University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval

Claudia Hauff, Dolf Trieschnigg, and Henning Rode

University of Twente
Human Media Interaction Group
Enschede, The Netherlands
{hauffc,trieschn,rodeh}@ewi.utwente.nl

Abstract. This paper describes the approach of the University of Twente at its first participation in GeoCLEF. A large effort went into the construction of a geographic thesaurus which was utilized to add geographic knowledge to the documents and queries. Geographic filtering was applied to the results returned from a retrieval by content run. Employing such a geographic knowledge base however showed no added value - the content-only baseline outperformed all geographically filtered runs.

1 Introduction

During GeoCLEF's pilot track in 2005 became clear that incorporating geographical knowledge into an information retrieval (IR) system is not always beneficial. The best performance of GeoCLEF 2005 was achieved using standard keyword search techniques [1,2,3]. Only Metacarta [4]'s approach using geographic bounding boxes did outperform their keyword-only approach. However, the resulting mean average precision did not exceed 17% and was far from the best submissions (36% and 39% respectively).

Despite the disappointing results of those efforts to incorporate some spatial awareness in IR systems, we believe that adding knowledge about locality can improve search performance. In our CLEF submission of this year, we have confined ourselves to the monolingual task and have only worked with the English queries and documents.

Our approach can be summarized as follows:

1. Carry out document retrieval to find "topically relevant" documents. For example, for the topic "Car bombings near Madrid" this step should result in a ranked list of documents discussing "car bombings", not necessarily near Madrid.
2. Filter this ranked list based on "geographical relevance". For each topically relevant document, determine if it is also geographically relevant. If not, it is removed from the list.

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 958–961, 2007.

The outline of this paper is as follows. In Section 2 the geographic gazetteer (a thesaurus like resource) we created is discussed. The following section describes the process of tagging the document collection and the queries (Section 3) with geographical information. Section 4 describes the document retrieval and filtering processes. Finally, in Section 5 the experiment results are listed followed by a discussion and conclusion in Section 6.

2 The Gazetteer

The gazetteer we used lists geographical references (strings of text) and links them to geographical locations, defined through longitude and latitude values. It also provides information about parent-child relationships between those references. A parent is defined as a region or a country, hence information such as "Madrid lies in Spain which is part of Europe" can be found in the gazetteer.

Our gazetteer was built up from freely available resources. To achieve world coverage, the following gazetteers were combined:

- GEOnet Names Server (http://earth-info.nga.mil/gns/html/),
- the Geographics Names Information System (http://geonames.usgs.gov/stategaz/),
- the World Gazetteer (http://www.world-gazetteer.com/).

An inference mechanism was employed to maximize the amount of parent-child information. Missing parent information of an entry was inferred from agreeing parent information of nearby location entries. The coverage of the merged gazetteer though is not uniform: whereas the USA and Western Europe are well represented, other regions - such as Canada, Northern Africa and a large part of Asia - are barely covered. Figure 1 shows the coverage of the gazetteer. Grid regions with few gazetteer entries are green (light), while red (darker) areas are densely covered.

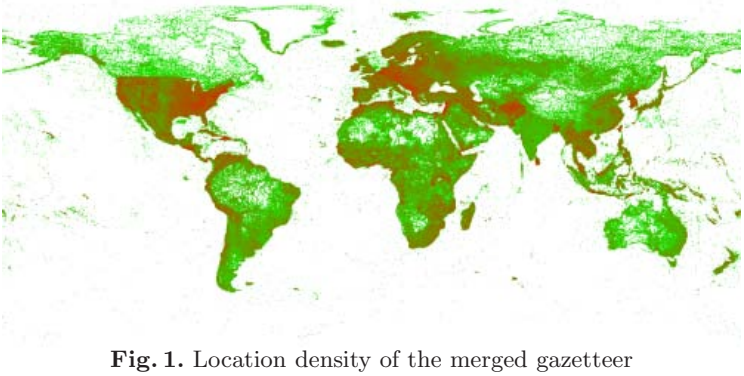

Fig. 1. Location density of the merged gazetteer

3 Corpus and Query Processing

In the preprocessing stage, the geographical range of each document is determined in a two-step process. First, potential location phrases in documents are identified by searching for the longest phrases of capitalized letter strings in each sentence. One additional rule is applied: if two found phrases are only separated by 'of', these phrases are treated as one (for example "Statue of Liberty"). In the second step, the list of potential locations found is matched against the geographic references in the gazetteer.

Contrary to last year, the GeoCLEF topics were not provided with geographic tags. We processed the topics' title section manually and tagged the location names and the type of spatial relation (around, north, east, etcetera). Given a query, it's potential locations were extracted. For a tagged query this is merely the text between the location tags. For an untagged query all capitalized letter phrases were treated as location candidates and matched against the gazetteer. If the gazetteer did not contain any candidate, the extraced phrases were used as Wikipedia queries and the returned wiki page was geotagged the same way as the corpus documents. Additionally, if the found location was a country, its boundaries (minimum and maximum latitude/longitude pairs in the gazetteer) were applied as location coordinate restrictions.

4 Document Retrieval and Filtering

The document collection was indexed with the Lemur Toolkit for Language Modeling and for retrieval purposes the language modelling approach with Jelineck-Mercer smoothing was chosen. Given a query the ranked list of results of a retrieval by content run is returned and subsequently filtered to remove documents outside the desired geographical scope. The locations found in the document are matched against the coordinate restrictions obtained from the query. A document is removed from the result list, if it does not contain any locations which fulfil the query coordinate restrictions.

For queries without coordinate restrictions, the sets of query and document locations are split into parents sets Q_p (query parents) and D_p (document parents). The children sets Q_c (query children) and D_c (document children) are the location names that appear in the gazetteer but not as a parent. In order to determine geographical relevance the intersection sets $I_p = Q_p \cap D_p$ and $I_c = Q_c \cap D_c$ are evaluated. If $Q_x \neq \emptyset$ with $x = \{p, c\}$, then $I_x \neq \emptyset$ must hold in order for the document to be geographically relevant.

5 Experiments and Results

We tested different variations of the usage of title, description and narrative as well as merging the filtered results with the content-only ranking by adding the top filtered-out results at the end of the ranking. The results are given in Table 1. The baseline run in each case is the content-only run.

Table 1. Results (Mean Average Precision) for the English task of GeoCLEF

run id	title	desc.	narr.	geo	merged	map
baseline	x					17.45%
utGeoTIB	x			x		16.23%
utGeoTIBm	x			x	x	17.18%
baseline	x	x				15.24%
utGeoTdIB	x	x		x		7.32%
baseline	x	x	x			18.75%
utGeoTdnIB	x	x	x	x		11.34%
utGeoTdnIBm	x	x	x	x	x	16.77%

6 Discussion and Conclusion

The results show no improvement over the baseline by the addition of geographical knowledge; on the contrary, the performance degrades significantly when including the description or narrative of a topic in a query. A manual evaluation of the relevant documents of the first eight GeoCLEF 2006 topics revealed, that the exact location phrases (e.g. *in the northern part of Iraq*) mentioned in the title query also occur in almost all relevant documents. This makes a geographically enhanced approach unnecessary and also explains the similar results between the baseline and the geographically filtered results for the title queries.

The performance drop of the description and narrative queries is suspected to be due to the fact that many queries contain a long list of possible location terms within them. For retrieval purposes, the location terms within the queries are treated like every other query keyword. However, they are different in the sense that their term frequency within the documents is of no importance; mentioning the location term once within the document already determines the location.

In conclusion, we are still unable to provide conclusive evidence for or against the usage of a geographical knowledge base in the ad hoc information retrieval task. In future work we will evaluate the quality of our geotagging process and its influence on retrieval performance. Furthermore probabilistic geotagging and retrieval models will be investigated.

References

1. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: Geoclef: the clef 2005 cross-language geographic information retrieval track overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
3. Guillén, R.: CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks. In: Working Notes for the CLEF 2005 Workshop (2005)
4. Kornai, A.: MetaCarta at GeoCLEF 2005. In: GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview (2005)