

# Indexing and Ranking in Geo-IR Systems

Bruno Martins and Mário J. Silva and Leonardo Andrade  
Faculdade de Ciências Universidade de Lisboa  
1749-016 Lisboa, Portugal  
{bmartins,mjs,leonardo}@xldb.di.fc.ul.pt

## ABSTRACT

This paper addresses document indexing and retrieval using geographical location. It discusses possible indexing structures and result ranking algorithms, surveying known approaches and showing how they can be combined to build an effective Geo-IR system.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

## General Terms

Algorithms, Design

## Keywords

Geo-IR, Indexing, Ranking, Searching

## 1. INTRODUCTION

Finding resources related to a specific geographic region is very difficult with IR systems based on keyword matches. Geo-IR addresses queries of the type `concept@location` (i.e. documents relevant with respect to some thematic concept and some geographic region) using, for instance, the geographical scopes (i.e. the region of interest) computed for each document, with basis on the geographical references made in the text [13, 14]. Relevance has now two different dimensions (thematic and geographical), raising the problems of defining geographical relevance, finding appropriate metrics for its computation, building efficient index structures for this information, and evaluating their performance.

Unlike the spatial information used in Geographic Information Systems (GIS), geographic information obtained from Web documents is often incomplete and fuzzy. Typical GIS queries specify complex spatial restrictions, while a Web search engine targets a wide variety of users who provide simpler queries. However, these users are also in need of effective retrieval mechanisms for queries with geo-spatial relationships. One of the main challenges lies on associating text indexes with geo-spatial indexes, providing the support for finding resources (i.e. local businesses, services or

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIR'05, November 4, 2005, Bremen, Germany.

Copyright 2005 ACM 1-59593-165-1/05/0011 ...\$5.00.

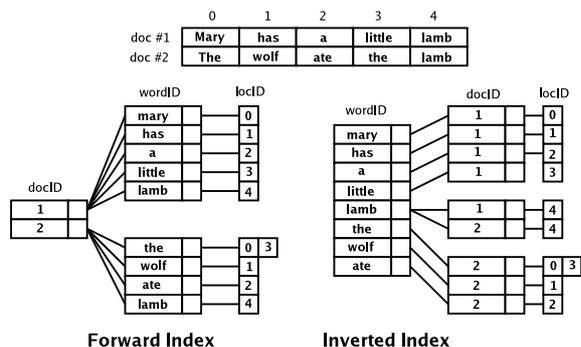


Figure 1: Indexes for text retrieval.

events) related to a particular location. This paper discusses possible indexing structures and ranking algorithms for Geo-IR on Web data, building on existing work from the areas of IR and GIS.

## 2. TEXT RETRIEVAL

Information retrieval (e.g. Web search engines) concerns essentially with two main activities: indexing and searching.

Indexing refers to representing data for the purpose of efficient retrieval, and is done after pre-processing operations have taken care of extracting appropriate items (i.e. tokenizing text). Various text indexing methods have been developed. Inverted indexes are the most popular technique [27], consisting of a set of inverted lists, one for each occurring word or index term. The inverted list for a term is a sorted list of positions, or hits, where the term appears in the collection. A hit consists of a document identifier and the position of the term within it, often containing additional information useful for ranking (e.g. HTML markup). Figure 1 shows a forward index (usually created as a first step in making an inverted index) and an inverted index for two example documents.

Searching involves the use of the structure built in the indexing stage for processing queries. A typical query contains terms and operators (i.e. disjunction, conjunction and filters). The indexes are examined to find matching documents, and a similarity score is computed between the query and each document. A ranked list is finally computed according to the similarity scores. The term weighting and document ranking function known as Okapi BM25 is the state-of-the-art in ranking results for text IR, and extensions to HTML documents have also been proposed [19].

In Web IR, citations and hypertext links are commonly combined with document content to improve ranked retrieval. PageRank is the most popular link-based ranking algorithm [17], and researchers have evaluated different techniques for combining standard text-based techniques with link-based ranking scores [8].

### 3. RETRIEVAL WITH GEO-SCOPES

In addition to having geo-scopes associated with the documents [14], and similarly to text IR, retrieving documents with basis on geographical criteria requires appropriate indexing and searching mechanisms. In geographic space, “everything is related to everything else, but near things are more related than distant things” [23]. We can hypothesize that the relevance of a location with respect to a query region increases with decreasing Euclidean distance between them [20]. The extent of overlap can also be used to measure spatial relevance. For instance, the greater the overlap between the two regions, the greater the assumed relevance [1, 7, 26]. Besides spatial distance, we can define notions of topological distance between locations. Examples include adjacency, connectivity or hierarchical containment. Hierarchical measures can, for instance, use the number of non-common parents between a pair of places within the hierarchies to which they belong [22], or the minimum number of direct relationships separating both places at an ontology [11]. Besides edge-counting, semantic similarity measures can also take into consideration hierarchy depth, or even things like language, population, and non-geographical relations. The problem of measuring similarity in hierarchical semantic structures has in fact been extensively studied [11]. Combinations of semantic and spatial methods can also be used to create hybrid metrics [6, 9], which in turn can be further combined with thematic similarity to create an integrated Geo-IR relevance ranking metric [9, 12]. A good motivation for using semantic similarity is that Euclidean space has been noted as unsuitable for modeling geographical proximity [28]. The concept of proximity is asymmetric, as people can consider  $A$  is near  $B$  while considering  $B$  is not near  $A$ . This asymmetry is related to the sizes and importance of geographical objects (e.g. total population or economic relevance), and the existing relationships with other geographical objects.

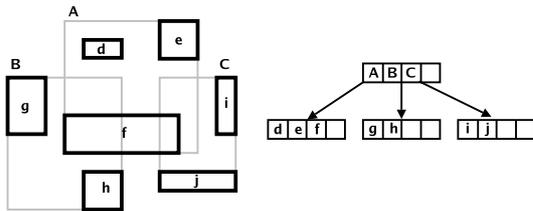


Figure 2: Rectangles arranged in a R-tree hierarchy.

Different multi-dimensional indexes have been proposed for managing spatial data, including grid indexes, quad-trees, R-trees, k-d-trees, and space filling curves such as Z-order [5]. Since geo-scopes can be seen as spatial footprints, these schemes can be used for document retrieval in a Geo-IR system. Still, most of the proposed approaches are tuned for handling complex spatial objects, e.g. polygons with hundreds of vertices [21], whereas we are interested in supporting relatively simple but frequent queries.

The most popular spatial indexing method is the R-Tree [5], a balanced tree derived from the B-tree which splits space in hierarchically nested, possibly overlapping, boxes – See Figure 2. Each internal node has between  $m$  and  $M$  children for some constants  $m$  and  $M$ . 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). To keep lookups fast, the R-tree insertion algorithm attempts to minimize the overlap and total area of nodes using var-

ious heuristics (for example, inserting a new rectangle in the subtree that would increase its overlap with its siblings by the least amount). One set of heuristics, called the R\*-tree [2], has been empirically validated as reasonably efficient for random collections of rectangles. Two R-trees  $T1$  and  $T2$  can be intersected by traversing the trees in tandem, comparing the current  $T1$  node with the current  $T2$  node and expanding the nodes only if their bounding boxes overlap. Traversing the trees in tandem has the potential for pruning much of the search, since if two nodes high in each tree are found to be disjoint, the rectangles stored in their subtrees will never be compared. An R-Tree efficiently supports operations such as enclosure (return all rectangles that contain the query rectangle or point), intersection (return all rectangles that intersect with the query rectangle), nearest neighbor (returns the rectangle or point nearest to the query point or rectangle) and closest pairs (returns the pairs of rectangles or points that have smallest Euclidean distance between them). These operations form the basis of many interesting Geo-IR access methods. Through this index structure (or at least using the index as filter to prune non-qualifying objects) we can build a system that ranks results according to distance, area of overlap, or combinations of these two measures. Spatial indexing methods like the R-Tree have little difficulty in computing the precise answer to a rectangular range query, but a radial query or a query involving non-rectangular polygons may require processing in two steps: first, filter out spatial objects that based on their minimum bounding box cannot possibly satisfy the constraint, and then compare the precise extent of the objects with the constraint. This aspect has not been taken into account in most published performance analysis, and designing retrieval functionalities making use of complex polygonal shapes can incur in performance problems. Nonetheless, previous studies found evidence that ranking retrieval sets by spatial similarity to a reference region is relatively insensitive to differences in the exact shape and size of the footprints [7].

Rees described c-squares, a promising a hierarchical spatial indexing system [18]. Rather than using bounding boxes or polygons, c-squares is built on the principle that the surface of the earth can be divided up to a grid of labeled squares, at one of a range of scales, and spatial data can be represented as a list of those squares, encoded as a textual string comprising numbers and a separator character. Spatial data can be matched to a designated search region (itself expressed as one or more c-square codes) using standard text search methods. Because the codes are hierarchical, a user’s search will be capable of matching codes at multiple resolutions. Also being standard textual strings, c-squares codes can be easily exchanged, and we can leverage on standard text searching approaches to build complex spatial access methods that deal with irregular polygons more effectively than bounding rectangles.

Besides spatial approaches, a Geo-IR system can also be built on top of simpler indexing schemes, similar to the inverted and forward indexes used for text retrieval. The idea is to associate each geo-scope with the list of documents concerning it, and associating each document to the correspondent geo-scope (or list of geo-scopes) – See Figure 3. The lists associated with each index key (a docID in the forward index and a scopeID in the inverted index) can be pre-ordered according to a relevance score (i.e. how related is a document to a geo-scope). This indexing scheme allows a very fast search for all the documents related to some specific geographic scope(s), or a search for the geographic scope(s) related to a given document(s), providing the necessary support for the development of interesting Geo-IR functionalities.

The trick in using these simple indexing schemes lays in transforming the location part of the query, together with the spatial operators (i.e. in, near or north-of), into a set of geo-scopes, as-

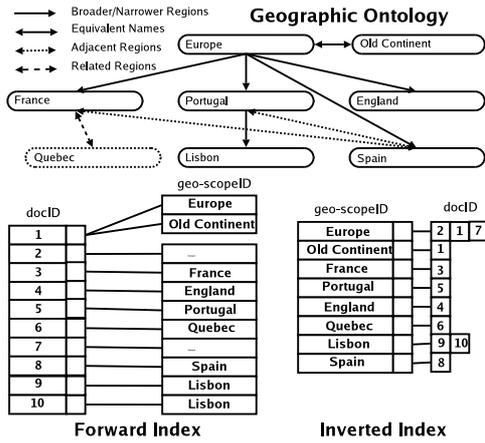


Figure 3: Simple indexes linking documents with geo-scopes.

signing each scope with a relevance score for ranking (i.e. scopes more related to the query are more relevant). However, for some queries, this can lead to large sets of geo-scopes. A similar problem has been reported for c-square strings, which can be very long for broad areas. There is a mechanism for “compressing” c-square codes, using “wildcards” to follow a cell at any level of the hierarchy. Similarly, different geo-scopes for narrow regions can be aggregated into a single broader geo-scope, provided that the indexes keep each document associated with multiple scopes (with relevance scores precomputed for each) corresponding to different hierarchical levels. A geo-retrieval algorithm can follow this general guideline:

1. Transform the location and the spatial operators in the query into a geo-scope or more, if the query cannot be disambiguated into a single geo-scope.
2. Rank each geo-scope in the set according to how relevant they are to the query location.
3. Get the ranked list of documents matching the set of geo-scopes. Ranking is based on the relevance of the documents to their corresponding scopes, obtained from the (pre-ordered) index, combined with the relevance score assigned to each scope from the query (e.g. a linear combination).

The translation of the query into a list of geo-scopes can be made through a geographical ontology, containing the relations among geographic concepts [3]. For instance “contained in Portugal” could be transformed into a set of scopes concerning the country and its administrative subdivisions. If the ontology also associates a spatial footprint to each geographical concept, then the ranking of geo-scopes can be made through spatial criteria, such as distance or overlap. Access to the information at the ontology must however be done efficiently, and we therefore require appropriate data structures for storing the information (e.g. sparse matrices for storing connectivity and spatial indexes for storing the footprints associated with the concepts at the ontology).

#### 4. COMBINING BOTH APPROACHES

In a previous study, Vaid et al. compared different ways to combine text indexes with geographical indexes, namely associating spatially ordered lists of documents to terms in a text index (text primary TS), associating text indexes to a spatial index (space primary ST), and keeping separate text and spatial indexes (T+S) [24].

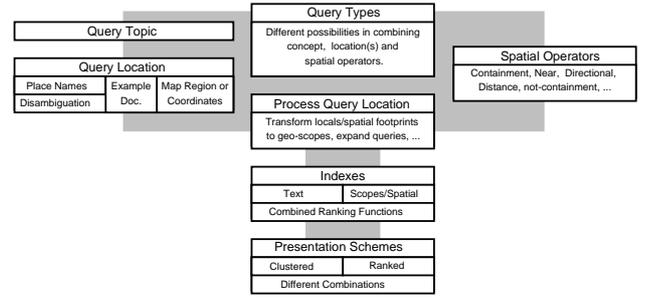


Figure 4: Queries, operators and result presentation schemes.

The authors concluded that the T+S scheme resulted in less storage costs, although it could lead to higher response times.

We plan on keeping separate indexes for text and geographical scopes. This division has many advantages: (i) all conventional (text) queries, which do not require the geographical index, can be efficiently processed by the inverted index, (ii) geographical “query-by-example” requests, searching for resources related to a given example (i.e. nearby), can be efficiently processed by the geographical index alone, (iii) queries combining thematic and geographical aspects are supported, (iv) updates in each index are handled independently, (v) new/existing data can be easily added to/removed from the system (i.e. introduction of a new geographical scope for a document without changing the text index), (vi) specific optimizations can be applied to each individual indexing structure, and (vii) different result ranking approaches are supported (i.e. geographical, thematic and combinations).

In what concerns geographical indexing, we plan on experimenting with both the spatial methods and the scope indexes presented on the previous section. The inverted index for text can be used to find documents matching the query terms and rank them according to thematic similarity. Separately, a geographical index can be used to find documents matching the query location (accounting for the spatial operators in the query), and results can be ranked according to geographical similarity. These two ordered sets can then be intersected to find those documents that both contain the query terms and have a geo-scope that matches the query footprint. Results can also be presented in several different ways, namely ranked according to thematic similarity, (hierarchically) clustered according to their geo-scopes, ranked according to geographical similarity, or ranked according to a combination of thematic and geographical similarity. The combination of thematic (tsim) with geographical similarity (gsim) can follow different schemes:

- Various linear combinations  $s = w1 * tsim + (1 - w1) * gsim$  for  $0 < w1 \leq 1$
- The product  $s = tsim * gsim$
- The maximum similarity  $s = \max(tsim, gsim)$
- The step-linear function  $s = tsim * H(gsim)$  where  $H(gsim) = 1$  for  $gsim > threshold$  and 0 otherwise. This is the same ranking according to thematic similarity, but only geographically relevant results are considered.

Figure 4 shows different combinations that can be made in terms of queries with geographical context, indexing schemes, and result presentation schemes. The “thematic” part of the query can be specified both through query terms and an example document. The “location” part can be specified by a location name, an example document (which is indexed according to a corresponding geo-scope), or a spatial footprint (i.e. a region selected from a map or a

pair of coordinates). Different operators for geographical concepts can also be used (i.e. find resources related to a location, find near resources, find resources within a given distance, or find resources at a given direction). The internals of the Geo-IR system can involve multiple data combinations and transformations (i.e. transforming location names to geographic scopes, transformations between spatial footprints and geographic scopes, and combining different ranking scores). Finally, results can be presented as a ranked list (together with information related to their geographical scopes) or clustered according to geographical scopes.

## 5. EVALUATION METHODOLOGY

Text indexing and spatial access methods have both been active areas of research for over two decades, and comparisons of different schemes have been described in the literature [5, 10, 27]. However, with very few exceptions [24, 25], reports on experiments combining these indexing schemes are very scarce. The whole subject of Geo-IR is still at an early stage of development, and very few studies have so far been performed on such systems. Geo-IR evaluation problems are discussed in a separate publication [16].

We are currently implementing some of the ideas discussed here, building on an existing Web search engine (www.tumba.pt), on a document indexing system specifically developed for the search engine [4], and on a software framework for assigning documents to their corresponding geographical scope [13, 14, 15]. Some of the proposed concepts were already tested at GeoCLEF, a Geo-IR track at the 2005 edition of the CLEF IR evaluation campaign. However, official results are still not available. We also plan on testing further developments using the GeoCLEF datasets.

Besides measuring relevance (through popular metrics such as precision and recall) according to a gold-standard collection of documents and queries, and using different combinations of scope assignment, indexing and ranking algorithms, future evaluation studies will also address computational aspects related to indexing and searching information with geographical context. Some of these particular aspects include:

- Response times for spatial indexes, scope indexes, and their combination with textual indexes, using different software architectures.
- Storage cost for spatial indexes and scope indexes.
- Response times involved in the use of the geographical ontology to disambiguate and expand location queries.
- Computational costs associated with having each document associated with several different geographic scopes (e.g. document fragments can have a specific geographical context) or having the same scope assigned to whole Web sites.

## 6. CONCLUSIONS

Geographic information retrieval builds on two well-established areas of information technology, namely information retrieval and geographical information systems. This paper presented methods developed in both areas for searching and indexing information, showing how they can be combined to address the problem of retrieval with geographic context. We are currently experimenting with some of the ideas reported here. Future studies will test if the combination of textual with geographical indexes is computationally feasible, and if it generates superior results for queries with a geographical context.

## 7. REFERENCES

- [1] K. Beard and V. Sharma. Multidimensional ranking in digital spatial libraries. *Special Issue of Meta-data - Journal of Digital Libraries*, 1(1):153–160, 1997.
- [2] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The  $R^*$ -tree: An efficient and robust access method for points and rectangles. In *Proceedings of SIGMOD-90, the 1990 Conference on Management of Data*, 1990.
- [3] M. Chaves, M. Silva, and B. Martins. A geographic knowledge base for semantic web applications. In *Proceedings of SBBD-05, the 20th Brazilian Symposium on Databases*, 2005.
- [4] M. Costa and M. J. Silva. Indexação distribuída de coleções Web de larga escala. *IEEE Latin America Transactions*, 2005. (Submitted for publication).
- [5] V. Gaede and O. Günther. Multidimensional access methods. *ACM Computing Surveys*, 30(2):170–231, 1998.
- [6] S. Göbel and P. Klein. Ranking mechanisms in meta-data information systems for geo-spatial data. In *Proceedings of EOGeo-2002, the 2002 Workshop on Earth Observation and Geo-Spatial Data*, 2002.
- [7] L. L. Hill. *Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface*. PhD thesis, University of Pittsburgh, 1990.
- [8] R. Jin and S. T. Dumais. Probabilistic combination of content and links. In *Proceedings of SIGIR-01, the 24th Conference on Research and Development in Information Retrieval*, 2001.
- [9] C. B. Jones, H. Alani, and D. Tudhope. Geographical information retrieval with ontologies of place. In *Proceedings of COSIT-2001, Spatial Information Theory Foundations of Geographic Information Science*, 2001.
- [10] H. P. Kriegel, M. Schiwietz, R. Schneider, and B. Seeger. Performance comparison of point and spatial access methods. In *Proceedings of SSD-90, the 1st symposium on Design and implementation of large spatial databases*, 1990.
- [11] Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 2003.
- [12] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and implementation of a geographic search engine. Technical Report TR-CIS-2005-03, 2005.
- [13] B. Martins and M. J. Silva. Geographical named entity recognition and disambiguation in Web pages, 2005. (To Appear).
- [14] B. Martins and M. J. Silva. A graph-based ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, 2005.
- [15] B. Martins and M. J. Silva. WebCAT: A Web content analysis tool for IR applications. In *Proceedings of WI-2005, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.
- [16] B. Martins, M. J. Silva, and L. Andrade. Challenges and resources for evaluating geographical IR. In *Proceedings of the Workshop on Geographic Information Retrieval at CIKM 2005*, 2005.
- [17] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library, November 1999. Working Paper.
- [18] T. Rees. C-Squares, a new spatial indexing system and its applicability to the description of oceanographic data. *Oceanography*, 16(1):11–19, 2003.
- [19] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of CIKM-04, the 13th Conference on Information and Knowledge Management*, 2004.
- [20] C. Schlieder, T. Voegelé, and U. Visser. Qualitative spatial reasoning for information retrieval by gazetteers. In *Proceedings of COSIT-02, the 2001 Conference on Spatial Information Theory*, 2001.
- [21] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. tien Lu. Spatial databases-accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):45–55, 1999.
- [22] M. Sintichakis and P. Constantopoulos. A method for monolingual thesauri merging. In *Proceedings of SIGIR-97, the 20th conference on Research and development in information retrieval*, 1997.
- [23] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240, 1970.
- [24] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the Web. In *Proceedings of SSTD-05, the 9th Symposium on Spatial and Temporal Databases*, 2005.
- [25] M. van Kreveld, I. Reinbacher, A. Arampatzis, and R. van Zwol. Distributed ranking methods for geographic information retrieval. In *Proceedings of EWCG-04, the 20th European Workshop on Computational Geometry*, 2004.
- [26] D. Walker, I. Newman, D. Medyckyj-Scott, and C. Ruggles. A system for identifying datasets for GIS users. *International Journal of Geographical Information Systems*, 6(6):511–527.
- [27] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and indexing documents and images*. Morgan Kaufmann, 1994.
- [28] M. F. Worboys. Metrics and topologies for geographic space. In *Advances in Geographic Information Systems Research II: Proceedings of the Symposium on Spatial Data Handling*, 1996.