Ontology-based Query Construction for GeoCLEF

Rui Wang[1] and Gϋnter Neumann[2]

[1] Saarland University
66123 Saarbrücken, Germany
rwang@coli.uni-sb.de
[2] LT-Lab, DFKI
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
neumann@dfki.de

Abstract. This paper describes our participation in GeoCLEF. Being different from the traditional information retrieval, we focus more on the query expansion instead of document ranking. We parse each topic into the event part and the geographic part and use different ontologies to expand both parts respectively. The results show great advantages of our strategy for this task.

Keywords: Geographical Information Retrieval, query expansion, ontology

1 Introduction

The goal of geographic information retrieval (GIR) is to retrieve documents for topics with a geographic specification (Mandl et al., 2007). For example, given the query "*riots in South American prisons*", the system is asked to retrieve all the relevant documents about these *events* (i.e. "*riots*") happening at those *places* (i.e. "*South American prisons*").

Traditional information retrieval consists of three main components: query expansion, document retrieval, and document ranking, of which the last component attracts the most attention (Singhal, 2001). As for GIR, since geographic variation is an important criterion for evaluating such systems, we assume that the query processing will have more impact on the final results. That is the basic motivation of our work. Furthermore, we show that ontologies both for events and geographic terms can improve the results greatly.

In the following, we will first describe our system in detail, and then the experiment results, followed by some discussions and a conclusion.

2 System Description

The architecture of our system is a straightforward pipeline system consists of query processing, document indexing, and document ranking. Since we focus mainly on the first component, we will not talk about the rest two in this report, which is a straightforward use of Lucene[1].

The query processing module can be further divided into three sub modules: topic parsing, keywords expansion, and query construction. We preprocess the input topics and documents with named-entity (NE) recognition. The documents are indexed after that; and the topics with NE annotations are sent to later processing stages. The following picture shows the workflow,

[1] http://lucene.apache.org/

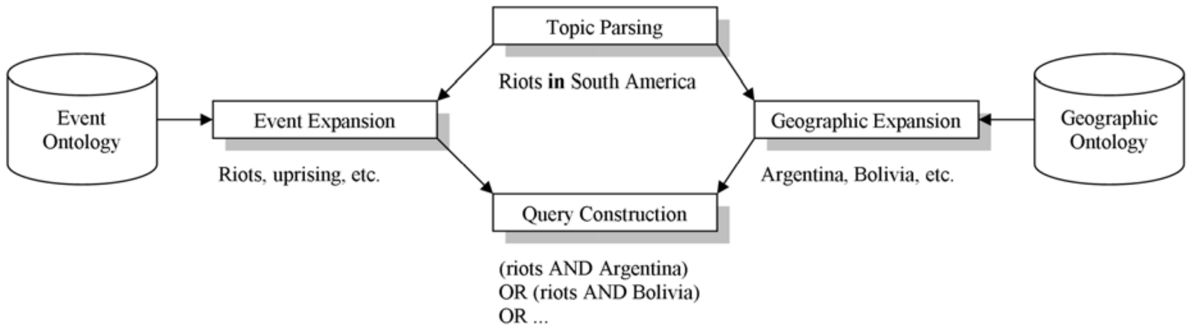


Fig. 1. Topic Parsing splits each topic into two parts, the *Event* part and the *Geographic* part, and send them to Event Expansion and Geographic Expansion components. These two components are assisted by Event Ontology and Geographic Ontology respectively. After the expansion, the query for the indexed documents will be constructed by Query Construction process.

2.1 Topic Parsing

As mentioned before, we preprocess the input topics with NE recognition and then identify the two parts of each topic, i.e. the *Event* part and the *Geographic* part. By doing this, we use prepositions as indicators for the division. Some topics are listed as follows,

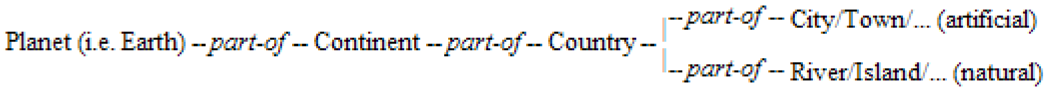
…
Riots ***in*** *South American prisons*
Nobel Prize winners ***from*** *Northern European countries*
Portuguese immigrant communities ***in*** *the world*
…
Most visited sights ***in*** *the capital* ***of*** *France and its vicinity*
…

In most cases, the prepositions do a good job as in the first three examples. Together with the NE information (i.e. location names), the two corresponding parts will be identified out. However, there are some cases, like the last example, which consist of several parts, if they are divided by prepositions. In practice, we take location names as the *Geographic* part (marked with double underline) and all the rest as the *Event* part (marked with underline). If there is no *Geographic* parts (e.g. errors from NE recognition), the last part will be assigned as the *Geographic* part heuristically. Consequently, in the last example, "*the capital*" could also be assigned as the *Geographic* part, if the NE recognizer takes it as the part of a location name.

2.2 Ontology-based Keywords Expansion

In this step, the *Event* part and the *Geographic* part will be tackled separately, assisted by two ontologies. We will first introduce these two ontologies and then how we use them.

Geographic Ontology. After referring several geographic taxonomies (Geonames[2], WorldGazetteer[3], etc.), we construct a geographic ontology using geographic terms and two relations. The backbone taxonomy of the ontology is as follows,


Planet (i.e. Earth) --part-of-- Continent --part-of-- Country --
--part-of-- City/Town/... (artificial)
--part-of-- River/Island/... (natural)

[2] Geonames geo coding web service: http://www.geonames.org/
[3] WorldGazetteer: http://www.world-gazetteer.com

Fig. 2. The basic structure of the geographic ontology consists of geographic terms referring different granularities of areas. Inside each Country, we have two categories of fine-grained places, i.e. artificial divisions and natural places. The basic relation in-between is the directional *part-of* relation, which means the geographic area on the left side contains the area on the right side.

In addition, extra geographic areas are connected with these basic terms using the same *part-of* relation. For example, the following geographic areas consist of the basic terms above,

Subcontinent: *the Indian subcontinent, the Persian Gulf*, etc.
Subcountry: *Lower Saxony, the Western USA*, etc.
Organization: *the Organization for Economic Co-operation and Development (OECD)*, etc.
Others: *Spanish islands*, etc.

An additional *equal* relation is utilized for synonyms and abbreviations of the same geographic area, e.g. *the United Kingdom, the UK, Great Britain*, etc.

Event Ontology. The event ontology is constructed using Wikipedia as an extra resource. Unlike the linguistic classification of events, we consider this ontology as a rather flat structure of two main categories, natural events and human activities. The first category mainly contains natural disasters, e.g. floods, earthquakes, etc; the second category takes all the rest, e.g. meetings, sports, wars, etc. The term event is a bit vague here, since we also include entities in the second category, e.g. Nobel Prize winners, politicians, etc. Two examples are given here,

Earthquakes*: San Francisco Earthquake (1906), Good Friday Earthquake Earthquake (1964)*, etc.
Nobel Prize winners: *Marie Curie (Russian Poland, Physics, 1903), Albert Einstein(Germany, Physics, 1921), Mother Teresa (Albania, Peace, 1979)*, etc.

Keywords Expansion. The population of the ontologies is done with either the narratives given or Wikipedia. The former can be done automatically from the texts after NE recognition; the latter has to be done manually. The usage of the event ontology is quite straightforward, that is directly taking all the terms contained in that category; the use of the geographic ontology follows the rule,

1. If the geographic part contains the granularity of the basic terms, e.g. country, city, etc, the ontology will provide all the geographic terms at that level;
2. Otherwise, the ontology will provide all the geographic terms below the level of that term.

Since the submissions can be either automatic or manual, we treat the automatic population of the ontologies as the former case and under the help of Wikipedia as the latter case.

2.3 Query Construction

After the expansion of both the events and the geographic terms, the query can be constructed using Boolean operators. In order to achieve both high precision and recall, we setup four levels of queries, giving different weights for the retrieved documents. The higher levels of queries aim to obtain accurate results, while the lower levels for the recall. The four levels are as follows,

Level 4 (1000): the event ontology AND the geographic ontology
Level 3 (100): the event terms AND the geographic ontology
Level 2 (10): the event terms AND the geographic terms
Level 1 (1): the event terms OR the geographic terms

Here, *event terms* and *geographic terms* mean those words appearing in the topics but not the narratives. In fact, both the event ontology and the geographic ontology can be further divided into two cases, the *automatic* meaning the ontology is constructed automatically using the narratives and the *manual* meaning the ontology is constructed also with Wikipedia information. In practice, when we made the submissions, we tested several combinations of queries from different levels (see the following section for details).

3 Submissions and Results

In the GeoCLEF track, we submitted 5 runs for the monolingual task of English. Different runs were constructed from combinations of different levels of queries. As mentioned above, the system can provide us 4 levels of queries with different weights for the retrieved documents. Therefore, to test how much each level can contribute to the final results, we setup the following 5 runs,

Run1 (M): Use queries from Level 1~4 and both ontologies are constructed with Wikipedia information
Run2 (A): Similar to Run1, but both ontologies are constructed with narratives
Run3 (M): Use queries from Level 1~3 and the ontology is constructed with Wikipedia information
Run4 (A): Similar to Run3, but the ontology is constructed with narratives
Run5 (A): Use queries from Level 1~2

Since we consider the ontologies constructed from Wikipedia are manual work, Run1 and Run3 are Manual (M) submissions and the other three are Automatic (A) submissions. The following table shows the final results of our five submissions,

Table 1. Results of our five submissions.

Submissions	R-Prec	MAP
Run1 (M)	33.38% (1/68)	29.18% (3/68)
Run2 (A)	33.19% (2/68)	29.24% (2/68)
Run3 (M)	31.70% (3/68)	30.37% (1/68)
Run4 (A)	31.41% (4/68)	27.73% (6/68)
Run5 (A)	20.95% (58/68)	16.07% (68/68)

We are quite satisfied with the results of our first participation, not only the performance, but also the impact of focusing on ontology-based query expansion for GIR. The best automatic submission will be Run2, which has both high R-Prec and MAP scores. For the best manual submissions, Run1 and Run3 have the best R-Prec and MAP scores respectively. Comparing automatic and manual submissions, the R-Prec has a slight difference, while for MAP, the difference is bigger. Consequently, the manual work of populating the ontology with Wikipedia information does help to improve the precision. At last, only using the terms in the topics without any help from the narratives or Wikipedia, the results are quite poor (Run5). This suggests the query expansion a step of great importance.

Taking a closer look at the results, we find that the system performs very well in some topics, but worse in some others. This may be because the improvement from the ontology is not stable, since different topics contain various events, which cannot be treated uniformly.

Additionally, since the only language dependent components of our system are the NE recognizer and the ontologies, we also did experiments on the German data sets. The SPPC system (Neumann and Piskorski, 2002) was used for German NE recognition and the ontologies were constructed with the help of German Wikipedia. The preliminary evaluation was not so satisfactory, so that we did not make submissions, but the framework of our system suggests an easy extension to other languages.

4 Conclusion and Future Work

In this paper, we showed our participation of the GeoCLEF track. Our approach focused on the query processing part instead of the document ranking as in traditional IR systems. In particular, we analyzed the topics and applied ontologies to expand the keywords in both the geographic part and the event part. We also setup four levels of queries in order to achieve both high precision and recall. The results suggest the success of our strategy and techniques.

In the future, we will take into account the document ranking part as well. One direction could be to use a context window to control the distance between the event and the geographic term in order to filter out some documents or give them lower weights. More experiments on other languages are also considered by us in the near future.

References

1. Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. 2007. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In CLEF2007 Working Notes.
2. Neumann, G. and Piskorski, J. 2002. A Shallow Text Processing Core Engine. Journal of Computational Intelligence, Volume 18, Number 3, 2002, pages 451-476.
3. Singhal, Amit. 2001. "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.
4. Geonames geo coding web service: http://www.geonames.org/
5. WorldGazetteer: http://www.world-gazetteer.com