

Indexing Structures for Geographic Web Retrieval

Leonardo Andrade¹ and Mário J. Silva¹

University of Lisboa, Faculty of Sciences

Abstract. Context-aware search in mobile web environments demands new retrieval methods that rank web resources based on the proximity to users' locations. This paper presents the indexing and ranking architecture of a new geographic web retrieval system that can accept the user location as input and ranks searched items based on the estimated distances between the users and the resources. We describe the criteria to be considered by a geographic similarity metric and the indexing data structures that we created for fast selection of web resources based on proximity.

1 Introduction

In ubiquitous computing environments, the number for geographic information retrieval applications is quickly growing. With the massification of third generation mobile technologies and wi-fi networks new web retrieval applications are no longer confined to the home desktop. For example, travelers who search for “restaurants” in a web search engine are not in general interested in the list of the most popular restaurants in the world - but in those closest to their present location.

A significant portion of documents have a local scope, understood as the geographic region covered by the document. It is estimated that one fifth of the queries submitted to search engines have a geographic context [1]. This context can also be seen as the scope of the query. However, classic systems designed for text retrieval perform poorly when a query related to a location is submitted.

To achieve good results, a natural solution is to build a system with geographic reasoning capabilities. It is necessary to build structures that index spatial information related to the documents and develop algorithms for geographic ranking. Previous works supported geographic indexing and searching under a framework based on classic GIS (Geographic Information Systems) [1, 2]. However, we do not rely on classic spatial indexing structures. Our geographic ranking algorithm combines spatial and non-spatial features, but the index structures are similar to those used in classic text retrieval.

This work is part of the GREASE project which researches on methods, algorithms and software architecture to develop geographic-aware IR systems [3]. GREASE developed and published an ontology of Portuguese geographic names and has been researching methods for classification of documents by geographic scope. Disambiguation of geographic names is performed while assigning scopes to documents. This paper discusses access methods to retrieve documents classified with scopes.

2 Geographic Ranking

Our geographic ranking methods account for geographic terms that may be present in documents and queries. From these terms, we derive the geographic the geographic scopes of documents and queries. The scopes have associated features that are used to compute multiple distance metrics, which are later combined to yield a final geographic similarity measure:

- **Spatial Distance.** According to Tobler’s “first law of geography” [4], *everything is related to everything else, but near things are more related than distant things.* We model the similarity between two geographic scopes as inversely proportional to the distance between the shapes defining the two regions. However, Euclidean or geodesic distances may not be the best metric. Experiments by Montello et al. suggest that cognitive perception does not always match the distance values [5]. Taking the transportation paths into account could be an interesting development (see Figure 1). However, this approach requires a large amount of information collected from the operators of the transportation grid.



Fig. 1. An example of the differences between Euclidean distances and transportation paths. When traveling by car, the shortest distance between the marked locations in the Lisbon area is three times larger than the distance “as the crow flies.” Map from <http://maps.google.com/>.

- **Spatial Overlap.** Two regions may overlap or be adjacent. Regions with larger overlapping areas are more similar. Given the polygons P_q , representing the scope of a query and P_d , representing the scope of a document, the spatial overlap between the two scopes S_q and S_d is defined as:

$$\text{SpatialOverlap}(S_q, S_d) = \frac{P_q \cap P_d}{P_q \cup P_d}$$

- **Ontology Distance** The distance between two scopes is computed as the number of hops in the ontology graph used by the IR system to represent geographic knowledge. Figure 2 depicts a geographic ontology. Non-spatial information present in the ontology, such as the population and political importance of a scope is also taken into account. In web document collections, the hyperlink count between geographic scopes, which may also be present in the ontology, is an interesting feature to account for as well.

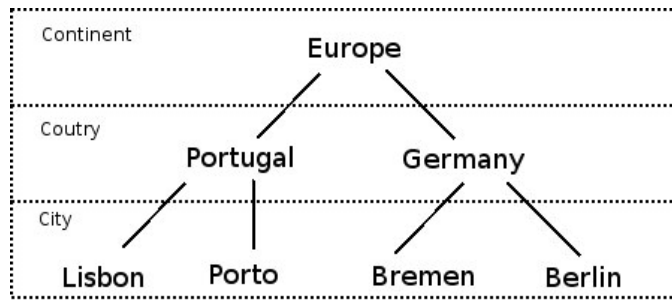


Fig. 2. Example of part of an ontology graph. This example represents a hierarchy of political divisions. The scope *Berlin* is closer to the scope *Germany* than the scope *Lisbon*, as the number of hops between the former two is smaller.

In our environments, the above distances and overlap are combined by means of a weighted sum. The following formula gives the geographic similarity between any two scopes s_1 and s_2 :

$$\begin{aligned} \text{GeographicSim}(s_1, s_2) = & w_1 \times \frac{1}{1 + \text{SpatialDistance}(s_1, s_2)} \\ & + w_2 \times \text{SpatialOverlap}(s_1, s_2) \\ & + w_3 \times \frac{1}{1 + \text{OntologyDistance}(s_1, s_2)} \end{aligned}$$

The geographic similarity is normalized to $[0, 1]$ because the geographic similarity factors in all the terms are in the interval $[0, 1]$ and we observe the constraint $w_1 + w_2 + w_3 = 1$.

The geographic similarity is combined with the textual similarity by means of a weighted combination to obtain the final similarity function used to rank documents given a query:

$$\text{Similarity}(q, d) = b \times \text{TextualSim}(T_q, T_d) + (1 - b) \times \text{GeographicSim}(S_q, S_d)$$

TextualSim computes the indexed terms similarity, usually as a variant of the *TFIDF* ranking in the vectorial space model [6]. T_q and T_d are the indexed terms of query q and document d , and S_q and S_d are their scopes, respectively. The balancing factor b can be tuned to assign more or less importance to geographic ranking and may be:

1. a static value;
2. a user-specified query parameter;
3. automatically assigned as a function of the depth of the geographic scope of the input query.

3 Searching and Indexing

The indexing of spatial information is well studied. Some of the previously developed Geo-IR systems incorporate concepts from Geographic Information Systems and Spatial Databases. Fixed grid models have been used to represent the geographic locations [7, 8]. Zhou et al. used R*-trees in a system where the textual indexes are combined with geographic footprints, which are represented as sets of Minimum Bounding Rectangles [2]. In these systems, the balance between high data definition and slow processing on one hand, and poor data definition and high computing speed on the other has to be considered. The index designer must choose between high precision, useful for good ranking performance, and the response time to deliver the results to the user. To tackle these limitations, our indexing scheme simply maps scopes to documents in an inverted index structure [9]. The inverted index that maps the documents to geo-scopes is depicted in Figure 3.

This simple approach does not have the features that classic solutions such as R-Trees offer (e.g. nearest neighbor finding, distance computation). It is thus necessary to build additional structures to enable geographic ranking. The similarities between geo-scopes can be precomputed and stored in a matrix. The criteria discussed in the previous Section are used in the computation. Table 1 gives an example of a similarity matrix.

As the geographic scopes collection grows, the similarity matrix can reach high dimensions. Processing all the scopes in the collection is costly. As a pruning method, an ordered list of the most related scopes is associated to each scope (Table 2). For each query, only the top- k most related scopes are used in the subsequent ranking computation.

We keep textual and geographic indexes separated. This decision enables efficient handling of queries that only have the textual or geographic parts. This makes index

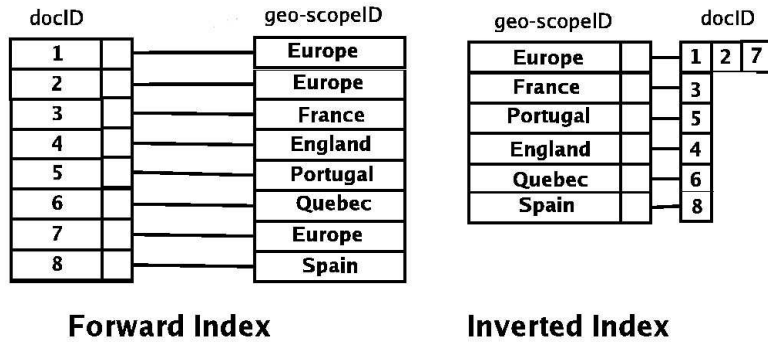


Fig. 3. The geographic scopes inverted index. The inverted index is used for fast document lookup. The forward index is accessed to obtain the scopes assigned to documents.

[scope]	Europe	Portugal	Spain	France	Quebec	England
Europe	1.0	0.2	0.3	0.4	0.02	0.4
Portugal	0.2	1.0	0.6	0.5	0.1	0.4
Spain	0.3	0.5	1.0	0.6	0.02	0.3
France	0.4	0.5	0.6	1.0	0.5	0.5
Quebec	0.08	0.1	0.02	0.5	1.0	0.04
England	0.4	0.4	0.3	0.6	0.04	1.0

Table 1. Similarities table. Similarity between two scopes may not be symmetric.

distribution for load balancing purposes less complex to perform. Previous works have achieved identical conclusions [7]. With this indexing structure, we provide very fast access methods to queries where users request ranked results sets ordered by relevance to a location. However, there is no support for for selectig results based on spatial boolean predicates. These could have to be supported by additional indexing structures, identical to those of GIS.

The size of these structures is easily manageable. In a system with about 500 scopes (e.g. all the Portuguese municipalities, districts, the NUTS I, II and III and some parishes) the size of the matrix will be around $500 \times 500 \times sizeof(integer) \simeq 1MB$. The most related scopes vector will have a smaller size.

- Europe** → France, England, Spain, Portugal, Quebec
- Portugal** → Spain, Europe, France, England, Quebec
- Spain** → France, Portugal, Europe, England, Quebec
- France** → Europe, Quebec, Spain, England, Portugal
- Quebec** → France, Europe, England, Portugal, Spain
- England** → France, Europe, Spain, Portugal, Quebec

Table 2. Most related scopes vectors.

4 Discussion

The methods proposed in this paper are being implemented for incorporation in a prototype of a new geographic search engine, Geotumba. A collection of 9.3 million documents has been scanned for geographic terms, yielding 8.7 million web pages with a geoscope assigned. Figure 4 shows a screen dump of the user interface developed by Freitas et al. [10].

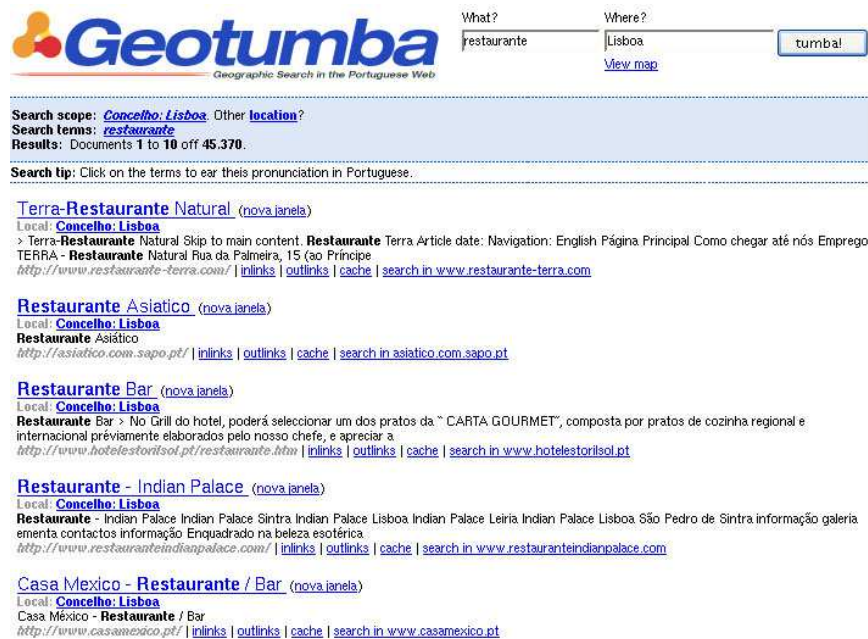


Fig. 4. A query example: “restaurant in Lisbon” with the top results shown.

Forward and inverted indexes with geographic scopes support geographic queries. The indexes are physically stored in inverted files separated from the document text indexes. Access times to these indexes are similar to other inverted files — 0.1 seconds on average on a Pentium 4 processor @ 2.4 GHz and 7200 rpm IDE disks.

The compressed inverted index of the 8.7 million documents collection labeled has 308 different geoscopes (the number of Portuguese municipalities), and is about 35 Megabytes.

5 Conclusions

In this paper, we identified important ranking criteria for a geographic similarity function. In the indexing and searching system, the following design principles were adopted:

1. The geographic scopes are not represented as footprints, but handled as special terms with geographic semantics.

2. The geographic similarity between scopes can be pre-computed and stored in a scopes matrix.
3. The geographic index can have a structure analogous to a text inverted index.
4. Textual and geographic indexes are kept separated.

This design is now under implementation and evaluation. Geotumba is still a work in progress.

6 Acknowledgements

We thank to Bruno Martins, for reviewing the paper and giving valuable suggestions. This work was partially financed by the Portuguese Fundação para a Ciência e Tecnologia through grant POSI / SRI / 40193 / 2001 (GREASE). Leonardo Andrade was supported by scholarships from GREASE and FIRMS — FCT (POSI/ISFL/13/408).

References

1. Sanderson, M., Kohler, J.: Analyzing geographic queries (2004)
2. Zhou, Y., Xie, X., Wang, C., Gong, Y., Ma, W.Y.: Hybrid index structures for location-based web search. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, New York, NY, USA, ACM Press (2005) 155–162
3. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P.: Adding Geographic Scopes to Web Resources. CEUS - Computers, Environment and Urban Systems, Elsevier Science (2005) In print.
4. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. *Economic Geography* **46** (1970) 234–240
5. Montello, D.R., Fabrikant, S.I., Ruocco, M., Middleton, R.S.: Testing the first law of cognitive geography on point-display spatializations (2004)
6. Ricardo Baeza-Yates, B.R.N.: Modern Information Retrieval. Addison Wesley Longman (1999)
7. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-textual indexing for geographical search on the web. In: Proceedings of SSTD-05, the 9th Symposium on Spatial and Temporal Databases. (2005)
8. Markowitz, A., Chen, Y.Y., Suel, T., Long, X., Seeger, B.: Design and implementation of a geographic search engine. Technical Report TR-CIS-2005-03 (2005)
9. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, New York (1994)
10. Freitas, S., Afonso, A.P., Silva, M.J.: Concepção e desenvolvimento de interfaces para o motor de busca geográfico geotumba! In: Encontro Nacional de Visualização Científica (ENVC 05). (2005)