Explicit Query Diversification for Geographical Information Retrieval

Davide Buscaldi, Paolo Rosso

Explicit Query Diversification for Geographical Information Retrieval

Davide Buscaldi[1] and Paolo Rosso[2]

[1] LIFO, Laboratoire d'Informatique Fondamentale d'Orléans,
Université d'Orléans, 45100 Orléans, France
davide.buscaldi@univ-orleans.fr
[2] Natural Language Engineering Lab,
ELiRF Research Group,
Dpto. de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, 46022 Valencia, Spain
prosso@dsic.upv.es

Abstract. In this paper we make a first attempt to evaluate the potential of diversity in the Geographical Information Retrieval task. This task represent an opportunity to take advantage of diversity, given that documents are not relevant only from a thematic point of view, but also spatially. A user of a GIR system may be interested in results that are geographically distributed and equally relevant. We attempted to diversify results explicitly, reformulating queries with meronyms of the places contained in the original queries, with the help of a geographical ontology. The obtained results show that a theoretical improvement is possible, but this approach may be effective only in the case that the relevant documents do not contain enough geographical data.

1 Introduction

Diversity search is an Information Retrieval (IR) paradigm that is somehow opposed to the classic IR vision of "similarity search", in which documents are ranked accordingly to their similarity to the query. In the case of diversity dearch, similarity to the query is not the only criterion to determine relevant results: they should be different one from each other under some aspect, in order to satisfy the user information needs from different points of view which may be known to the user or not. For instance, if the user submit an ambiguous query, it is possible that he is not aware of its ambiguity, and the system should return a mixture of documents which may provide a complete picture of all interpretations, allowing the user to take a further step and decide which aspect of the query is more relevant to him. However, ambiguity is not the only source of diversity. Information is often temporally and/or geographically constrained, such that the results of a given query may be diversified in the temporal or spatial dimension, in order to provide the user with a picture of the evolution of a topic in time or giving him an idea of how the topic may be relevant to a specific sub-region of a region described in the query. For instance, the temporal diversification of the query

"Countries of the European Union" may result in a list of documents where each document describes the countries entering into the Union in a specific year (the complete set of retrieved documents show the history of the adhesions); the geographical diversification of the same query may return documents where the perspective is switched to the membership of a single country (the complete set of retrieved documents provide a full coverage, from a geographical point of view, of the topic).

In most of the research works over diversity the objective has been to provide multiple distinct interpretations for ambiguous queries [1,2,3]; less works have dealt with the representation of sub-topics within search results for queries with broad thematic scope [4]. Spatial diversity has been successfully applied to image search in [5]; Tang and Sanderson [6] showed that spatial diversity is appreciated by users. Clough et al. [7] analysed query logs and found that in the case of place names ambiguities users tend to reformulate queries more often.

The objective of this paper is to determine the potential of geographical diversity in the context of Geographical Information Retrieval (GIR). In GIR, queries are geographically constrained: therefore, it is possible, with the help of a geographical ontology, determine the sub-topics directly from the query (for instance: Europe is diversified in all the component countries) and build a set of reformulated queries, one for each subtopic. With the help of GeoWorSE[8], a GIR-enabled search engine, and the evaluation framework (queries and documents) of GeoCLEF[3] we attempted to determine the effects of the diversified sub-queries on the retrieval results.

The remainder of this paper is structured as follows: in Section 2 we describe the retrieval framework, in Section 3 we describe the collection used and the experiments carried out; in Section 4 we present the results of the experiments and an analysis of these results; finally in Section 5 we draw some conclusions and set the path for future works.

2 The GeoWorSE Retrieval System

This system is built around the Lucene search engine and a geographical ontology based on Geonames[4] and WordNet [9]. It is based on the enrichment of the index with terms that are not contained in the examined document but which can be inferred from the geographical entities in the document text.

During the indexing phase, the documents are examined in order to find location names (*toponyms*) by means of the Stanford conditional random fields-based NER system. When a toponym is found, in the case it has more than one referent according to the geographical ontology, a disambiguator [10] determines the correct reference for the toponym, depending on the other toponyms contained in the document. Then, holonyms and synonyms of the toponym are extracted from the ontology and added to an *expanded* index, together with

[3] http://ir.shef.ac.uk/geoclef/
[4] http://www.geonames.org

the original toponym. For instance, consider the following text from document GH950630-000000 in the GeoCLEF collection:

> ...The *British* captain may be seen only once more here, at next month's world championship trials in **Birmingham**, where all athletes must compete to win selection for *Gothenburg*...

Let us suppose that in the ontology there are two possible referents for "Birmingham": "Birmingham/Alabama", and "Birmingham/England". "Gothenburg" is found only once but with synonyms *Goteborg* (the original Swedish name) and the alternate spelling "Goetenborg". Let us suppose that the disambiguator correctly identifies "Birmingham" with the English referent, then its holonyms are *England*, *United Kingdom*, *Europe*. In the case of "Gothenburg" we obtain *Sweden* and *Europe* as holonyms, "Goetenborg" and "Goteborg" as synonyms. Therefore, the words added to the expanded index for the above paragraph are: *Birmingham*, *England*, *United Kingdom*, *Europe*,*Gothenburg*, *Goteborg*, *Goeteborg*, *Sweden*.

Then, the *geo* index contains the geographical coordinates associated to the above toponyms; finally, all document terms are stored in the *text* index. The *text* and expanded indices are used during the search phase; the *geo* index was not used for search in this work. In Figure 1 we show the architecture of the indexing module.

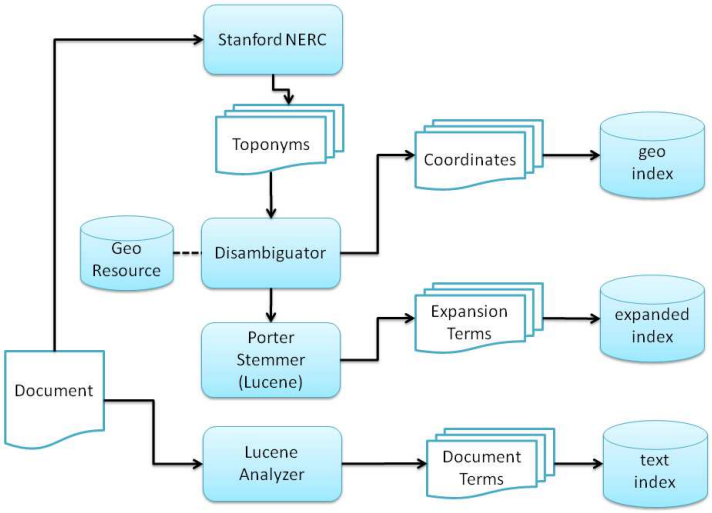

Fig. 1. Diagram of the Indexing module

The topic text is searched by Lucene in the text index. All the toponyms are extracted by the NER system and searched for by Lucene in the *expanded* index The result of the search is a list of documents ranked using the $tf \cdot idf$ weighting scheme, as implemented in Lucene.

2.1 Query Diversification

The query terms $t_0 \dots t_n \in Q$ are grouped into two subsets, a *content* set C_q, containing the words which represent the "focus" or thematic part of the query, and a *geographic footprint* set G_q, which contains the place names P identifying the geographical constraint associated to the query. For every toponym $t \in G_q$, we search the geographical ontology for *meronyms* $M_t = \{m_0, \dots m_k\}$ (i.e., places contained in the place represented by toponym t). The diversified queries are assembled by taking the terms in C_q together with a meronym $m \in M_t$, for every t. Therefore, the number or queries built in such a way is $\sum_{t \in G_q} |M_t|$. For instance, the query "golf tournaments in Europe" would be diversified into: "golf tournaments Spain", "golf tournaments Italy", "golf tournaments UK", "golf tournaments France", etc.

Among all the produced queries, we selected the ν most promising queries as the ones having the highest mutual information (MI) between the content terms and the terms in the geographic footprint:

$$I(C_q; G_q) = p(C_q \cap G_q) \log \frac{p(C_q \cap G_q)}{p(C_q)p(G_q)} \tag{1}$$

Where probabilities are calculated as the number of hits (obtained with the baseline ranking and the indicated set of terms) divided the number of documents in the collection. If there are less than ν possible reformulations, all reformulated queries are taken into account. This selection process has the objective of identifying the relative importance of the geographical aspects underlying the original query.

3 Experimental Setup

Our experiments were conducted over the GeoCLEF 2005-2008 test collection, including a total of 100 topics with the relative relevance judgements. The document collection consists of 169,477 documents and is composed of stories from the British newspaper "The Glasgow Herald", year 1995, and the American newspaper "The Los Angeles Times", year 1994. We run the experiments using only the topic title. Of the 100 topics, it was possible to build reformulation for 45 of them. This means that of the 100 topics, 55 did not include a place name or the included place names did not have meronyms in the geographical ontology (this may happen if the place can be approximated to a point or a line, such as cities or rivers). We used two baselines: the first baseline is constituted by the result obtained with the original query, without reformulation. The second

baseline is made of the merged results of reformulations, using the cmbMNZ fusion algorithm [11].

The potential of diversification was examined using an oracle, that is, a system which returns the results that could be obtained if the best reformulation is known apriori. Since we are still at the beginning of our work on diversity search, we implemented a naïve round robin technique (subsequently indicated as "RR") for the fusion of the results of the reformulated queries, consisting in building a list by taking one document in turn from each individual list and alternating them in order to construct the final merged output. This is how a user would behave while examining different sets of results (examining the top ones from each set, then the second best results, and so on). Duplicate results are removed. In this way, the merged result set can be compared with the ones obtained with the baselines.

The metrics used in the evaluation are: Mean Average Precision (MAP), Mean Relevance Rank (MRR), Precision at 5 (P@5), and Normalized Cumulative Discounted Gain (NDCG). NDCG ability to handle degrees of relevance was not exploited since the relevance judgements in GeoCLEF are binary judgements.

4 Experimental Evaluation

We carried out two evaluations, one with $\nu = 5$ (Table 1) and another with $\nu = 10$ (Table 2). In all measures, the baseline is better than the fused results, while the oracle is always better than the baseline. It is interesting to note that the round-robin technique allowed to obtain better results than CombMNZ with MRR and NDCG (although the difference in NDCG is not statistically relevant). NDCG has been observed in [12] to be the measure that most effectively models user preferences.

Table 1. Results with $\nu = 5$

	base	CombMNZ	RR	Oracle
MAP	0,2074	0,1935	0,1818	0,2543
MRR	0,5301	0,4923	0,5185	0,7131
NDCG	0,4710	0,4605	0,4644	0,5401
P@5	0,3435	0,3087	0,2696	0,4304

Table 2. Results with $\nu = 10$

	base	CombMNZ	RR	Oracle
MAP	0,2074	0,1862	0,1777	0,2612
MRR	0,5301	0,4323	0,4948	0,7512
NDCG	0,4710	0,4555	0,4616	0,5510
P@5	0,3435	0,3217	0,2783	0,4522

We analysed the data and found some queries that obtained always a significant improvement over the baseline with the RR fusion, and others for which the oracle was not able to obtain a result better than the baseline. These "critical" topics are shown in Table 3.

Table 3. "Critical" topics

Topics mostly benefitted by query reformulation (group 1)	
10.2452/GC-001	Shark Attacks off Australia and California
10.2452/GC-006	Oil Accidents and Birds in Europe
10.2452/GC-008	Milk Consumption in Europe
10.2452/80-GC	Politicians in exile in Germany
Topics negatively affected by query reformulation (group 2)	
10.2452/GC-048	Fishing in Newfoundland and Greenland
10.2452/GC-013	Visits of the American president to Germany
10.2452/GC-010	Flooding in Holland and Germany
10.2452/51-GC	Oil and gas extraction found between the UK and the Continent

In order to understand the reason of such behaviour, we examined the distributions of places in the set of relevant documents in order to understand whether geographical diversity is supported by the data contained in the test collection or not. For each query q we carried out a k-means clustering, with $k = \nu$, of the points contained in the set R_q of relevant documents. The desired behaviour was to obtain clusters centred on geographic areas corresponding to the places identified in the query diversification process.

We found that reformulation of queries in group 1 was effective because actually the centroids *did not* match the diversified places, while for queries in group 2, the data showed clusters centred mostly on relevant areas (we plotted these clusters in Figure 2 and Figure 3 for topics $10.2452/GC-006$ and $10.2452/GC-010$, respectively). It can also be observed that many places are distributed accordingly the sources of the news (Glasgow and Los Angeles). Therefore, the diversification of the queries based on the geographical ontology seems to be effective only when the data do not offer enough clues to group results from a geographical viewpoint.

5 Conclusions and Further Work

We developed a simple method to geographically diversify GIR queries, based on the knowledge provided by a geographical ontology. With this method, if the original query contains the name of a region which includes n places, the $\nu \leq n$ most significant places (according to the mutual information between the query content and the geographical constraint) are selected, and ν queries are submitted to the search engine. We evaluated this method over the GeoCLEF test set. The results showed that an oracle selecting always the best result obtains

Fig. 2. Distribution of places in documents judged relevant for topic $10.2452/GC-006$. Cluster centroids indicated with star-shaped markers. Data are sparse and do not reflect the geographic footprint of the query.

Fig. 3. Distribution of places in documents judged relevant for topic $10.2452/GC-010$. Cluster centroids indicated with star-shaped markers. Data mostly reflect the geographic footprint of the query.

better results in all measures than the baseline, indicating that a theoretical improvement is possible; however, the tested fusion methods are not able to capture this potential. The error analysis showed that apriori diversification of the query was useful when the geographical data are sparse, and therefore it is necessary to "drive" the query towards possible relevant results. If the geographical data in the relevant documents are dense enough to support the diversification of results, then diversity can be inferred from data and query reformulation adds noise.

In order to validate these conclusions, we will have to carry out more experiments. We will have to design a data-driven diversification algorithm (or use an existing one, such as the one proposed by [1]) and verify that in this way it is possible to exploit the geographical diversity contained in the data to improve the results in GIR. We should also evaluate the results using metrics specifically aimed to diversity.

References

1. Santos, R.L.T., Peng, J., Macdonald, C., Ounis, I.: Explicit Search Result Diversification through Sub-queries. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S.M., van Rijsbergen, K., eds.: ECIR. Volume 5993 of Lecture Notes in Computer Science., Springer (2010) 87–99
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, New York, NY, USA, ACM (2009) 5–14
3. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '08, New York, NY, USA, ACM (2008) 659–666
4. Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A.: Efficient computation of diverse query results. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Washington, DC, USA, IEEE Computer Society (2008) 228–236
5. Paramita, M.L., Tang, J., Sanderson, M.: Generic and Spatial Approaches to Image Search Results Diversification. In: ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, Springer-Verlag (2009) 603–610
6. Tang, J., Sanderson, M.: Spatial Diversity, Do Users Appreciate It? In: GIR10 Workshop. (2010)
7. Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., Paramita, M.: Multiple Approaches to Analysing Query Diversity. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '09, New York, NY, USA, ACM (2009) 734–735
8. Buscaldi, D., Rosso, P.: Using GeoWordNet for Geographical Information Retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. (2009) 863–866
9. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38**(11) (1995) 39–41
10. Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Systems **22**(3) (2008) 301–313
11. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of the 2nd Text REtrieval Conference (TREC-2). (1994) 243–249
12. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do user preferences and evaluation measures line up? In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '10, New York, NY, USA, ACM (2010) 555–562