The University of Lisbon at GeoCLEF 2006

Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade and Mário J. Silva
University of Lisbon, Faculty of Sciences
1749-016 Lisboa, Portugal
{bmartins,ncardoso,mchaves,leonardo,mjs}@xldb.di.fc.ul.pt

Abstract

This paper details the participation of the XLDB group from the University of Lisbon at the GeoCLEF task of CLEF 2006. We tested text mining methods that make use of an ontology to extract geographic references from text, assigning documents to encompassing geographic scopes. These scopes are used in document retrieval through a ranking function that combines BM25 text weighting with a similarity function for geographic scopes. We also tested a topic augmentation method, based on the geographic ontology, as an alternative to the scope-based approach. We analyze the obtained results and discuss directions for future improvements.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Evaluation, Geographical IR, Text Mining, Geographic Relevance, GeoCLEF

1 Introduction

This paper reports the experiments of the XLDB group from the University of Lisbon on the GeoCLEF track of CLEF 2006. Our main objective was comparing two strategies, specific for geographic IR, with a more standard IR approach. These specific strategies were i) using text mining for extracting and combining geographic references from the texts, in order to assign documents to geographic scopes, together a with a ranking function that combines scope similarity with a state-of-the-art text ranking scheme, and ii) augmenting the geographical terminology used in the topics through the use of an ontology.

2 System Description

Figure 1 outlines the architecture of the prototype system that was used in our experiments. Many of the components came from a Web search engine developed by our group that is currently being extended with geographic IR functionalities (a demonstration is available online at `local.tumba.pt`). For CLEF, the search engine crawler was replaced by a simpler component, responsible for loading documents into the repository. The user interface was replaced by two other components, one that generates queries from CLEF topics, and another that outputs results in the CLEF format.

This research was partially supported Fundação para a Ciência e Tecnologia, under grants POSI/SRI/40193/2001 and SFRH/BD/10757/2002.

The components related to geographic text mining are shown in the gray boxes of Figure 1. They are essentially a pipeline of operations for associating documents to appropriate geographic scopes, and mechanisms for processing topics (i.e. geographic queries) also according to scopes. In order to assist in recognizing geographical terminology, both over documents and topics, the system relies on an ontology that encodes place names and the semantic relationships among them. An R*-tree index structure is used to store the spatial information (centroids and bounding boxes) defined at the ontology [2]. The other information (e.g. place names and relationships) is kept on specialized indexes, built using traditional data structures such as lists and hash tables. Topics are transformed into triples of the form $< what, relation, where >$, where *what* corresponds to the non-geographical aspect, *where* corresponds to a geographic area of interest (i.e. geographic scope), and *relation* specifies a spatial relationship connecting *what* and *where* [13]. Finally, for ranking results, the system uses a linear combination of the BM25 text weighting scheme [14] with a similarity function for geographic scopes.

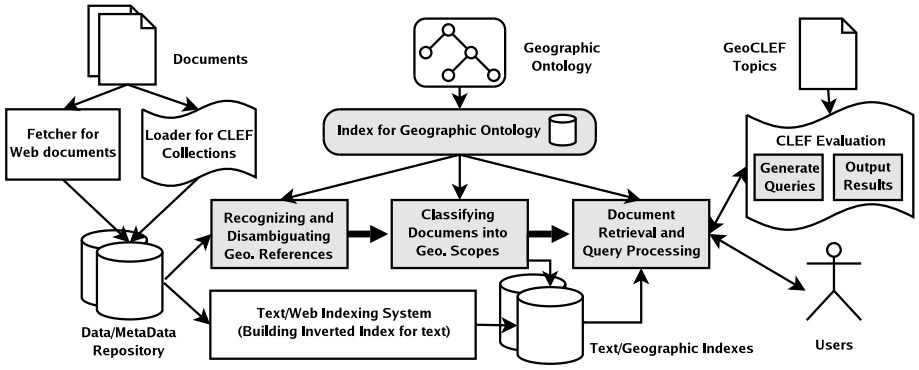


Figure 1: Architecture of the geographical IR prototype that we used in GeoCLEF 2006.

2.1 Text retrieval and ranking

The text retrieval module briefly described here, as well as the blind feedback expansion scheme mentioned in Section 3, was also used in our submissions to the CLEF 2006 Ad-hoc task. The reader should refer to Cardoso et al. [4] for additional details and a discussion on the obtained results. Text retrieval relies on an inverted index for the document collections (we separately indexed the Portuguese and the English collections), providing the support for simple, ranked retrieval. We used the BM25 ranking scheme [15], where the score for each document corresponds to the weighted sum of the terms that occur in both the document and the query. Each term t_i has a weight according to the formula:

$$BM25(t_i) = \frac{(k_1+1) \times term\_freq(t_i)}{k_1 \times ((1-b) + b \times \frac{doc\_len}{avg\_doc\_len}) + d} \times \log(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5}) \quad (1)$$

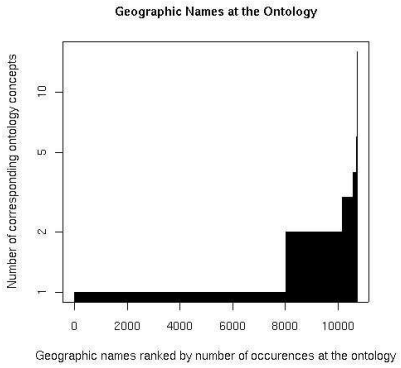
The k_1 and b parameters were set to the standard values of 2.0 and 0.75, respectively. An extension of the BM25 scheme for structured documents, proposed by Robertson et al., was also applied [14]. We assumed that the first three sentences of each document should be weighted as more important, following the intuition that in news articles, the first sentences usually summarize the entire document. We gave a weight of 3 to the first sentence, and a weight of 2 to the following two sentences, mapping the original documents into more verbose ones where content is repeated according to the weighting values.

2.2 Geographic ontology

The ontology is a central component of the system, offering the support for geographic reasoning. It models both the vocabulary and the relationships between geographic concepts, providing a hierarchical naming scheme with transitive "sub-region-of" and name alias capabilities. For our experiments, we developed

an ontology with global geographic information in multiple languages, by integrating data from several public sources [5]. Some characterization statistics are listed at the right part of Figure 2. The considered information includes names for places and other geographic features, adjectives of place, place type information (e.g. street or city), relationships among concepts (e.g. adjacent or sub-region-of), demographics data, occurrence statistics for the geographic names over a large collection of Web documents, spatial coordinates (i.e. centroids) and bonding boxes for the geographic concepts. We would be happy to contribute this resource for future editions of GeoCLEF, or customize it as appropriate. Since our participation at GeoCLEF 2005, some minor bugfixes were made to this resource, and we also added considerably more spatial information (coordinates and bounding boxes) and adjacency relations.

Figure 2: Statistical characterization of the geographic ontology.



Ontology statistic	Value
Ontology concepts	12,654
Geographic names	15,405
Unique geographic names	11,347
Concept relationships	24,570
Concept types	14
Part-of relationships	13,268
Adjacency relationships	11,302
Concepts with spatial coordinates	4,204 (33.2%)
Concepts with bounding boxes	2,083 (16.5%)
Concepts with demographics	8,206 (64.8%)
Concepts with corpus frequency	10,057 (79.5%)

Each geographic concept can be described by several names. The chart presented left of Figure 2 illustrates the ambiguity present in these names, by plotting for each name the number of different corresponding concepts. Even in our medium sized ontology, place names with multiple occurrences are not just a theoretical problem (more than 25% of the place names correspond to multiple ontology concepts).

Note that some geographic concepts do not have spatial coordinates or population information. In these cases, we propose to interpolate values from sibling concepts at the ontology (e.g. the centroid of a given region can be approximated by the average of all centroids from its sub-regions, and the population of a region can be computed by the sum of the population counts for all its sub-regions). This aspect assumes a particular importance, as we propose using these values for the computation of a similarity function.

2.3 Recognizing place references and assigning documents to geographic scopes

In the text mining approach, each document was assigned to a single encompassing geographic scope, according to the document's degree of locality. Each scope corresponds to a concept at our ontology. Scope assignment was performed off-line, as a pre-processing task that had two stages. First, we used a named entity recognition procedure, specifically tailored for recognizing and disambiguating geographic references over text, which relies on place names at the ontology together with lexical and contextual clues. Each reference was matched into the corresponding ontology concept (e.g. a string like "city of Lisbon" would be matched into a concept identifier at the ontology). Next, we combined the references extracted from each document into a single encompassing scope, using a previously described algorithm that explores relations among geographic concepts [12]. This is essentially a graph-ranking approach similar to PageRank, assigning ontology concepts to confidence scores and then selecting the highest scoring concept as the scope. For instance, if a document contained references to the cities of "Alicante" and "Madrid", it would be assigned to the scope "Spain", as both cities have a part-of relationship with that country.

On the English collection from the CoNLL-03 contest [17], our system has a precision of 0.85 and a recall of 0.79, in the simple task of recognizing place references (reference disambiguation cannot be evaluated with this resource, as it lacks the associations from places to ontology concepts). The best reported system from CoNLL-03 achieved over 0.95 in both precision and recall, showing that our system can still be improved. As for the scope assignment procedure based on graph-ranking, it achieved an accuracy of 0.92 on the well-known Reuters-21578 newswire collection [12].

2.4 Processing GeoCLEF topics

GeoCLEF topics were also assigned to corresponding geographic scopes, so that we can match them to the scopes of the documents. Topic titles were first transformed into $< what, relation, where >$ triples, where *what* specifies the non-geographical aspect of the topic, *where* specifies the geographic area of interest (latter disambiguated into a scope), and *relation* specifies a spatial relationship connecting *what* and *where*. The algorithm for doing this is described in a separate publication [13]. Two different types of relationships could be found at the topics, namely "near" and "contained at." Topic GC40 (cities near active volcanoes) could not be processed through this mechanism, and was therefore treated as non-geographical (i.e. with no *where* and *relation* terms). Some topics (e.g. topic GC29, diamond trade in Angola and South Africa) were assigned to multiple scopes, according to the different locations referenced in the *where* part.

2.5 Geographical Similarity

Geographic relevance ranking requires a mechanism for computing the similarity among the scopes assigned to the documents and the scopes assigned to the topics. Geographic scopes correspond to concepts at the ontology, and we can use the different types of information, available at our ontology, to compute similarity. Taking inspiration in previous works [1, 8, 9, 16], we chosen to use the following heuristics:

2.5.1 Topological distance from hierarchical relations

Topological part-of relations, defined at the ontology, can be used to infer similarity. We have, for instance, that Alicante is part of Spain, which in turn is part of Europe. Alicante should therefore be more similar with Spain than with Europe. We used the formula below, similar to Lin's similarity measure [11], to compute the similarity according to the number of transitively common ancestors for the two scopes.

$$OntSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is the same or equivalent to } scope_2 \\ \frac{2 \times NumCommonAncestors(scope_1, scope_2)}{NumAncestors(scope_1) + NumAncestors(scope_2)} & \text{otherwise} \end{cases} \quad (2)$$

For example, considering the ontology on Figure 3, the similarity between the scopes corresponding to "Alicante" and "Spain" is $\simeq 0.67$, while the similarity between "Alicante" and "Europe" is 0.4.

2.5.2 Spatial distance

Spatially near concepts are in principle more similar. However, people's notion of distance depends on context, and $scope_1$ being near to $scope_2$ depends on their relative sizes and on the frame of reference.

We say that distance is 0, and therefore similarity is 1, when one of the scopes is a sub-region of the other. We also normalized distance according to the diagonal of the minimum bounding rectangle for the scope of the topic (i.e. $scope_2$ in the formula bellow), this way ensuring that different frames are treated appropriately. We employed a double sigmoid function with the center corresponding to the diagonal of the bounding rectangle. This function has a maximum value when the distance is at the minimum, and smoothly decays to 0 as the distance increases, providing a non-linear normalization. The curve is illustrated at Figure 3. The formula is given below, where D is the spatial distance between $scope_1$ and $scope_2$ and D_{MBR} is the diagonal distance for the minimum bounding rectangle corresponding to $scope_2$.

$$DistSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is part of or parent of } scope_2 \\ 1 - \left(\frac{1 + sign(D - D_{MBR}) \times (1 - \exp(-(\frac{D - D_{MBR}}{D_{MBR} \times 0.5})^2))}{2}\right) & \text{otherwise} \end{cases} \quad (3)$$

2.5.3 Shared population

When two regions are connected through a part-of relationship, the fraction of the population from the more general area that is also assigned to the more specific area can be used to compute a similarity measure.

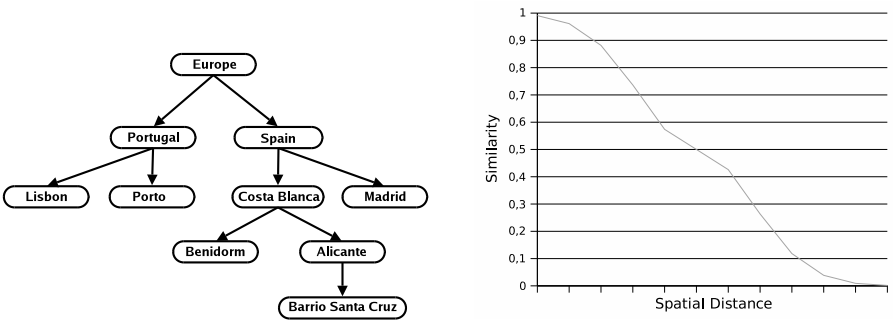


Figure 3: An example ontology with hierarchical part-of relations (on the left) and the double sigmoid function used to normalize the spatial distance (on the right).

This metric corresponds to the relative importance of one region inside the other, and it also approximates the area of overlap. The general formula is given below:

$$PopSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is the same or equivalent to } scope_2 \\ \frac{PopulationCount(scope_1)}{PopulationCount(scope_2)} & \text{if } scope_1 \text{ is part of } scope_2 \\ \frac{PopulationCount(scope_2)}{PopulationCount(scope_1)} & \text{if } scope_2 \text{ is part of } scope_1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

2.5.4 Adjacency from ontology

Adjacent locations are, in principle, more similar than non-adjacent ones. Using the adjacency relationships from the ontology, we can assign a score of 1 if the two scopes are adjacent, and 0 if not.

$$AdjSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is adjacent to } scope_2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

2.6 Score combination for geographic retrieval and ranking

The previously discussed measures, computed by different mechanisms, need to be combined into an overall similarity measure, accounting for textual and geographical aspects. We tried a linear combination due to it's simplicity. Normalization is a crucial aspect, making different scores comparable. The previously given geographic measures already produce values in the interval $[0,1]$. For the BM25 formula, we used the normalization procedure presented by Song et al. [18], corresponding to the formula below:

$$NormBM25(doc, query) = \frac{\sum_{t_i \in doc} BM25(t_i) \times weight(query, t_i)}{\sum_{t_i \in doc} \log(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5})(k_1 + 1)} \quad (6)$$

The $weight(query, t_i)$ parameter corresponds to 1 if term t_i is in the query, and 0 otherwise. The final ranking score combined the normalized BM25 value with the similarity between the geographic scope of the document and the most similar scope of the query (note that each query could have more than one geographical scope assigned to the *where* term). It is given by the formula below:

$$\begin{aligned} Ranking(doc, query) = \quad & (0.5 \times NormBM25(doc, query)) + \\ & (0.5 \times MAX_{s \in scopes_{query}}(GeoSim(scope_{doc}, s))) \end{aligned} \quad (7)$$

where the geographical similarity *GeoSim* is given by:

$$GeoSim(s_1,s_2) = \begin{array}{l}(0.5 \times OntSim(s_1,s_2)) + (0.2 \times DistSim(s_1,s_2)) + \\ (0.2 \times PopSim(s_1,s_2)) + (0.1 \times AdjSim(s_1,s_2))\end{array} \quad (8)$$

The combination parameters were based on the intuition that *topology matters and metric refines* [7], in the sense that we gave more weight to the similarity measures derived from topological relations at the ontology. Still, for future work, we plan on using a systematic approach for finding the optimal combination. We also plan on using the confidence scores from the geographic scopes (recall than $scope_{doc}$ was assigned with a given confidence score) in ranking, adjusting the weight of *GeoSim* accordingly.

3 Description of the runs submitted

Table 1 summarizes the submitted runs, a total of eight with half for the Portuguese and half for the English monolingual tasks. We did not submit runs for other languages, restricting our efforts to the Portuguese (Público and Folha) and English (LA Times and Glasgow Herald) document collections.

Table 1: Runs submitted to GeoCLEF 2006.

Run Number	Description
1 (PT and EN)	Baseline using manually-generated queries from the topics and BM25 text retrieval.
2 (PT and EN)	BM25 text retrieval. Queries were generated from blind-feedback expansion of *what* terms at the topic title, together with the original *where* and *relation* terms also at the topic title.
3 (PT and EN)	Geographic relevance ranking using geographic scopes. Queries were generated from blind-feedback expansion of *what* terms at the topic title. The *where* terms in the topic title were matched into scopes.
4 (PT and EN)	BM25 text retrieval. Queries were generated from blind-feedback expansion of *what* terms at the topic title, together with the augmentation of *where* and *relation* terms also at the topic title.

In runs 2, 3, and 4, the non-geographical terms of each topic (i.e. the *where* terms obtained from the topic titles) were expanded through a blind feedback mechanism [6]. Essentially, the method adds the 15 top-ranked terms from the top 10 ranked documents of an initial ranking [4].

In run 3, ranked retrieval was based on the combination of BM25 with the similarity score computed between the scopes assigned to the topics and the scope of each document, as described in Section 2.6.

In run 4, the *where* terms were augmented, using information from the ontology to get semantically related place names, either topologically or by proximity. As stated by Li [10], a hierarchical structure can be used to expand place names in two directions, namely downward and upward. Downward expansion is appropriate for queries involving a "contained-at" spatial relation, extending the influence of a place name to all of its descendants, in order to encompass subregions of the location specified in the query. Upward expansion can be used to extend the influence of a place name to some or all of its ancestors, and then possibly downward again into other sibling places. This can be used for queries involving a "near" spatial relation, although many irrelevant place-names can this way also be included. We have chosen not to use upwards expansion, instead using adjacency relations from the ontology and near concepts computed from the spatial coordinates. The general augmentation procedure involved the following steps:

1. Use the ontology to get concepts that correspond to sub-regions of the *where* term(s) obtained from the topic title (i.e. topologically related concepts).
2. If the *relation* term obtained from the topic title corresponds to the "near" relationship, use the ontology to get the adjacent regions to the *where* term(s).
3. If the *relation* term obtained from the topic title corresponds to the "near" relationship, use the ontology to get the top k nearest locations from the *where* term(s).
4. Rank the list of concepts that was obtained from the previous three steps according to an operational notion of importance. This ranking procedure is detailed in a separate publication [13], taking into account heuristics such as concept types (e.g. countries are preferred to cities, which in turn are preferred to small villages), demographics, and occurrence frequency statistics for the place names.
5. Select the place names from the 10 top ranked concepts to augment the original topic.

4 Results

In table 2, we summarize the `trec_eval` output for the official runs we submitted. For the definition of the various measures, run `trec_eval -h`.

Table 2: Results obtained for the different submitted runs.

Measure	Run 1		Run 2		Run 3		Run 4	
	PT	EN	PT	EN	PT	EN	PT	EN
num-q	25	25	25	25	25	25	25	25
num-ret	5232	3324	23350	22483	22617	21228	10483	10652
num-rel	1060	378	1060	378	1060	378	1060	378
num-rel-ret	607	192	828	300	519	240	624	260
map	0,301	0,303	0,257	0,158	0,193	0,208	0,293	0,215
R-prec	0,359	0,336	0,281	0,153	0,239	0,215	0,346	0,220
bpref	0,321	0,314	0,254	0,140	0,208	0,191	0,306	0,199
gm-ap	0,203	0,065	0,110	0,027	0,074	0,024	0,121	0,047
ircl-prn.0.00	0,708	0,677	0,553	0,367	0,715	0,503	0,716	0,543
ircl-prn.0.10	0,601	0,581	0,487	0,254	0,485	0,443	0,577	0,380
ircl-prn.0.20	0,512	0,415	0,438	0,215	0,365	0,320	0,499	0,287
ircl-prn.0.30	0,437	0,382	0,357	0,210	0,288	0,293	0,455	0,266
ircl-prn.0.40	0,390	0,339	0,292	0,171	0,199	0,234	0,389	0,223
ircl-prn.0.50	0,347	0,304	0,256	0,162	0,163	0,221	0,305	0,215
ircl-prn.0.60	0,265	0,267	0,220	0,143	0,095	0,164	0,235	0,197
ircl-prn.0.70	0,145	0,200	0,160	0,120	0,059	0,121	0,163	0,170
ircl-prn.0.80	0,080	0,156	0,115	0,107	0,034	0,089	0,101	0,124
ircl-prn.0.90	0,012	0,117	0,069	0,076	0,004	0,032	0,021	0,113
ircl-prn.1.00	0,002	0,116	0,012	0,056	0,000	0,025	0,003	0,094
P5	0,488	0,384	0,416	0,208	0,432	0,240	0,536	0,288
P10	0,496	0,296	0,392	0,180	0,372	0,228	0,480	0,240
P15	0,472	0,243	0,360	0,171	0,341	0,195	0,440	0,224
P20	0,442	0,224	0,350	0,156	0,318	0,170	0,424	0,212
P30	0,399	0,197	0,324	0,144	0,287	0,147	0,369	0,184
P100	0,218	0,072	0,193	0,073	0,162	0,068	0,218	0,084
P200	0,119	0,037	0,130	0,044	0,090	0,040	0,118	0,049
P500	0,048	0,015	0,063	0,022	0,039	0,019	0,050	0,021
P1000	0,024	0,008	0,033	0,012	0,021	0,100	0,025	0,010

In both the Portuguese and English subtasks, run 1 achieved the best results, corresponding to MAP scores of 0.301 and 0.303, respectively. Contrary to our expectations, run 4 also outperformed run 3, showing that a relatively simple augmentation scheme for the geographic terminology at the topics can outperform the text mining approach. In GeoCLEF 2005, our best run achieved a MAP score of 0.2253 (also a baseline with manually-generated queries). Also in our GeoCLEF 2005 submissions, an automatic technique that involved geographic scope assignment, although with a much simpler retrieval scheme, achieved a MAP score of 0.1379 [3]. The best system in GeoCLEF 2005 achieved a MAP score of 0.3936.

Figure 4 shows the average precision for the 25 individual topics, for runs 3 and 4 and in the Portuguese and English subtasks.

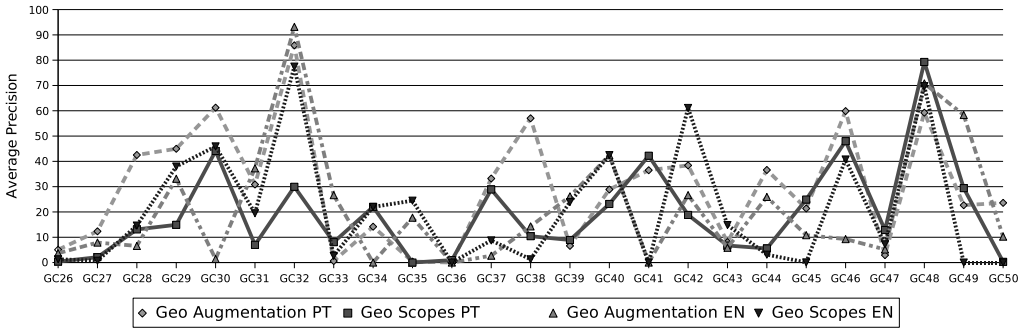


Figure 4: Average precision for the 25 topics in runs 3 and 4, for both the Portuguese and English subtasks.

We analyzed the documents retrieved for some of the topics, together with the scopes that had been assigned to them, particularly focusing on GC32 and GC48. It is our belief that run 3 performed worse due to errors in scope assignment, and to the fact that having each document assigned to a single geographic scope can be too restrictive. We are now performing additional experiments using the GeoCLEF 2006 relevance judgments, reassigning geographic scopes to the documents and this time allowing multiple scopes for each one. The proceedings paper will also report these new experiments.

5 Conclusions

We mainly tested two different approaches at GeoCLEF 2006, namely the relatively simple augmentation of geographic terms in the topics, through the use of a geographic ontology, and a text mining approach based on extracting geographical references from documents, in order to assign each to a corresponding geographic scope. In the latter approach, relevance ranking was based on a linear combination of the BM25 text weighting scheme with a similarity function for scopes. In both cases, the obtained results were of acceptable quality, although somewhat inferior to our expectations. Particularly, the text mining approach failed in providing better results than the augmentation method. This point requires more investigation, and we are already making additional experiments with the relevance judgments for the GeoCLEF 2006 topics.

References

[1] H. Alani, C. B. Jones, and D. Tudhope. Associative and spatial relationships in thesaurus-based retrieval. In *Proceedings of ECDL-00, the 4th European Conference on Digital Libraries*, 2000.

[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 International Conference on Management of Data*, 1990.

[3] N. Cardoso, B. Martins, M. Chaves, L. Andrade, and M. J. Silva. The XLDB group at GeoCLEF 2005. In *Working Notes for the CLEF 2005 Workshop*, 2005.

[4] N. Cardoso, M. J. Silva, and B. Martins. The university of lisbon at CLEF 2006 Ad-Hoc task. In *Working notes for the CLEF 2006 workshop*, 2006.

[5] M. Chaves, B. Martins, and M. J. Silva. GKB - Geographic Knowledge Base. DI/FCUL TR 05–12, Department of Informatics, University of Lisbon, July 2005.

[6] E. N. Efthimiadis. A user-centred evaluation of ranking algorithms for interactive query expansion. In *Proceedings of SIGIR-93, the 16th Conference on Research and Development in Information Retrieval*, 1993.

[7] M. J. Egenhofer and D. M. Mark. Naive geography. In *Proceedings of COSIT-95, the 1st Conference on Spatial Information Theory*, 1995.

[8] M. Gutiérrez and A. Rodríguez. Querying heterogeneous spatial databases: Combining an ontology with similarity functions. In *Proceedings of the ER Workshop on Conceptual Modeling of GIS*, 2004.

[9] C. B. Jones, H. Alani, and D. Tudhope. Geographical information retrieval with ontologies of place. In *Proceedings of COSIT-01, the 7th Conference on Spatial Information Theory*, 2001.

[10] Y. Li, A. Moffat, N. Stokes, and L. Cavedon. Exploring probabilistic toponym resolution for geographical information retrieval. In *Proceedings of GIR-06, the 3rd Workshop on Geographical Information Retrieval*, 2006.

[11] D. Lin. An information-theoretic definition of similarity. In *Proceedings of ICML-98, the 15th International Conference on Machine Learning*, 1998.

[12] B. Martins and M. J. Silva. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, 2005.

[13] B. Martins, M. J. Silva, S. Freitas, and A. P. Afonso. Handling locations in search engine queries. In *Proceedings of GIR-06, the 3rd Workshop on Geographical Information Retrieval*, 2006.

[14] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of CIKM-04, the 13th international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM Press.

[15] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC-3. In *Proceedings of TREC-3, the 3rd Text REtrieval Conference*, pages 21–30, 1992.

[16] A. Rodríguez and M. Egenhofer. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *International Journal of Geographic Information Science*, 18(3), 2004.

[17] T. K. Sang, E. F., and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning*, 2003.

[18] R. Song, Ji-RongWen, S. Shi, G. Xin, Tie-YanLiu, T. Qin, J. Z. Xin Zheng, G. Xue, and W.-Y. Ma. Microsoft research asia at the Web track and TeraByte track of TREC 2004. In *Proceedings of TREC-04, the 13th Text REtrieval Conference*, 2004.