Relevance Ranking for Geographic IR

Leonardo Andrade and Mário J. Silva
University of Lisboa, Faculty of Sciences
1749-016 Lisboa, Portugal
{leonardo, mjs}@xldb.di.fc.ul.pt

ABSTRACT

In this paper, we introduce a geographic similarity operator that computes the relatedness between two geographic places and describe how it is combined with textual ranking. The effectiveness of the geographic ranking is evaluated on the GeoCLEF 2005 collection. We considered various strategies for query formulation and for combining textual ant geographical ranking. For some queries, geographic ranking significantly improves results, while for other queries it does not have a positive impact.

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

In classic IR, retrieved documents are ranked by their similarity to the text of the query. In a search engine with geographic capabilities, the semantics of geographic terms should be considered as one of the ranking criteria. The problem of weighting the geographic importance of a document can be reduced to computing the similarity between two geographic locations, one associated with the query and other with the document.

The degree of relatedness between locations has been previously addressed [2, 3, 11]. Geographic reasoning methods consider topological relationships and spatial proximity. For instance, Nedas and Egenhofer proposed a similarity operator rather than an equivalence operator in the context of GIS databases [7].

The SPIRIT project proposed a set of geographic query operators that use different ranking metrics, depending on the operator considered [12]. For example, in the *near* operator, the Euclidean proximity is taken into account, but in queries containing the *north of* expression, the angular difference is considered. This approach is similar to the 1995 work by Ray Larson, where a set of distinct geographic operators for querying has been proposed but no relevance ranking was considered [4]. Recently, Zhou et al. proposed geographic operators with other metrics, such as spatial overlap and distance [13]. In the work of Jones et al., a solution based on combining measures as distance and ontological relations is introduced [3].

Our approach relies on a geographic ontology, offering the support for geographic reasoning. Essentially, the ontology provides a hierarchical naming scheme for geographic concepts, with transitive "sub-region-of" and name alias capabilities. It describes global geographical information in multiple languages, and integrates data from several public sources [1].

Each document has a single encompassing geographic scope, according to the document's degree of locality. Each scope corresponds to a concept at our ontology. The task of assigning scopes is performed off-line, as a pre-processing operation of our GIR system. It can be seen as having two stages. First, we use a named entity recognition procedure, specifically tailored to the task of recognizing and disambiguating geographical references occurring in the documents. Each reference is matched into the according ontology concepts (e.g. the natural language string "city of Lisbon" is matched into the corresponding concept id at the ontology). Next, we combine the references extracted from the document into a single encompassing geographic scope [5]. If the document contains the references "city of Lisbon" and "city of Porto", the algorithm assigns to the document a scope corresponding to Portugal.

Our work focuses on the ranking module. We handle queries with associated geographic locations and retrieve documents previously annotated with geographic places. The documents are ranked by a combination of textual and geographic relevance.

2. GEOGRAPHIC QUERY PROCESSING

It is estimated that one fifth of the queries submitted to search engines have geographic meaning. Among them, eighty percent can be associated with a geographic place [9]. Geographic queries are split into two parts: the textual part, composed of undifferentiated (terms not associated to a concept on the geographic ontology), and the geographic part, where one or more ontology concept identifiers are derived from the query terms.

Each query can be parsed to a triple <*what*, *relation*, *where*>, where the *what* term is used to specify the general non-geographical aspect of the information need, the *where* term is used to specify the geographical areas of interest, and the *relation* term is used to specify a spatial relationship connecting *what* and *where*. Another

Workshop on Geographic Information Retrieval, SIGIR '06, USA

paper submitted to this workshop describes algorithms for parsing these queries [6]. Complex queries, containing multiple spatial relationships (e.g. “monuments south of Porto and north of Lisbon”) can be seen as combinations of such triples. In this paper we focus on implementation of a similarity function for geographic scopes. The function that computes the final relevance weight given a query q for each document d is:

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

where T_q (the *what* element) and T_d are the text of the query and the document, respectively. S_q (the *where* element) and S_d represent the geographic scopes. The textual similarity is weighted by the BM25 formula [8], normalized using a method that maps the BM25 weight into the [0,1] interval [10]. The *GeographicSim* function is already normalized to the [0,1] interval, as shown below.

2.1 Geographic Relevance Ranking

Egenhofer introduced the main notions of common-sense knowledge about the spatial world [2]. Taking these concepts into account, we compute the similarity between two geographic scopes from information in the ontology. Given a query scope S_q and a document scope S_d:

Inclusion tests if S_d is inside S_d, and weights the relationship degree between both scopes by the number of descendants in the ontology:

$\mathit{Inclusion}(S_q,S_d) = \frac{\mathit{NumDescendants}(S_d)+1}{\mathit{NumDescendants}(S_q)+1}$ if $S_d \subseteq S_q$;
0 otherwise

This formula returns values in]0,1], yielding the maximum value when both scopes are equal and the minimum when S_d has no descendants. $\mathit{NumDescendants}(S)+1$ returns the number of scopes spatially inside S plus the scope itself, as derived from the spatial world (“sub-region-of” relationships in the ontology).

Proximity is the inverse of distance:

$\mathit{Proximity}(S_q,S_d) = \frac{1}{1+\mathit{Distance}(S_q,S_d)/\mathit{Diagonal}(S_q)}$

where the Euclidean distance is normalized by the diagonal of the minimum bounding rectangle (MBR) of the query scope.

Siblings is a binary function that tests if S_q and S_d are siblings in the ontology graph:

$\mathit{Siblings}(S_q,S_d) = 1$ if $\exists S_x : \mathit{parent}(S_q) = S_x \wedge \mathit{parent}(S_d) = S_x$;
0 otherwise.

The functions above reflect Egenhofer’s aphorism of *Topology Matters, Metric Refines* [2]. The topological notion of *Inclusion* is refined by the descendant scopes count. The *Siblings* function represents the intuitive notion of *Boundaries being sometimes entities, sometimes not* also by the same author.

In the *GeographicSim* final formula these geographic similarity notions are combined as a weighted sum:

$$\mathit{GeographicSim}(S_q,S_d) = bb \times \{\mathit{Inside}(S_q,S_d) + \mathit{Proximity}(S_q,S_d)\} + (1-bb) \times \mathit{Siblings}(S_q,S_d)$$

$0 \leq bb \leq 1$ so that the final value lies in [0,1].

As first and second terms are inter-dependable —e.g.: $\mathit{Proximity}(S_d,S_q) \neq 1 \Longrightarrow \mathit{Inclusion}(S_d,S_q) = 0$ — only one balancing coefficient bb is necessary.

3. EVALUATION

We used a manually built ontology covering the whole planet [1]. The ontology has 12653 distinct geographic scopes. Some scopes present in the geographic ontology used to classify the documents did not have a shape associated with. For such cases, the shape assigned was the interpolation of a the medium point of all the corresponding ancestors in the ontology.

One of the key design decisions was to use the detailed shape of scopes, not only the minimum bounding rectangles (MBRs), when computing the inclusion and distance. A common solution to simplify the indexing and ranking computation is to use MBRs and fixed grid schemes, but this may lead to poor ranking results, as this approaches may be sometimes abusive over-simplifications of the real world, (see Figure 1).

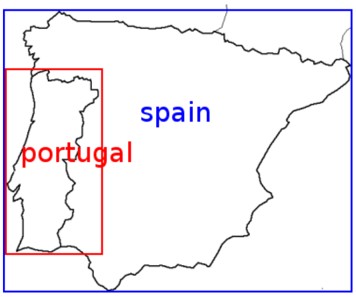

Figure 1: The bounding boxes of Portugal and Spain. This figures illustrates the problems of adopting MBRs as representations for the polygons. In the example, every Portuguese city is inside Spain, which is not true in the real-world.

The geographic scopes have three different geometric types: 1) Polygons — Continents, Countries, Administrative Regions; 2) Lines — Rivers; 3) Points — Cities, Towns.

Over this testbed, two evaluations were performed:

1. Observing the behavior of the criteria chosen for ranking in a selection of typical queries.

2. Testing on GeoCLEF 2005 topics and relevance judgments.

3.1 Relevance Criteria Analysis

The parameters of the similarity operator were set to *b=0.10; bb=0.90*, result of previous manual tuning. The textual part of the query was ranked by BM25 with the default parameter values (*K1=2.0, b=0.75*) with multiple fields weighting. All the tests were performed in the Portuguese language collection of GeoCLEF — 210734 documents from Portuguese and Brazilian newspapers. The top results are shown for some selected queries (submitted in Portuguese).

In Table 1, the query for searching for restaurants near the Danube river returns a set of documents about cities near that scope. As this European river is represented by a line, the decisive weighting term is the proximity measure — the inclusion is not measurable. In these experiments, for the distance computation, the document scopes are represented by the centroid of the shape. To observe the differences when distance is computed from MBRs, we submitted the same query using the bounding box as the shape (see Table 2). Some places, such as Pisa and San Marino, are actually far away from the Danube, but if the river is represented as a bounding box, these places are considered inside the river, and become highly ranked.

Rank	ASSIGNED SCOPE	Text	Geo	Final
1	Wien	0.795	0.473	0.505
2	Bayern	0.411	0.420	0.419
3	Strasbourg	0.715	0.299	0.341
4	Austria	0.639	0.290	0.325
5	Brno	0.804	0.263	0.317
6	Vienna	0.421	0.290	0.303
7	Austria	0.370	0.290	0.298
8	Bulgaria	0.580	0.266	0.298

Table 1: Results for the query: Text="restaurant" AND GeographicSim(Danube). Each row denotes a document retrieved.

Rank	ASSIGNED SCOPE	Text	Geo	Final
1	Wien	0.795	0.5	0.529
2	Brasov	0.781	0.5	0.528
3	San Marino	0.764	0.5	0.526
4	Switzerland	0.756	0.5	0.525
5	Berne	0.734	0.5	0.523
6	Monza	0.716	0.5	0.521
7	Pisa	0.700	0.5	0.520
8	Liguria	0.674	0.5	0.517

Table 2: Results for the query: Text="restaurant" AND GeographicSim(Danube), considering the shape corresponding to the Danube river as an MBR instead of a line.

Rank	ASSIGNED SCOPE	Text	Geo	Final
1	India	0.803	0.572	0.595
2	Brazil	0.449	0.530	0.522
3	Manaus	0.497	0.500	0.499
4	Ceara	0.491	0.500	0.499
5	Nicaragua	0.444	0.503	0.497
6	Luanda	0.451	0.500	0.495
7	Paulista	0.421	0.500	0.492
8	Zaire (*Angola region*)	0.409	0.500	0.490

Table 3: Results for the query: Text="malaria epidemic" AND GeographicSim(Tropics).

#	Query input	GeoSim	MAP
1	Title	NO	**0.1823**
2	Title without place names	YES	**0.1657**
3	Title	YES	**0.1785**
4	Title with place names not mandatory	NO	**0.1705**
5	Title with place names not mandatory	YES	**0.1850**

Table 4: Medium Average Precision with different configurations for the 25 queries from GeoCLEF 2005.

In the query representing the search for "malaria epidemics" in the tropical part of the globe (Table 3) all the places have the same topological relation to the query scope. However, their relevances are different because the refinement introduces weighting scopes by the number of descendants.

3.2 Experiments with GeoCLEF 2005

To evaluate the effectiveness of the proposed geographic relevance ranking methods against simple text relevance, we tested our similarity measures with the GeoCLEF 2005 collection and relevance judgments. The configuration was the same as above, but geographic ranking parameters were set to $b = 0.6; bb = 0.9$ (the best setting obtained by manual tuning). We set $Diagonal(S_q) = 1$ in all experiments. The test collection was the English language GeoCLEF corpus, with 166743 documents from *Los Angeles Times* and *Glasgow Herald* newspapers. Among these documents, 140159 had geographic scopes assigned with the algorithm described in Section 1.

A geographic scope was manually assigned from the geographic term present in the topic (e.g: for the query "Rice Imports in Japan" the associated scope is "Japan"). For some queries, no geographic scope was assigned (e.g.: the place name "Scottish Trossachs") because this place name can not be associated with a scope from the ontology having coordinates. In these queries, a manually assigned *GeographicSim* could not be computed. Of the 25 queries, 15 had a scope. The stop words in the title were removed, and no query expansion was performed. We choose the Mean Average Precision (MAP) to measure the effectiveness. This metric, used regularly in IR evaluation, takes into account the precision of the result set according to the position of the relevant documents retrieved. We run five experiments (summarized in Table 4):

1. Query input: the topic title. No geographic relevance ranking was computed. Example: *golf tournaments europe*
   With an average MAP on the 25 queries of 0.1823, this query served as a baseline for the experiments. As the effectiveness of the BM25 formula is well known, the textual ranking may be effective in ranking the documents where the term "Europe" has more weight, contributing indirectly to the geographic relevance ranking. The problem with this query is the exclusion of all the documents that do not contain the "Europe" term. For example, a document about golf tournaments in France would be excluded from the result set.

2. Query input: the topic title, with the geographic term deleted but with *GeographicSim* computation by the use of a geographic scope. Example: *golf tournaments* + *GeographicSim(europe)*
   The MAP was the lowest of the experiments (0.1657). The cause of degradation is the low BM25 performance, when geographic terms are stripped from the query. The gains that might be obtained with geographic similarity do not compensate the loss of effectiveness of BM25.

3. Query input: the topic title, with *GeographicSim* computation. Example: *golf tournaments europe* + *GeographicSim(europe)*
   The MAP was higher than on the previous experiment (0.1785), although worse than on experiment 1. The importance of the textual ranking is evident again.

4. Query input: the topic title, with the geographic name optional. No geographic relevance ranking computation. Example: *golf tournaments OR golf tournaments europe.*
   As the search system is Boolean, the term "europe" in this example is not mandatory, but a document with this term will be ranked higher. The MAP of this experiment was lower than on experiment 1. This is due to retrieving documents with no geographic filtering, such as descriptions of "golf tournaments" outside Europe.

5. Query input: the topic title, with geographic name optional and *GeographicSim*. Example: *golf tournaments OR golf tournaments europe + GeographicSim(europe)*
   The MAP was the highest of all experiments (0.1850). This query model combines a larger number of documents retrieved with both textual and geographic ranking.

Some queries are more influenced by the use the geographic similarity operator. The queries where the geographic similarity operator performs better are those where a relevant document is less likely to have an exact match with the query place name in the text. For example, having the query *golf tournaments in Europe*, it is unlikely for a document about a golf tournament in France to have the term "Europe" in its text, because golf courses are usually referred by their name and country, with no reference to the continent. The same consideration is valid to the query *child labor in Asia*. On the other hand, a newspaper article about *rice imports in Japan* will likely contain the geographic term "Japan."

The goal of the experiments with the GeoCLEF collection was to observe trends when the geographic ranking is included in the query processing. Pre-processing of the textual part of the queries in the experiments above was simplistic, contrary to other approaches of proven effectivness in CLEF tasks, such as query expansion and stemming. As a result, comparisons between MAP values of the described runs with the MAP of runs generated by GeoCLEF 2005 participants can not be used to assert the merits of different strategies evaluated. In addition, as the differences in MAP for the 25 topics are very small, statistical significance is not sufficient for deriving definitive conclusions.

Another characteristic of the geographic relevance ranking is its high sensitivity to the correctness of scope assignments. Relevant documents annotated with the wrong scope are not considered relevant to a geographic query. We could observe that some of the scope assignment for these experiments had errors that may have negatively influenced the performance of the geographic ranking.

4. CONCLUSIONS

The following conclusions result from the evaluation described above:

- Geographic ranking by inter-scope similarity is effective for some geographic queries, with a good overall performance in the 2005 GeoCLEF query set.
- Some queries are "more geographic" than others. The optimal balance between textual and geographic ranking is query-dependent.
- Textual relevance ranking is also good on geographic ranking, when computed over sufficient geographic terms.

5. ACKNOWLEDGEMENTS

We would like to thank to Bruno Martins for discussing early versions of this work and to the reviewers for useful 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 —(POSI/ISFL/13/408).

6. REFERENCES

[1] M. S. Chaves, M. J. Silva, and B. Martins. A Geographic Knowledge Base for Semantic Web Applications. In *Proc. of the 20th Brazilian Symposium on Databases, Uberlândia, Minas Gerais, Brazil*. In press, October, 3–7 2005.

[2] M. J. Egenhofer and D. Mark. Naive geography. In A. U. Frank and W. Kuhn, editors, *Spatial Information Theory: a theoretical basis for GIS*, volume 988 of *Lecture Notes in Computer Science*, pages 1–16. Springer-Verlag, Berlin, 1995.

[3] 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.

[4] R. R. Larson. Geographic information retrieval and spatial browsing, 1995. Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information, pages 81–123.

[5] B. Martins and M. J. Silva. A graph-ranking algorithm for geo-referencing documents. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 741–744, Washington, DC, USA, 2005. IEEE Computer Society.

[6] B. Martins, M. J. Silva, S. Freitas, and A. P. Afonso. Handling locations in search engine queries, 2006. Workshop on Geographical Information Retrieval).

[7] K. Nedas and M. Egenhofer. Spatial similarity queries with logical operators. In *SSTD '03 – Eighth International Symposium on Spatial and Temporal Databases*, July 2003.

[8] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M.Gatford. Okapi at trec-3. Third Text REtrieval Conference (TREC-3), 2000.

[9] M. Sanderson and J. Kohler. Analyzing geographic queries, 2004.

[10] R. Song, J.-R. Wen, S. Shi, G. Xin, T.-Y. Liu, T. Qin, X. Zheng, J. Zhang, G. Xue, and W.-Y. Microsoft research asia at web track and terabyte track of trec 2004. In *TREC*, 2004.

[11] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970.

[12] 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.

[13] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 155–162, New York, NY, USA, 2005. ACM Press.