Geographic Expansion of Queries to Improve the Geographic Information Retrieval Task

José M. Perea-Ortega[1] and L. Alfonso Ureña-López[2]

[1] Languages and Information Systems Department
University of Sevilla, Spain
jmperea@us.es

[2] Computer Science Department
University of Jaén, Spain
laurena@ujaen.es

Abstract. Geographic Information Retrieval (GIR) is concerned with improving the quality of geographically-specific Information Retrieval (IR), focusing on access to unstructured documents. Since GIR can be considered as an extension of IR, the application of Natural Language Processing (NLP) techniques, such as query expansion, can lead to significant improvements. In this paper we propose two NLP techniques of query expansion related to the augmentation of the geospatial part that is usually identified in a geographic query. The aim of both approaches is to retrieve possible relevant documents that are not retrieved using the original query. Then, we propose to add such new documents to the list of documents retrieved using the original query. In this way, the geo-reranking process takes into account more possible relevant documents. We have evaluated the proposed approaches using GeoCLEF as evaluation framework for GIR systems. The results obtained show that the use of proposed query expansion techniques can be a good strategy to improve the overall performance of a GIR system.

Keywords: Geographic Information Retrieval, Query Expansion, GeoCLEF.

1 Introduction

Natural Language Processing (NLP) techniques, such as query expansion, are an integral part of most Information Retrieval (IR) architectures. Since Geographic Information Retrieval (GIR) can be considered as an extension of IR [11], the application of these techniques in GIR systems can lead to significant improvements. While 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 [2].

G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 94–103, 2012.

Specifically, GIR is concerned with improving the quality of geographically-specific information retrieval, focusing on access to unstructured documents [11,14]. The IR community has primarily been responsible for research in the GIR field, rather than the Geographic Information Science (GIS) community. The type of query in a IR engine is based usually on natural language, in contrast to the more formal approach common in GIS, where specific geo-referenced objects are retrieved from a structured database. In a GIR system, a geographic query can be structured as a triplet of <*theme*><*spatial relationship*><*location*>, where <*theme*> is the main subject of the query, <*location*> represents the geographic scope of the query and <*spatial relationship*> determines the relationship between the subject and the geographic scope. For example, the triplet for the geographic query "*airplane crashes close to Russian cities*" would be <*airplane crashes*><*close to*><*Russian cities*>. Thus, a search for "*castles in Spain*" should return not only documents that contain the word "*castle*", also those documents which have some geographical entity related to Spain. For this reason, it is important to pay attention to finding effective methods for query expansion to improve the quality of the retrieved documents. From an IR point of view, query expansion refers to the process of automatically adding additional terms to the query, in an effort to improve the relevance of the retrieved results. From a GIR point of view, geographic expansion techniques can be used to augment any geographic term identified in queries, thereby increasing the likelihood of finding relevant documents with geographic entities that match the geographic scope identified in the query.

To carry out query expansion for geographic queries, we can take into account both lexical-syntactic features and geographical aspects. In this paper we propose two query expansion techniques based on the addition of synonyms of the geospatial scope identified in the query and, on the other hand, the addition of geographic terms that match with the geospatial scope of the query. Then, we merge the documents retrieved by using the original query with those new documents retrieved by using the proposed query expansions. Finally, we apply a reranking function based on the textual and geographical similarity between each retrieved document and the query. To carry out the evaluation, we have used the most important evaluation framework in this context: GeoCLEF[1] [8,17]. The results show that the proposed query expansion techniques retrieved relevant documents that were not retrieved by using the original query, so that after applying the reranking function, our GIR system was able to improve its overall performance.

The remainder of this paper is structured as follows: in Section 2, the most important works related to the query expansion in GIR are expounded; in Section 3 we describe the GIR system used for the experiments carried out in this work; in Section 4 and Section 5, the evaluation framework is briefly described and the experiments and an analysis of the results are presented, respectively. Finally, in Section 6, some conclusions and future work are expounded.

[1] `http://ir.shef.ac.uk/geoclef/`

2 Related Work

Jansen et al. [10] define the concept of query reformulation as the process of altering a given query in order to improve search or retrieval performance. Sometimes, query reformulation is applied automatically by search engines as with *relevance feedback* technique. It is a method that allows users to judge whether a document is relevant or not, so that automatic rewritings can be generated depending on it. At other times, query reformulation is carried out analysing the top retrieved documents without the user's intervention, taking into account term statistics. However, it has been found that users rarely utilize the relevance feedback options [22] and usually reformulate their needs manually [3].

The focus of this paper is geographic queries. According to Gravano [9], search engines are criticised because of their ignorance to the geographical constraints on users' queries and, therefore, retrieve less relevant results. This could be attributed to the way search engines handle queries in general as they adopt a keywords matching approach without spatially inferring the scope of the geographic terms. However, it shall be noted that a number of services to deal with this issue have recently been proposed in major search engines, but not in the general purpose tools.

Several authors have studied what users are looking for when submitting geographic queries [21,7,12]. One of the main conclusions of these studies is that the structure of geographic queries consists of thematic and geographical parts, with the geo-part occasionally containing spatial or directional terms. From a geographical point of view, Kohler [13] provides a research about georeformulation of queries. She concludes that the addition of more geo terms in the query is commonly used to differentiate between places that share the same name. This is also known as query expansion using geographic entities.

In the literature, we can find various works that have addressed the spatial query expansion. Cardoso et al. [5] present an approach for geographical query expansion based on the use of feature types, readjusting the expansion strategy according to the semantics of the query. Fu et al. [6] propose an ontology-based spatial query expansion method that supports retrieval of documents that are considered to be spatially relevant. They improve search results when a query involves a fuzzy spatial relationship, showing that proposed method works efficiently using realistic ontologies in a distributed spatial search environment. Buscaldi et al. [4] use WordNet[2] during the indexing phase by adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms, proving that such method is effective. Li et al.[15] describe two types of geo-query expansion: *downward expansion* and *upward expansion. Downward expansion* extends the influence of a geo-term to some or all of its descendants in the hierarchical gazetteer structure, to encompass locations that are part of, or subregions of, the location specified in the query. *Upward expansion* extends the influence of a geo-term to some or all of its ancestors, and then possibly downward again into other siblings of the original node. This facilitates

[2] http://wordnet.princeton.edu/

the expansion of geo-terms in the query to their nearby locations. Finally, Stokes et al. [23] conclude that significant gains in GIR will only be made if all query concepts (not just geospatial ones) are expanded.

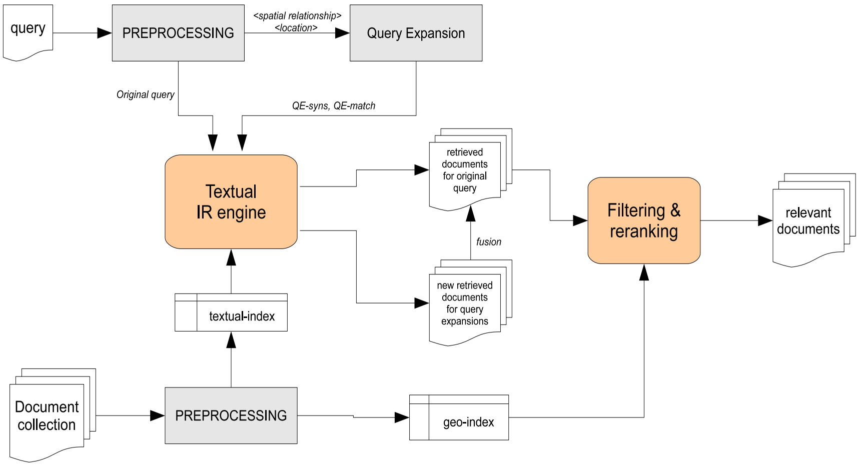

Fig. 1. Overview of the SINAI-GIR system

3 GIR System Overview

In this Section we describe an example of a GIR system. Specifically, we have used our own GIR system called SINAI-GIR [20]. GIR systems are usually composed of three main stages: preprocessing of the document collection and queries, textual-geographical indexing and searching and, finally, reranking of the retrieved results using a particular relevance formula that combines textual and geographical similarity between the query and the retrieved document. The GIR system used in this work follows a similar approach, as can be seen in Figure 1.

On the one hand, each query is preprocessed and analyzed, identifying the geographic scope and the spatial relationship that may contain. On the other hand, the document collection is also preprocessed, detecting all the geographic entities and generating a geo-index with them. In this phase, the stop words are removed and the stem of each word is taken into account. Then, each preprocessed query (including their geographic entities) is run against the search engine.

Regarding query processing, it is mainly based on detecting the geographic entities. We have used Geo-NER[19] to recognize spatial entities in the collection and queries, a Named Entity Recognizer (NER) for geographic entities based on GeoNames[3] and Wikipedia. This phase also involves specifying the triplet explained in Section 1, which will be used later during the filtering and reranking

[3] `http://www.geonames.org`

process. To detect such triplet, we have used a Part Of Speech tagger (POS tagger) like TreeTagger[4], taking into account some lexical syntactic rules such as *preposition + proper noun*, for example. Moreover, the stop words are removed and the Snowball stemmer[5] is applied to each word of the query, except for the geographical entities. During the text retrieval process, we obtain 1,000 documents for each query. We have used Terrier[6] as a search engine. According to a previous work [18], it was shown that Terrier is one of the most used IR tools in IR systems in general and GIR systems in particular, obtaining promising results. The weighting scheme used has been *inL2*, which is implemented by default in Terrier. This scheme is the Inverse Document Frequency (IDF) model for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization [1].

In addition to the original preprocessed query, each query expansion is also launched against the search engine. Then, the new documents retrieved by each expansion are added to the list of documents retrieved for the original query with the lowest Retrieval Status Value (RSV) found in such list. In this way, the reranking process will take into account more possible relevant documents to rerank.

Finally, in the last phase, the fusion list is filtered and reranked, leaving only 1,000 documents to return and making use of the reranking function that combines both similarities textual and geographical between the query and each document. Many GIR systems, for example those of Li et al. [16] and Andrade and Silva [2], combine the scores of textual terms and geographic terms using linear combinations of the form:

$$sim(Q,D) = \alpha \times sim_{text}(Q,D) + (1-\alpha) \times sim_{geo}(Q,D) \tag{1}$$

where $sim_{text}(Q,D)$ is the score assigned by the search engine to each document, i.e. the RSV score. For the experiments carried out in this work, we have tried several values of α, obtaining the best performance with $\alpha = 0.5$. On the other hand, $sim_{geo}(Q,D)$ is the geographic similarity between a document (D) and a query (Q) and it is calculated using the following formula:

$$sim_{geo}(Q,D) = \frac{\sum\limits_{i \in geoEnts(D)} match(i,GS,SR) \cdot freq(i,D)}{|geoEnts(D)|} \tag{2}$$

where the function $match(i,GS,SR)$ returns 1 if the geographic entity i satisfies the geographic scope GS for the spatial relationship SR and 0 otherwise. $freq(i,D)$ means frequency of the geographic entity i in document D, and $|geoEnts(D)|$ represents the total number of geographic entities identified in the document D. To explain the performance of the *match* function, we can

[4] TreeTagger v.3.2 for Linux. Available in `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[5] Available in `http://snowball.tartarus.org`

[6] Version 2.2.1, available in `http://terrier.org`

use the following query: "*Hurricane Katrina in the United States*". In this case, it is a geographic query because we can recognize a geographical scope (*United States*) and a spatial relationship (*in*). The theme or subject of the query would be *Hurricane Katrina*. Therefore, when the system finds a geographic entity i (for example, *New York*) in a retrieved document (D) which belongs to United States, then $match(NewYork, UnitedStates, in) = 1$. If the geographic entity did not belong to the geographic scope (GS), then the *match* function would return 0 (for example, $match(Madrid, UnitedStates, in) = 0$). In short, the $match(i, GS, SR)$ function receives as input the geographic entity i of the document, the geographic scope (GS) of the query and the spatial relationship (SR) identified in the query. This function is based on manual rules such as "*if* $SR = in$ *and* $i \in GS$ *then return 1, else return 0*". Obviously, this function makes use of an external geographical database like GeoNames in order to check if a city belongs to a country or a continent, for example.

4 GeoCLEF: The Evaluation Framework

In order to evaluate the proposed query expansions, we have used the GeoCLEF framework [8,17], an evaluation forum for GIR systems held between 2005 and 2008 under the CLEF[7] conferences. GeoCLEF provides a document collection that consists of 169,477 documents, composed of stories and newswires from the British newspaper *Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994), representing a wide variety of geographical regions and places. On the other hand, there are a total of 100 textual queries or topics provided by GeoCLEF organizers (25 per year). They are composed of three main fields: *title* (T), *description* (D) and *narrative* (N). For the experiments carried out in this work, we have only taken into account the *title* field because it represents in a similar way how a user would launch a geographic query to a search engine. Some examples of GeoCLEF topics are: "*vegetable exporters of Europe*", "*forest fires in north of Portugal*", "*airplane crashes close to Russian cities*" or "*natural disasters in the Western USA*".

Regarding the evaluation measures used, results have been evaluated using the relevance judgements provided by the GeoCLEF organizers and the TREC evaluation method. The evaluation has been accomplished by using the Mean Average Precision (MAP) that computes the average precision over all queries. The average precision is defined as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved.

5 Experiments and Results

As previously mentioned, two types of query expansions for the GIR task are proposed in this work. Both strategies use the geographic scope identified in the

[7] http://www.clef-initiative.eu/

query. The aim of these query expansions is to improve the retrieval process trying to find relevant documents that are not retrieved using the original query. Specifically, we propose the following query expansions:

- QE-syns: the geographic part is expanded using only synonyms of the geographic scope identified in the query.
- QE-match: the geographic part is expanded using locations or places that match with the geographic scope and the spatial relationship identified in the query.

Table 1. Example of query expansions generated for the query "*Visits of the American president to Germany*"

Query Expansion	**Text of the query**
original	`visit American presid Germany`
QE-syns	`#and(visit American presid #or(Germany #3(Federal Republic of Germany) Deutschland FRG ) )`
QE-match	`#and(visit American presid #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen ) )`

Table 1 shows an example of the query expansions generated for the query "*Visits of the American president to Germany*". As can be seen, QE-syns and QE-match expand only the geographical part of them, making use of the synonyms of the geographic scope identified in the query and with places that match with the geospatial scope of the query, respectively. Table 2 shows the results of each query expansion strategy compared with those obtained using original queries without applying any reranking process.

Table 2. Results of each query expansion strategy compared with those obtained using original queries without applying any reranking process

Query set	**MAP orig query baseline *inL2***	**MAP QE-syns baseline *inL2***	**MAP QE-match baseline *inL2***
2005	0.3514	0.2242	0.0952
2006	0.2396	0.2064	0.1811
2007	0.2311	0.1687	0.1874
2008	0.2484	0.1619	0.1906

As has been explained in Section 3, the new documents retrieved using the query expansions proposed (QE-syns and QE-match) are added to the list of documents retrieved using the original query and, therefore, a *fusion* list of documents is generated. This list is reranked taking into account the formula expounded in 1. Table 3 shows the results obtained using this *fusion* list compared with those obtained using the original query. Moreover, it is also shown the total number of relevant documents for each query set (*Total num rel*), the number of relevant documents retrieved (*Num rel ret*) by the original query and the *fusion* list and the MAP score obtained for each experiment.

Table 3. Summary of the experiments and results

Query set	**Total num rel**	**Num rel ret orig query**	**Num rel ret fusion list**	**MAP orig query baseline**	**MAP orig query reranked**	**MAP fusion list reranked**
2005	1028	**908**	904	0.3514	**0.3608**	0.3606
2006	378	284	**291**	0.2396	0.2417	**0.2419**
2007	650	543	**570**	0.2311	0.2448	**0.2464**
2008	747	588	**614**	0.2484	0.2606	**0.2614**

Analyzing these results, we can observe that for three of the four query sets, the proposed query expansions added relevant documents that were not retrieved using the original query. Specifically, for the 2007 and 2008 query sets, the expansion techniques provided 27 and 26 new relevant documents that were not retrieved using the original query, respectively. For the 2006 query set, the *fusion* list provided 7 new relevant documents. Obviously, these results have a positive impact on the calculation of the MAP score. As can be seen in Table 3, the MAP value obtained using the reranked *fusion* list improved 2.62%, 0.96%, 6.62% and 5.23% the MAP score obtained using the original query without applying the reranking process for the 2005, 2006, 2007 and 2008 query sets, respectively. The differences achieved by the reranked *fusion* list were smaller when we applied the reranking process to the list of documents retrieved for the original query solely: +0.08%, +0.65% and +0.31% for the 2006, 2007 and 2008 query sets, respectively.

6 Conclusions and Further Work

In this paper we propose two NLP techniques of query expansion related to the augmentation of the geospatial part that is usually identified in a geographic query. The aim of both approaches is to retrieve possible relevant documents that are not retrieved using the original query. Then, we propose to add such new documents to the list of documents retrieved using the original query. In this way, the geo-reranking process takes into account more possible relevant documents. We have evaluated the proposed approaches using GeoCLEF as evaluation framework for GIR systems. The results obtained show that the use of proposed query expansion techniques can be a good strategy to improve the overall performance of a GIR system.

For future work, we will analyze the different types of geographic queries and then we will study in depth when is more suitable to apply these techniques in a GIR system depending on the type of the query. We will also try to analyze the performance of the expansion of the thematic part of the query, using synonyms of the keywords, for example.

Acknowledgments. This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government, and a grant from the Instituto de Estudios Giennenses through the Geocaching Urbano project (RFC/IEG2010).

References

1. Amati, G.: Probabilistic Models for Information Retrieval based on Divergence from Randomness. Ph.D. thesis, School of Computing Science, University of Glasgow (2003)
2. Andrade, L., Silva, M.J.: Relevance ranking for geographic ir. In: Purves, R., Jones, C. (eds.) GIR. Department of Geography, University of Zurich
3. Anick, P.: Using terminological feedback for web search refinement: a log-based study. In: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 88–95. ACM, New York (2003)
4. Buscaldi, D., Rosso, P., Arnal, E.S.: Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. 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, pp. 939–946. Springer, Heidelberg (2006)
5. Cardoso, N., Silva, M.J.: Query expansion through geographical feature types. In: Purves, R., Jones, C. (eds.) GIR, pp. 55–60. ACM (2007)
6. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-Based Spatial Query Expansion in Information Retrieval. In: Meersman, R., Tari, Z. (eds.) OTM 2005, Part II. LNCS, vol. 3761, pp. 1466–1482. Springer, Heidelberg (2005)
7. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: Proceedings of the First International Workshop on Location and the Web, pp. 49–56. ACM, Beijing (2008)
8. Gey, F.C., Larson, R.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, pp. 908–919. Springer, Heidelberg (2006)
9. Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 325–333 (2003)

10. Jansen, B.J., Booth, D.L., Spink, A.: Patterns of query reformulation during web searching. JASIST 60(7), 1358–1371 (2009)
11. Jones, C.B., Purves, R.S.: Geographical information retrieval. International Journal of Geographical Information Science 22(3), 219–228 (2008)
12. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic intention and modification in web search. International Journal of Geographical Information Science 22(3), 229–246 (2008)
13. Kohler, J.: Analysing search engine queries for the use of geographic terms. Master's thesis, University of Sheffield - United King (2003)
14. Larson, R.: Geographic information retrieval and spatial browsing. In: Smith, Gluck, M. (eds.) Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information, pp. 81–124 (1996)
15. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring probabilistic toponym resolution for geographical information retrieval. In: Purves, R., Jones, C. (eds.) GIR. Department of Geography, University of Zurich (2006)
16. Li, Z., Wang, C., Xie, X., Wang, X., Ma, W.Y.: Indexing implicit locations for geographical information retrieval. In: Purves, R., Jones, C. (eds.) GIR. Department of Geography, University of Zurich (2006)
17. Mandl, T., Carvalho, P., Di Nunzio, G.M., Gey, F., Larson, R.R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 808–821. Springer, Heidelberg (2009)
18. Perea-Ortega, J.M., García-Cumbreras, M.Á., García-Vega, M., Ureña-López, L.A.: Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (eds.) NLDB 2008. LNCS, vol. 5039, pp. 142–147. Springer, Heidelberg (2008)
19. Perea-Ortega, J.M., Martínez-Santiago, F., Montejo-Ráez, A., Ureña-López, L.A.: Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) 43, 33–40 (2009)
20. Perea-Ortega, J.M., Ureña-López, L.A., García-Vega, M., García-Cumbreras, M.A.: Using Query Reformulation and Keywords in the Geographic Information Retrieval Task. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 855–862. Springer, Heidelberg (2009)
21. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings Workshop on Geographical Information Retrieval SIGIR (2004)
22. Spink, A., Jansen, B.J., Ozmultu, C.H.: Use of query reformulation and relevance feedback by excite users. Internet Research: Electronic Networking Applications and Policy 10(4), 317–328 (2000)
23. Stokes, N., Li, Y., Moffat, A., Rong, J.: An empirical study of the effects of nlp components on geographic ir performance. International Journal of Geographical Information Science 22(3), 247–264 (2008)