collection. The main functionalities are: creation of indexes, addition of documents to indexes, loading and storing indexes, etc. Figure 7 shows a screen-shot of the query user interface. This interface was developed as a web application using the Open Layers API (OSGeo, 2011). This API provides a number of utilities for manipulating maps and adding content to the map.

Figure 7. Query User Interface.



LBD LOCAL - Mozilla Firefox

http://localhost:8080/gir/#

e.g. "hotels" or "sunny places"

Select an indexed toponym on the drop down lists following the order shown

1st: Europe

2nd: Italian Republic

3rd: --

□ Only textual Query

How to use:

1. Type a word to search

and (or)

2. Type or select a location

3. Search!!!

In the previous section, we have presented the types of queries that can be solved with this system. These queries have two different aspects: a textual aspect and a spatial aspect. In our prototype, the query user interface allows the user to indicate both aspects. The spatial context can be introduced in three ways that are mutually exclusive:

- *Typing the location name*. In this case, the user types the location name in a text box. This is the most inefficient way because the system has to obtain all the geographic references associated with the place name typed by the user, which is a time-expensive process.
- *Selecting the location name in a tree*. In this case, the user sequentially selects a continent, a country within this continent, a region within the country, and a populated place within the region. If the user wants to specify a location name of a higher level than a populated place, it is not necessary to fill in all the levels. The operation is very easy and intuitive because the interface is implemented with a custom-developed component using the AJAX technology that retrieves in the background the location names for the next level. When the user selects a place in the component, the map on the right zooms in automatically to the selected place.
- *Selecting the spatial context of interest in the map*. The user can navigate using the map on the right to visualize the spatial context of interest. After that, a rectangle can be drawn over it. The system will use this rectangle as the query window if the user did not type a place name or did not select a location name.

CONCLUSION

Nowadays, Geographic Information Systems constitute a consolidated area in computer science. Many impressive research results have been presented and, more importantly, an effective technology transfer has improved the management of geographic information in traditional information systems. For example Spatial Data Infrastructures are prominent examples of this technology transfer.

The work presented in this paper can also be easily integrated in a spatial data infrastructure. First, the query processing functionality can be implemented as a Web Processing Service (OGC, 2007), which aims at the standardization of the way that GIS calculations are made available on the Internet. This service can then be used to index document collections such as administrative archives.

Furthermore, some of the internal components of the architecture could also be implemented using Web Processing Services. As an example, in Ladra et al. (2008) we show how the Geographic Space Ontology Service, which is enclosed in the intermediate layer of the architecture, can be integrated in a Spatial Data Infrastructure using the OGC Web Processing Service. Moreover, even though our prototype uses a database as Gazetteer Service, other implementations could easily use a Gazetteer from a spatial data infrastructure.

Finally, the GIR architecture proposed in this chapter is a perfect complement to OGC catalogues. A GIR system like the one described in this chapter can be built over a collection of OGC catalogues. In this case, instead of indexing digital documents, the system would index metadata records which are composed of textual descriptions and geographic references of the datasets and services. The user interface of the GIR system would allow a user to query the catalogues with keywords and a geographic reference and it would return a list of metadata records ranked by their relevance. An additional advantage of a GIR system is that an OGC catalogue is built by a human being that decides what is relevant and that categorizes the documents using thesaurus. Furthermore, an OGC catalogue is oriented to structured searches applying filters to metadata fields. On the other hand, a GIR system categorizes the documents automatically deciding what is relevant using the contents and allowing non-structured searches.

The application of this research to improve the task of information retrieval turns out to be a rather challenging problem. Due the importance of this task (millions of users perform queries on-line each day), many research efforts have been devoted to this new research topic and a new research area, named Geographic Information Retrieval, has emerged covering the topics in the intersection between GIS and IR. In this chapter, we have presented the state-of-the-art in this new field. Our own architecture (Brisaboa et al., 2010) was used as the framework that embraces the majority of the topics in the area.

Many new research topics have emerged in this young area. First, Geo-referencing techniques must be improved to solve the ambiguity problems. This is a crucial task as its influence in the precision and recall of the system is extremely high. Second, the development of new index structures is also very important as they are a key factor in the performance of the GIR systems. The use of ontologies in these structures, enhancing their semantic, represents an exciting problem that was just sketched with some initial proposals. As we shown in Brisaboa et al. (2010), structures considering the semantic of the space present valuable improvements for GIR systems over classical spatial index structures (e.g., query expansion, relevance ranking, etc.). Efficient implementations of these structures will also be crucial in their applicability. Finally, some efforts must be devoted to improve the system usability. The spatial component of GIR systems entails the need of user interfaces that allow users to properly express the spatial scope and represent the results in a user-friendly way.

REFERENCES

Alias-i (2011). LingPipe, Natural Language Tool. Retrieved June 9, 2011 from http://www.alias-i.com/lingpipe

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. In *ACM Special Interest Group on Information Retrieval Conference, SIGIR'04* (pp. 273-280). ACM Press.

Andrade, L., & Silva, M.J. (2006). Relevance Ranking for Geographic IR. In *Workshop on Geographic Information Retrieval, GIR'06*. ACM Press.

Apache (2011). The Apache Lucene project. Retrieved June 9, 2011 from http://lucene.apache.org

Apache Lucene (2011). Apache Lucene - Scoring. Retrieved June 9, 2011 from http://lucene.apache.org/java/2_2_0/scoring.html

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

Böhm , C., Klump, G., & Kriegel, H.-P. (1999). XZ-Ordering: A Space-Filling Curve for Objects with Spatial Extensión. In *Advances in Spatial Databases Conference, SSD'*99 (pp. 75-90). Springer.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Conference on Applied Natural Language Processing, ANLP'92* (pp. 152-155). Association for Computational Linguistics.

Brisaboa, N.R., Luaces, M.R., Places, A.S., & Seco, D. (2010). Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *GeoInformatica 14* (3), 307-331.

Chen, Y.-Y., Suel, T., & Markowetz, A. (2006). Efficient query processing in geographic web search engines. In *ACM Special Interest Group on Management of Data, SIGMOD'06* (pp. 277-288). ACM Press.

Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Message Understanding Conference*.

European Commission (2011). INSPIRE Directive. Retrieved June 9, 2011 from http://www.ec-gis.org/inspire

Fu, G., Jones, C.B., & Abdelmoty, A.I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems, ODBASE'05* (pp. 1466-1482). Springer.

Garbin, E., & Mani, I. (2005). Disambiguating toponyms in news. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT-EMNLP'05* (pp. 363-370). The Association for Computational Linguistics.

Geonames (2011). GeoNames Gazetteer. Retrieved June 9, 2011 from http://www.geonames.org

Godoy, F., & Rodríguez, A. (2004). Defining and comparing content measures of topological relations. *GeoInformatica 8* (4), 347-371.

Gospodnetic, O, & Hatcher, E. (2005). *Lucene IN ACTION*. Manning.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition 5* (2), 199-220.

GSDI (2011). Global Spatial Data Infrastructure Association, Online documentation. Retrieved June 9, 2011 from http://www.gsdi.org

ISO/IEC (2002). *Geographic Information - Reference Model*. International Standard, ISO 19101.

Jones, C.B., Alani, H., & Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In *International Conference on Spatial Information Theory, COSIT'01* (pp. 322-335). Springer.

Jones, C.B., Abdelmoty, A.I., & Fu, G. (2003). Maintaining ontologies for geographical information retrieval on the web. In *On The Move to Meaningful Internet Systems, ODBASE'03* (pp. 934-951). Springer.

Jones, C.B., Abdelmoty, A.I., Fu, G., & Vaid, S. (2004). The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *International Conference on Geographic Information Science* (pp. 125-139). Springer.

Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M. J., & Weibel R. (2002). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. In *ACM Special Interest Group on Information Retrieval Conference, SIGIR'* 02 (pp. 387-388). ACM Press.

Ladra, S., Luaces, M.R., Pedreira, O., Seco, D. (2008). A Toponym Resolution Service Following the OGC WPS Standard. In *International Sympoium on Web and Wireless Geographic Information System, W2GIS'08* (pp. 75-85). Springer.

Larson, R.R. (1995). Geographic information retrieval and spatial browsing. *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, 81-123.

Lieberman, M.D., Samet, H., Sankaranarayanan, J., & Sperling, J. (2007). STEWARD: Architecture of a Spatio-Textual Search Engine. In *ACM International. Symposium on Advances in GIS, ACMGIS'07 (pp. 186-193)*. ACM Press.

Martins, B., Silva, M.J., & Andrade, L. (2005). Indexing and ranking in Geo-IR systems. In *Workshop on Geographic Information Retrieval, GIR'05* (pp. 31-34). ACM Press.

Martins, B. (2008). *Geographically Aware Web Text Mining*. Unpublished doctoral dissertation. Universidade de Lisboa, Portugal.

Morton, G. M. (1966). *A computer Oriented Geodetic Data Base and a New Technique in File Sequencing*. Technical Report, IBM Ltd.

NIMA (2011). National Imagery and Mapping Agency, Vector Map Level 0. Retrieved June 9, 2011 from http://www.mapability.com

Nelson, R.C., & Samet, H. (1986). A consistent hierarchical representation for vector data. In *Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'86* (pp. 197-206). ACM Press.

Nievergelt, J., Hinterberger, H., & Sevcik, K. C. (1981). The Grid File: An Adaptable, Symmetric Multi-Key File Structure. In *European Cooperation in Informatics, ECI'81* (pp. 236-251). Springer.

OGC (2002). *OpenGIS Web Map Service Implementation Specification*. OpenGIS Project Document 01-068r3. Open GIS Consortium, Inc.

OGC (2003). *OpenGIS Reference Model*. OpenGIS Project Document 03-040. Open GIS Consortium, Inc.

OGC (2006). *OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 2: SQL option*. OpenGIS Project Document 06-104r3. Open GIS Consortium, Inc.

OGC (2007). *OpenGIS Web Processing Service Implementation Specification*. OpenGIS Project Document 05-007r7. Open GIS Consortium, Inc.

OSGeo (2011). Open Layers API. Retrieved June 9, 2011 from http://openlayers.org

Petasis, G., Cucchiarelli, A., Velardi, P., Paliouras, G., Karkaletsis, V., & Spyropoulos, C.D. (2000). Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *ACM Special Interest Group on Information Retrieval Conference, SIGIR'* 00 (pp. 128-135). ACM Press.

Pustejovsky, J. , Knippen, R., Littman, J., & Saurí, R. (2005). Temporal and event information in natural language text. *Computers and the Humanities 39*, 123-164.

Rauch, E., Bukatin, M., & Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Human Language Technology Conference, HLT-NAACL'03* (pp. 50-54). Association for Computational Linguistics.

Salton, G. (1963). Associative Document Retrieval Techniques Using Bibliographic Information. *Journal of the ACM 10* (4), 440-457.

Smith, D.A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *European Conference on Research and Advanced Technology for Digital Libraries, ECDL'01* (pp. 127-136). Springer.

Tjong, E.F., Sang, K., & Meulder, F.D. (2003). Introduction to the CoNLL-03 shared task: Language-independent named entity recognition. In *Conference on Natural Language Learning, CoNLL'03* (pp. 142-147).

Vaid, S., Jones, C. B., Joho, H., & Sanderson, M. (2005). Spatio-Textual Indexing for Geographical Search on the Web. In *Symposium on Spatial and Temporal Databases, SSTD'05* (pp. 218-235). Springer.

van Kreveld, M., Reinbacher, I., Arampatzis, A., & van Zwol, R. (2005). Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *GeoInformatica 9* (1), 61-84.

Worboys, M. F. (2004). *GIS: A Computing Perspectiva*. CRC.

W3C (2011). World Wide Consortium - OWL 2 Web Ontology Language Document Overview. Retrieved June 9, 2011 from http://www.w3.org/TR/owl2-overview

Yu, B., & Cai, G. (2007). A Query-Aware document Ranking Method for Geographic Information Retrieval. In *Workshop on Geographic Information Retrieval, GIR'07*. ACM Press.

Zhou, Y., Xie, X., Wang, C., Gong, Y., & Ma, W. Y. (2005). Hybrid index structures for location-based web search. In *ACM International Conference on Information and Knowledge Management, CIKM'05* (pp. 155-162). ACM Press.

KEY TERMS & DEFINITIONS

Information Retrieval (IR): research area related with the access to non-structured repositories of information.

Geographic Information Retrieval (GIR): research area related with the properly management of the geographic information available in IR repositories.

Architecture: formal description of the set of services and structures that compose a system.

Geo-reference: geographic references contained in repositories of information in textual form (e.g., place names, postal codes, etc.)

Geo-referencing process: common task in GIR involving the location of geo-references and its translation to a formal model of the geographic space.

Workflow: sequence of tasks that have to be performed to achieve a goal.

Relevance: measure of the importance of a resource regarding to a specific query.