Using Geographic Signatures as Query and Document Scopes in Geographic IR

Nuno Cardoso, David Cruz, Marcirio Chaves and Mário J. Silva

Faculty of Sciences, University of Lisbon, LASIGE
{ncardoso,dcruz,mchaves,mjs}@xldb.di.fc.ul.pt

Abstract. This paper reports the participation of the University of Lisbon at the 2007 GeoCLEF task. We adopted a novel approach for GIR, focused on handling geographic features and feature types on both queries and documents, generating signatures with multiple geographic concepts as a scope of interest. We experimented new query expansion and text mining strategies, relevance feedback approaches and ranking metrics.

1 Introduction

This paper presents the participation of the XLDB Group from the University of Lisbon at the 2007 GeoCLEF task. We experimented with novel strategies for geographic query expansion, text mining, relevance feedback and ranking metrics in a renewed GIR system. The motivation for this work derived from the results obtained in last year's participation, which revealed limitations on our previous GIR model [1].

First, our former GIR models focused on capturing and handling geonames and associated features for geographic reasoning, but ignored other terms with important geographic connotation, such as spatial relationships (e.g. in, near, on the shores of) and feature types (e.g. cities, mountains, airports). These terms may play an important role on the definition of the geographic relevance criteria of queries, and on the recognition of geonames in documents. At least, in the GeoCLEF 2007 topics, 13 out of the 25 topics of the Portuguese subtask contained feature types on the topic's title. So, for GeoCLEF 2007 we rebuilt the query processing modules so that all geographic information present on a query is captured, giving special attention to feature types and spatial relationships, as guides for the geographic query expansion [2].

Second, we rely on text mining methods to capture and disambiguate geonames extracted from the text, so that geographic scopes can be inferred for each document [3]. These methods involve geoname grounding into geographic concepts included in a geographic ontology, and disambiguation of hard cases through reasoning based on surrounding geonames also extracted from the text [1,4].

In CLEF 2006, we used a graph-ranking algorithm to analyse the captured features and assign one single feature as the scope of each document [5]. However, this proved to be too restrictive in some cases (other partial geographic contexts of the document were ignored), and also too brittle (incorrectly assigned scopes often lead to poor results). For example, too generic scopes were assigned to documents with geonames that do not correspond to adjacent areas: a document describing a football match between Portugal

and Hungary would have the common ancestor node (Europe) as a very strong candidate for final scope.

We therefore introduced a more comprehensive way to represent query and document scopes, generating geographic signatures for each document (D_{Sig}) and query (Q_{Sig}). A geographic signature is a list of geographic concepts that characterize a document or a query, allowing them to have several geographic contexts. The D_{Sig} is generated for each document by a text mining module, while the Q_{Sig} is generated through a geographic query expansion module. As a consequence of this novel geographic signature focused approach, the geographic ranking step now has the burden of evaluating relevance considering queries and documents with multiple geographic concepts as their scope, which required the development of new combination metrics for computing geographic relevance. In contrast, the similarity metric used last year only compared the (single) geographic concept as the scope of a document against the (single) geographic concept as the scope of a query.

The rest of this paper is organised as follows: Section 2 depicts our assembled GIR system, and describes in detail each module. Section 3 presents our experiments and analyses the results, and Section 4 ends with some conclusions and discussion topics.

2 System Description

Figure 1 presents the architecture of the GIR system assembled for GeoCLEF 2007, which has been presented in [6]. The GeoCLEF topics are automatically parsed by QueOnde and converted into <*what, spatial relationship, where*> triplets. The QuerCol module performs term and geographic query expansion, producing query strings consisting of query terms and a query geographic signature (Q_{Sig}). CLEF documents are loaded into a repository, becoming available to all modules. Faísca is a text mining module specially crafted to extract and disambiguate geonames, generating geographic signatures for each document (D_{Sig}). Sidra5 is the index and ranking module that generates text indexes from the documents and geographic indexes from their geographic signatures. Sidra5 also receives the queries generated by QuerCol as input, and generates final GeoCLEF runs in the `trec_eval` format.

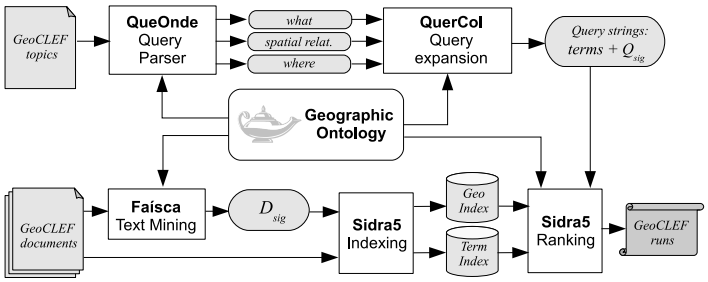


Fig. 1. Architecture of the GIR system assembled for GeoCLEF 2007.

2.1 Geographic Ontology

All modules rely on a geographic ontology for geographic reasoning, created using our own geographic knowledge base, GKB [7]. The GKB 2.0 model now supports relationships between feature types, a better property assignment for features and feature types, and a better control of information sources [8]. Most of the ontology enrichment was carried out in the physical domain, with the addition of new feature types like airports, circuits and mountains, along with their instances.

For the purpose of our participation in GeoCLEF 2007, we made two improvements in the ontology: i) update of the GKB conceptual model to directly support multilingual names for geographic references, and ii) the addition of new features that we found missing after inspecting the GeoCLEF topics.

2.2 Query Parser and Query Expansion

We developed QueOnde, a new geographic query parsing module, which automatically converts query strings into <*what, spatial relationship, where*> triplets with the help of the geographic ontology and a set of manually-crafted context rules. These are used for capturing and disambiguating spatial relationships, features and feature types. For GeoCLEF, we consider the topic titles as query strings.

QuerCol is a geographic query expansion module, introduced in last year's participation [1,9]. QuerCol expands the thematic (*what*) and the geographic (*where*) parts of a query separately. The *what* is expanded through blind relevance feedback [10], while the *where* expansion is based on the available ontological information for the captured geographic concepts.

For GeoCLEF 2007, QuerCol was improved to handle feature types and spatial relationships, and to choose the appropriate geographic expansion strategy based on the features and feature types present in a query [2]. To better illustrate the reasoning task assigned to QuerCol, note that, when feature types are given in a query, they may mean two things: i) the user is disambiguating the geoname, because it can be associated to other geographic concepts (e.g., *City of Budapest* and *Budapest Airport*); or ii) the user is designating a set of concepts as a scope of interest (e.g., *Airports of Hungary*). In i), the feature type is disambiguating the geographic concept given by the feature *Budapest* as the scope of interest, while in ii), the feature type is designating a group of geographic concepts of the scopes of interest. QuerCol will choose the correct interpretation, and perform additional geographic reasoning to obtain the corresponding geographic concepts of the scope.

We now present a complete example of QueOnde and QuerCol integration to produce the Q_{Sig}: consider the following example taken from the GeoCLEF topic #74, *Ship traffic around Portuguese islands*: QuerOnde splits the topic title as a triplet, with *Ship traffic* as the thematic part, *in* as the spatial relationship, and *Portuguese islands* as the geographic part, sub-divided into *Portugal* as a grounded geoname and mapped into the corresponding ontological concept, and *islands* as a feature type. Given this query type, QueOnde therefore reasons that the scope of interest contains all geographic concepts of type *island* that have a *part-of* relationship with geographic concept *Portugal*. In the

end, the Q_{Sig} is composed by the geographic concepts *São Miguel, Santa Maria, Formigas, Terceira, Graciosa, São Jorge, Pico, Faial, Flores, Corvo, Madeira, Porto Santo, Desertas* and *Selvagens*.

2.3 Faísca

The text mining module Faísca parses the documents for geonames, generating the D_{Sig}. Faísca relies on a gazetteer of *text patterns* generated from the geographic ontology, containing all concepts represented by their feature name and respective feature types. The text patterns are in [<*feature type*> $ <*feature name*>] and [<*feature name*> <*feature type*>] format (the former being more common in Portuguese texts, and the latter on English texts). Each pattern is assigned to a single *identifier* of the corresponding geographic concept in the ontology.[1] This immediately captures and grounds all geonames into their unique concept identifiers, without depending on hard-coded disambiguation rules. In the end, we have a *catch-all* pattern, which is used when the geoname found in the document does not contain any kind of external hints on its feature type. For these cases, we assign all identifiers of geographic concepts having that geoname.

The D_{Sig} generated by Faísca consists of a list of geographic concept identifiers and a corresponding *confidence measure* ($ConfMeas$) normalized to [0,1], representing the confidence on the feature being part of the document scope. $ConfMeas$ is obtained through an analysis of the surrounding concepts on each case, in a similar way as described by Li et al. [11]. Geonames on a text are considered as qualifying expressions of a geographic concept when a direct ontology relationship between the geonames is also observed. For example, the geoname *Adelaide* receives an higher $ConfMeas$ value on the document signature if an ontologically related concept, such as *Australia*, is nearby on the text. If so, the feature *Australia* is not included in the D_{Sig}, because it is assumed that it was used to disambiguate *Adelaide*, the more specific concept. Below is an example of D_{Sig} for document `LA072694-001`:

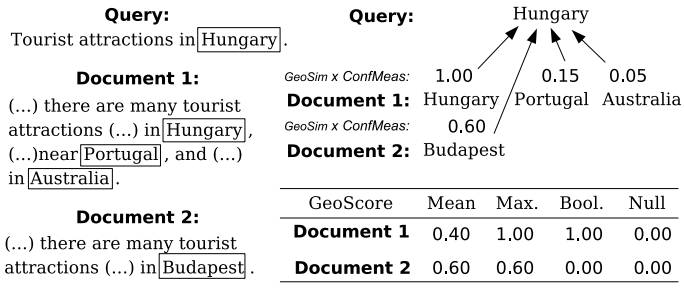
```
LA072694-0011: 5668[1.00]; 2230[0.33]; 4555[0.33]; 4556[0.33];
```

2.4 Sidra5

Sidra5 is a text indexing and ranking module with geographic capabilities based on MG4J [12]. It uses a standard inverted term index provided by MG4J, and a geographic forward index of [*docid*, D_{Sig}] that maps the id of a document to the corresponding D_{Sig}. Sidra5 first uses the *what* part of the query on the term index to retrieve the top 1000 documents. Afterwards, it retrieves the D_{Sig} of each document with the help of the geographic index. The document score is obtained by combining the Okapi BM25 *text score* [13], normalized to [0,1] ($NormBM25$) as defined by Song et al. [14], and a *geographic score* normalized to [0,1] ($GeoScore$) with equal weights:

$$Ranking(query,doc) = \mathbf{0.5} \times NormBM25(query,doc) + \mathbf{0.5} \times GeoScore(query,doc) \quad (1)$$

[1] The character $ means that an arbitrary term or group of terms is allowed to be present between the feature and the feature type, in order to avoid different stopword and adjective patterns.



GeoScore	Mean	Max.	Bool.	Null
Document 1	0.40	1.00	1.00	0.00
Document 2	0.60	0.60	0.00	0.00

Fig. 2. Example of the computation of the four *GeoScore* combination metrics.

The calculation of *GeoScore* begins with the computation of the geographic similarity *GeoSim* for each pair (s_1, s_2), where s_1 *in* Q_{Sig} and s_2 *in* D_{Sig}, through a weighted sum of four heuristic measures (discussed in our 2006 GeoCLEF participation [1]): Ontology (*OntSim*), Distance (*DistSim*), Adjacency (*AdjSim*) and Population (*PopSim*).

$$\begin{aligned} GeoSim(s_1,s_2) = &\mathbf{0.5} \times OntSim(s_1,s_2) + \mathbf{0.2} \times DistSim(s_1,s_2) + \\ &\mathbf{0.2} \times PopSim(s_1,s_2) + \mathbf{0.1} \times AdjSim(s_1,s_2) \end{aligned} \quad (2)$$

Having geographic signatures with multiple geographic concepts requires using aggregation metrics to calculate *GeoScore* from the different *GeoSim* values that a (*query*, *doc*) pair can generate. We experimented four metrics: Maximum, Mean, Boolean and Null.

$GeoScore_{Maximum}$ is the maximum *GeoSim* value computed for a (*query*, *doc*) pair.

$$GeoScore_{Maximum}(query,doc) = max\left(GeoSim(s_1,s_2) \times ConfMeas(s_2)\right), s_1 \in Q_{sig} \wedge s_2 \in D_{sig}$$

$GeoScore_{Mean}$ is the average *GeoSim* values computed for a (*query*, *doc*) pair.

$$GeoScore_{Mean}(query,doc) = avg\left(GeoSim(s_1,s_2) \times ConfMeas(s_2)\right), s_1 \in Q_{sig} \wedge s_2 \in D_{sig}$$

$GeoScore_{Boolean}$ equals 1 if there is a common concept in a (*query*, *doc*) pair, and equals 0 otherwise.

$GeoScore_{Null}$ is always 0, turning off the geographic scores. This is used as a baseline metric for comparing results obtained with the other metrics.

The computation of the four *GeoScore* metrics is illustrated in Figure 2, which presents a fictional query and two document surrogates, along with the $GeoSim \times ConfMeas$ values and final *GeoScore* values.

3 Experiments and Results

Our experiments aimed at:

1. evaluating if this novel approach obtains better results than treating geonames as terms in a standard IR approach;

2. determining which *GeoScore* combination metrics is best.
3. measuring the importance of the geographic query expansion before or after the relevance feedback step.

All runs were generated in the following way: first, the topic titles are used for an initial retrieval, generating *initial runs*. The results of the initial runs are then used for query expansion through blind relevance feedback, generating final queries. These final queries are then used for a final retrieval, generating the *final runs*. More details on the run generation setup can be found in [6]. The generated runs represent three main experiments:

1. The *Terms only* experiment, that uses the names of the Q_{Sig} geographic concepts as standard terms in the generation of the initial and final runs. This means that this experiment uses only classical text retrieval. Nonetheless, the Q_{Sig} were generated by QuerCol through geographic query expansion.
2. The *Geo.QE* experiments, that uses text and geographic scores as described in Section 2.4. This experiment has two types of runs: *Geo. QE before RF*, where the fully expanded Q_{Sig} is used for the generation of the initial run and final run, and the *Geo. QE after RF*, that uses only the fully expanded Q_{Sig} on the generation of the final run; the initial run uses only the geographic concepts found on the initial query as the Q_{Sig} for the generation of the initial run.
3. The *Terms/GIR* experiment, that uses the initial run generated by the *Terms only* experiment to base the relevance feedback step, and afterwards uses the fully expanded Q_{Sig} for the generation of the final run, in the same way as the *Geo.QE* experiments generate their final runs after the relevance feedback step.

The results of our experiments are described on Table 1. We obtained significantly better results for the initial run by using geonames as terms instead of the respective geographic concepts (0.210 versus 0.126), which shows that this is an important result for the final results. The fact that the initial and final run of the *Terms Only* experiment was consistently better than the *Geo.QE* experiments, suggesting us to bootstrap a *Geo.QE* experiment with the initial run fom the *Terms Only*, producing the *Terms/GIR* experiments. In the end, it obtained the highest MAP value from all our experiments (0.268 for the $GeoScore_{Boolean}$ metric).

Regarding the combination metrics, the $GeoScore_{Mean}$ produces poor MAP values because long document signatures tend to cause query drifting. $GeoScore_{Maximum}$ and $GeoScore_{Boolean}$ revealed to be much more robust, and the $GeoScore_{Boolean}$ metric has the best MAP values for Portuguese. This is explained in part because the $GeoScore_{Maximum}$ is highly dependent on the heuristics used, and these are dependent on the quality of the ontology, while the $GeoScore_{Boolean}$ metric is more straightforward on assigning maximum scores for geographically relevant documents.

We also noticed that using fully expanded Q_{Sig} produces better initial runs (0.126 versus 0.084 for Portuguese). This shows that the query signatures produced by QuerCol contribute to more relevant documents on the top of the retrieval results, which is helpful for the blind relevance feedback step. Yet, we did not observe this on the English subtask, prompting us to to do further analysis to understand the reasons for this observation.

Table 1. MAP results obtained for the experiments.

	GeoScore	Terms only	Geo.QE before RF	Geo.QE after RF	Terms/GIR
Initial run		**0.210**	0.126	0.084	**0.210**
Final Run	Maximum	**0.233**	0.122	0.104	0.205
	Mean		0.022	0.021	0.048
	Boolean		**0.135**	**0.125**	**0.268**
	Null		0.115	0.093	0.221

a) Results for the Portuguese monolingual subtask.

Initial run		**0.175**	0.086	0.089	**0.175**
Final Run	Maximum	**0.166**	0.093	0.104	**0.218**
	Mean		0.043	0.044	0.044
	Boolean		**0.131**	**0.135**	0.204
	Null		0.081	0.087	0.208

b) Results for the English monolingual subtask.

4 Conclusions and Discussion

We tested a novel approach for GIR and evaluated its merits against standard IR approaches. We finally outperformed the standard IR approach, albeit in an unexpected way: the best experiment setup is to generate an initial run with classic text retrieval, and use the full geographic ranking modules for the generation of the final run. These results show that there are more efficient ways to introduce geographic reasoning on an IR system, and shed some light on what may be the main problem of many GIR approaches that fail to outperform standard IR approaches.

One should question if the full segregation of the thematic part and the geographic part, from query processing to document ranking, is really the best approach. In fact, as far as we know, there is no published work about a thorough evaluation on the effect of such segregation, claiming that this procedure clearly benefits GIR. A more detailed analysis showed that some terms added by relevance feedback were in fact geonames, and we noticed that geonames may also be good terms for standard IR.

An analysis for each topic reveals that our GIR system is very dependent on the quality of the geographic ontology, and has some limitations in the text mining step. For instance, 25% of all relevant documents (and, as such, with enough geographic evidence to define its scope) had an empty D_{Sig}. Also, we found that most geographic concepts found on the retrieved documents were not relevant for the document scope, or were not in the context of the topic. We also evaluated the results by query type, as the geographic query expansion shifted its strategy according to the spatial relationships, features and feature types found on the queries. We did not observe significative differences on the MAP values by query type.

As future work, we should revise our QR approach and use all query terms for the thematic and geographic expansion steps. The text mining module should also be improved to recognize more geonames and other named entities with a strong geographic connotation (e.g., monuments), and to better detect the roles of each geoname and its contribution for the scope of the document. In conclusion, a longer D_{Sig} does not imply a better D_{Sig}.

Acknowledgements

We thank Joana Campos for improving Faísca for the GIR prototype, Catarina Rodrigues for managing the geographic data, and Diana Santos for relevant suggestions. This work was jointly funded by the European Union (FEDER and FSE) and the Portuguese government, under contracts POSI/ISFL/13/408 (FIRMS-FCT) and POSC/339/1.3/C/NAC (Linguateca), and supported by grants SFRH/BD/29817/2006, POSI/SRI/47071/2002 (GREASE) and PTDC/EIA/73614/2006 (GREASE II) from FCT, co-financed by POSI.

References

1. Martins, B., Cardoso, N., Chaves, M., Andrade, L., Silva, M.J.: The University of Lisbon at GeoCLEF 2006. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Accessing Multilingual Information Repositories: 7th Workshop of the Cross-Language Evaluation Forum (CLEF'2006). Revised Selected Papers. Volume 4730 of LNCS., Springer (2007) 986–994
2. Cardoso, N., Silva, M.J.: Query Expansion through Geographical Feature Types. In: 4th Workshop on Geographic Information Retrieval (GIR'2007), Lisbon, Portugal, ACM (2007)
3. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N.: Adding Geographic Scopes to Web Resources. CEUS - Computers, Environment and Urban Systems **30** (2006) 378–399
4. Cardoso, N., Martins, B., Andrade, L., Chaves, M.S., Silva, M.J.: The XLDB Group at GeoCLEF 2005. In Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., de Rijke, M., eds.: Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum (CLEF'2005). Revised Selected Papers. Volume 4022 of LNCS., Springer (2006) 997–1006
5. Martins, B., Silva, M.J.: A Graph-Based Ranking Algorithm for Geo-referencing Documents. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05), Houston, Texas, USA (2005)
6. Cardoso, N., Cruz, D., Chaves, M., Silva, M.J.: The University of Lisbon at GeoCLEF 2007. In Peters, C., et al., eds.: Working Notes of CLEF'2007, Budapest, Hungary (2007)
7. Chaves, M.S., Silva, M.J., Martins, B.: A Geographic Knowledge Base for Semantic Web Applications. In Heuser, C.A., ed.: Proceedings of the 20th Brazilian Symposium on Databases, Uberlândia, Minas Gerais, Brazil (2005) 40–54
8. Chaves, M.S., Rodrigues, C., Silva, M.J.: Data Model for Geographic Ontologies Generation. In Ramalho, J.C., Lopes, J.C., Carriço, L., eds.: XML: Aplicações e Tecnologias Associadas (XATA'2007), Lisbon, Portugal (2007) 47–58
9. Cardoso, N., Silva, M.J., Martins, B.: The University of Lisbon at CLEF 2006 Ad-Hoc Task. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Accessing Multilingual Information Repositories: 7th Workshop of the Cross-Language Evaluation Forum (CLEF'2006). Revised Selected Papers. Volume 4730 of LNCS., Springer (2007) 51–56
10. Rocchio Jr., J.J.: Relevance Feedback in Information Retrieval. In Salton, G., ed.: The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, USA (1971) 313–323
11. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In: 3rd Workshop on Geographical Information Retrieval, (GIR'06), Seattle, Washington, USA (2006)
12. Boldi, P., Vigna, S.: MG4J at TREC 2005. In: Proceedings of the 14th Text REtrieval Conference (TREC'2005), NIST SP 500-266 (2005) `http://mg4j.dsi.unimi.it`.
13. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference (TREC-3). (1992) 21–30
14. Song, R., Ji-RongWen, Shi, S., Xin, G., Tie-YanLiu, Qin, T., Xin Zheng, J.Z., Xue, G., Ma, W.Y.: Microsoft Research Asia at the Web Track and TeraByte Track of TREC 2004. In: Proceedings of the 13th Text REtrieval Conference (TREC'2004). (2004)