Information Processing and Management xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Information retrieval with geographical references. Relevant documents filtering vs. query expansion ☆

Miguel Á. García-Cumbreras *, José M. Perea-Ortega, Manuel García-Vega, L. Alfonso Ureña-López

Department of Computer Science, University of Jaén, Jaén, Spain

ARTICLE INFO

Article history:
Received 13 July 2008
Received in revised form 13 April 2009
Accepted 14 April 2009
Available online xxxx

Keywords:
Geographic Information Retrieval
Query expansion
Document filtering
CLEF

ABSTRACT

This is a thorough analysis of two techniques applied to Geographic Information Retrieval (GIR). Previous studies have researched the application of query expansion to improve the selection process of information retrieval systems. This paper emphasizes the effectiveness of the filtering of relevant documents applied to a GIR system, instead of query expansion. Based on the CLEF (Cross Language Evaluation Forum) framework available, several experiments have been run. Some based on query expansion, some on the filtering of relevant documents. The results show that filtering works better in a GIR environment, because relevant documents are not reordered in the final list.

1. Introduction

Geographical information is recorded in a wide variety of media and document types. Over the past few decades, information technology for accessing geographical information has focused on the combination of digital maps and databases that characterize the majority of Geographic Information Systems (GIS) (Bolstad, 2005; Chang, 2007; Wise, 2002). It is only in recent years that much attention has been paid to the development of computer systems to retrieve geographically specific information from the relatively unstructured documents that compose the Web (Jones & Purves, 2008; Larson, 1996; McCurley, 2001).

Much current research in Geographic Information Retrieval (GIR) can be regarded as an extension of the field of Information Retrieval (IR) (Baeza-Yates & Ribeiro-Neto, 1999), and indeed has primarily been undertaken by researchers from the IR rather than the GIS research community. It includes all of the research areas that have traditionally made up the core of research in Information Retrieval, except that it has an emphasis on spatial and geographic indexing and retrieval (Larson, 1996). GIR is therefore concerned with improving the quality of geographically specific information retrieval with a focus on access to unstructured documents (Jones & Purves, 2008).

This paper aims at a brief description of GIR and a thorough analysis of two techniques, namely query expansion and filtering of relevant documents.

Because a GIR system can be seen as an IR system from a functional point of view, GIR systems can work with the same document collections as IR systems, accepting that the queries usually include location entities (something that has a dis-

☆ This work has been supported by the Spanish Government (MCYT) with Grant TIN2006-15265-C06-03.
* Corresponding author.
E-mail addresses: magc@ujaen.es (M.Á. García-Cumbreras), jmperea@ujaen.es (J.M. Perea-Ortega), mgarcia@ujaen.es (M. García-Vega), laurena@ujaen.es (L. Alfonso Ureña-López).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.04.006

tinct, separate existence, concerning a location) and spatial awareness (a well thought-out awareness of things in the space around us). For instance,

- "Wine regions around rivers in Europe."
- "Cities within 100 km of Frankfurt."
- "Whisky making in the Scottish Islands."
- "Events at St. Paul's Cathedral."

Evaluation campaigns such as the Cross Language Evaluation Forum[1] (CLEF) have tasks to evaluate geographical IR systems, all grouped under the name "GeoCLEF"[2]. The aim of GeoCLEF is to provide the necessary framework in which GIR systems are evaluated for search tasks involving spatial and multilingual aspects. We have participated in the last three editions of GeoCLEF (Perea-Ortega, Cumbreras, Vega, & Lpez, 2007, 2008; Vega, Cumbreras, Lpez, & Perea-Ortega, 2006).

There are several approaches to solve the GIR task, including simple IR systems that do not use geographical terms or spatial references. More complex systems rely on Natural Language Processing (NLP) methods to detect locations and spatial references. Some techniques that are often applied are "*geographic entities extraction*", "*semantic analysis*", "*geographical data bases*" (thesauri, gazetteers), "*query expansion*" methods and "*geographical disambiguation*".

Most of the systems presented in GeoCLEF, for the English monolingual task between 2005 and 2008, preprocess the collections and the queries applying a *stop-word* list (Salton, 1971) (to delete the most frequent words in each language) and a stemmer (Porter, 1980) (to extract the stem of each non-*stop-word*).

They also use a Named Entity Recognizer (NER), a module that detects and recognizes named entities. Some groups have implemented their own NER module using geographical databases and thesauri (FerrTs & Rodrfguez, 2007; Larson, 2007), but most groups use an external one, like *LingPipe*[3] (Andogah & Bouma, 2007; Buscaldi & Rosso, 2007; Hu & Ge, 2006).

Although GIR is not typically multilingual, in our GIR system a translation module is necessary because it works with multilingual queries.

The paper is organized as follows: Section 2 describes the complete system and the operations; Section 3 presents the resources used in this empirical comparison and the experimental results; and in Section 4 the conclusions of this work are written.

2. System overview

The SINAI GIR system is made up of five main subsystems: *Translator*, *Entity Extraction*, *Geo-Information Extraction*, *Information Retrieval* and *Filtering and Re-ranking*. They are introduced in the following section. For query expansion experiments, we have added a *Query Expansion* subsystem. In addition, we make use of the *Geonames*[4] gazetteer as geographic knowledge base for the whole system, and Lemur[5] as IR index-search engine.

First of all, each translated query is preprocessed and analyzed by the *Geo-Information Extraction* subsystem, identifying their geo-entities and spatial relationships. Each query will be run later against the IR subsystem. On the other hand, the document collection is preprocessed and its entities are extracted by the *Entity Extraction* module. Finally, the documents recovered by the IR subsystem are filtered and re-ranked by means of the *Filtering and Re-ranking* subsystem, making use of all geo-information extracted from document collection and queries.

2.1. System design and operations

As we can see in the Fig. 1, the main modules of our system are:

- **Translator subsystem**. We have developed and used the SINTRAM[6] translator (GarcíaCumbreras, Ureña-López, Martínez-Santiago, & PereaOrtega, 2006). This own subsystem translates different queries into English and implements some heuristics in order to combine various translations of the same query. A comprehensive evaluation showed that Systran[7] worked best for German and Portuguese (Vega et al., 2006).
- **Entity Extraction subsystem**. GATE[8] is an infrastructure for developing and deploying software components that process human language. It includes a NER module to detect and to recognize entities.

[1] http://www.clef-campaign.org/.
[2] http://ir.shef.ac.uk/geoclef/.
[3] http://www.alias-i.com/lingpipe.
[4] http://www.geonames.org/. Geonames is a geographic database which contains over eight million geographical names and consists of 6.3 million unique features whereof 2.2 million populated places and 1.8 million alternate names.
[5] http://www.lemurproject.org/.
[6] SINai TRAnslation Module.
[7] http://www.systransoft.com.
[8] http://gate.ac.uk/.

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

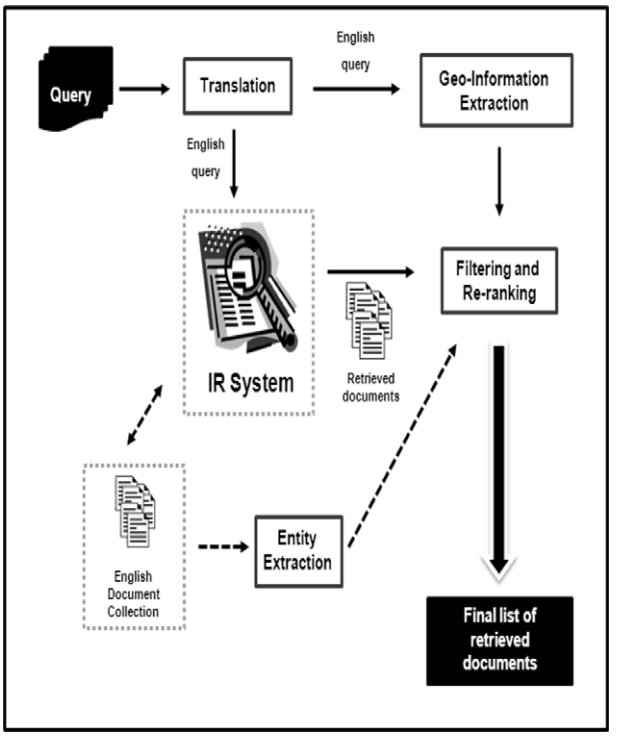

Fig. 1. Architecture of the SINAI GIR system.

- **Geo-Information Extraction subsystem**. The aim of this module is to recognize spatial relationships and locations in the queries:
  - As in the *Entity Extraction* subsystem, we have used GATE in order to detect the geo-entities. All locations detected are also verified using the GeoNames gazetteer.
  - In order to find spatial relationships, we have used manual rules. A spatial relationship is a constraint that appears in a spatial query to select spatial objects (Clementini, Sharma, & Egenhofer, 1994). They may include such simple selections as, "Lakes *in* the state of Maine", or more complex ones like "Find the *shortest path* from Boston to Bangor". Spatial relationships are divided into three main categories (Clementini & DiFelice, 2000): topological (e.g., *museums in Rome*), metric (e.g., *Rome is not far from L'Aquila*) and projective (e.g., *the shops on the right of the road*). Our system detects and recognizes topological spatial relationships. Some examples of them are: *in, of, near, north of, next to, in or around, in the west of*, etc.

- **Information Retrieval subsystem**. We have used Lemur as IR index-search engine.
- **Filtering and Re-ranking subsystem**. It is intended to filter the list of relevant documents recovered by the IR subsystem, establishing what of them are valid, depending on the locations and the spatial relationships detected in the query. Another important function is to establish the final ranking of documents, based on manual rules and the initial position of the document in the ranking. This process is explained in more detail in the next section.

2.2. Relevant documents filtering

This process filters the list of relevant documents recovered by the IR system, validating them and obtaining a new ranking. This validation uses two important indexes:

- **Documents Index**. It stores the preprocessed information of documents collection. We have preprocessed the collection with linguistic tools to remove *stop-words* and to mark stems. This index contains all the stems from the collection, including the original entities. We have used Lemur to build the documents index.
- **Geographical Index**. It stores all the locations detected in the collection by the *Entity Extraction* subsystem. All entities typified as *LOC* (location) by the NER are checked using the GeoNames gazetteer. Organizations are not included.

On the other hand, before the filtering process, the *Geo-Information Extraction* subsystem determines the type of each location detected in the query. We have considered four main types of locations:

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

- **Continent**. GeoNames uses the *CONT* code for the continent type. It recognizes seven entities as continents: Europe, Africa, Asia, North America, South America, Oceania and Antarctica.
- **Country**. We have considered administrative divisions of a country as country type. GeoNames uses several codes in order to classify them: *ADM1*, *ADM2*, *ADM3*, *ADM4*, *ADMD*, etc. We have also considered political entities (*PCL* code) as country type.
- **City**. Several codes are used by GeoNames to identify a city or a village: *PPL* for populated place, *PPLC* for capitals of a political entity, *PPLL* for populated localities, *PPLS* for populated places, etc. We use all of them in order to classify an entity as city type.
- **Place**. It is any entity that has not been considered as some type before.

Fig. 2 describes the basic architecture for the filtering process.

In the beginning, the rules applied in the filtering process were very restrictive, and few heuristics were used. For instance, we considered a document as valid if it contained only one location which appeared in the query. Moreover, the filtering process did not make a re-ranking with the valid documents. They are included with the score assigned by the IR system.

In the last experiments carried out in 2008, the filtering process includes new heuristics and a re-ranking process. Depending on the location type and the spatial relationship detected in the query, the documents recovered by the IR system are filtered and re-ranked. Some examples of these filtering rules are:

- If the entity type is *country* and it has associated "*in the north of*" as spatial relationship, the system obtains the maximum and minimum latitudes of all locations which belong to that country, using again the GeoNames gazetteer. In order to estimate the *mid-latitude* of a country, the system subtracts the maximum and minimal latitudes. Any location which is above of this *mid-latitude* will be considered in the northern part of the country and, therefore, the document will be valid. Some approaches take a more sophisticated point of view of partitioning space or regions according to direction (van Kreveld & Reinbacher, 2004).
- If the entity type is *city* and its spatial relation associated is "*near*", there is a difficult problem to solve. Spatial reasoning is a complex activity that involves at least two levels of representation and reasoning: a geometric level where metric, topological, and projective properties are handled (Herskovits, 1986); and a functional level where the normal function of an entity affects the spatial relationships attributed to it in the context (for example, the meaning of "*near*" for a bomb is

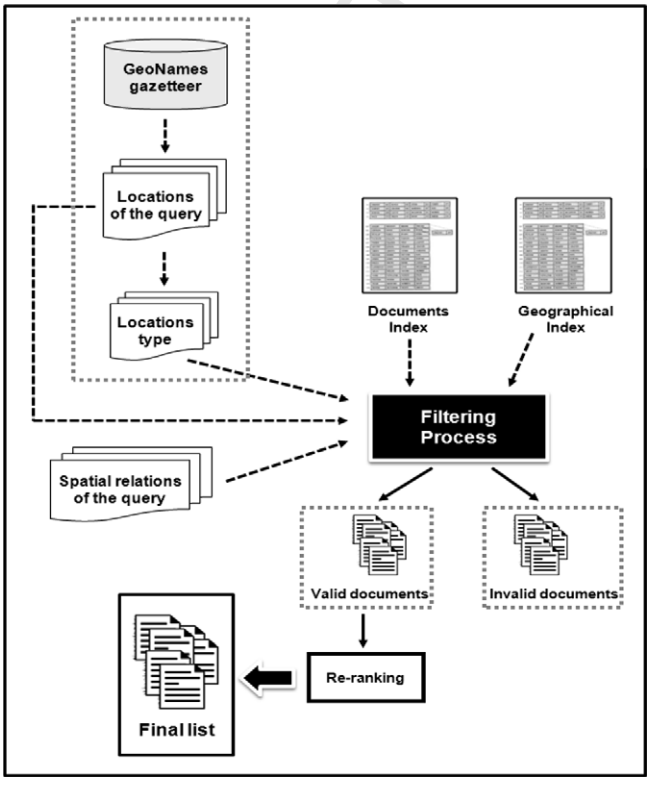

Fig. 2. Basic architecture of the filtering process.

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

quite different from the meaning of "*near*" for other objects of the same size (Coventry, 1998; Garrod, Ferrier, & Campbell, 1999)). Without much further on this issue, only for *point-to-point* distances, we have decided to consider that a location is *near* to another when their distance is less than 50 km. This value has been determined based on experiments carried out in the framework proposed in this paper. In order to measure the *point-to-point* distance in kilometers between two locations, we have used the *Great-Circle* formula Sinnott (1984):

$$D = arcos((\sin a)(\sin b) + (\cos a)(\cos b)(\cos P))$$

where

D is the distance, in kilometers;
a is the latitude of point A;
b is the latitude of point B;
P is the longitudinal difference between points A and B.

- If the entity type is *continent* or *country* and its spatial relationship associated is "*in*", "*of*", "*at*", "*on*", "*from*" or "*along*", the system will accept the document wether it has at least one location that belongs to that continent or country.

After the filtering process, in order to make the final ranking, the system handles two documents lists:

- The list of valid documents. In this list there are only documents recovered by the IR system which satisfy at least one of the filtering rules.
- The list of invalid documents. This list includes the documents which have not satisfied any filtering rule.

The valid documents appear in the final list, whereas the invalid ones are deleted. The initial score of each document is provided by the IR system. For the re-ranking of the valid documents, the system increases their initial scores depending on the filtering rules which comply and the position of the document in the ranking:

$$RSV_{new} = \log(RSV_{old} + 1) + w_{rule}$$

where w_{rule} is the weight associated to the filtering rule that has been applied to the document. These weights have been optimized and established empirically using the framework proposed in this work.

2.3. Query expansion

We have developed a *Query Expansion* architecture in order to carry out the query expansion approach. The aim is to expand queries with many relevant words as possible. This applies not only to geographical terms, like countries, towns and cities, but also to synonyms, terms from the thesauri collections and any word in the same semantic domain.

The *Query Expansion* subsystem is made up of three main modules:

- **Named Entity Recognition**. It detects and recognizes the location entities in the queries in order to expand the topics with geographical data. Examples of location terms are: towns, cities, capitals, countries and even continents.
- **Geographical Information**. The purpose of this module is to expand the locations detected in the previous module, using geographical information of Geonames. To do this, it is also necessary to detect and recognize spatial relationships in the query.
- **Thesaurus Expansion**. This module has its own thesauri collection, generated from the GeoCLEF training corpus according to a very high word co-location rate (Croft & Yufeng, 1994; Jones, 1971). This module finds the best thesaurus terms and adds them to the query. All terms are considered, including the geographic entities.

Fig. 3 describes the architecture for the query expansion.

3. Experiments and results

In this section the experiments carried out during the years 2006–2008 are described. We have used the evaluation framework provided by GeoCLEF (Gey et al., 2006; Mandl et al., 2007). Its aim is to evaluate GIR systems for search tasks, involving spatial and multilingual aspects. This framework provides an English document collection and different textual queries in several languages in order to test each GIR system:

– **The document collection**. It consists in 169,477 documents, composed of stories and newswires from the British newspaper *Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994) (Braschler & Peters, 2004). This collection contains stories covering international and national news, therefore representing a wide variety of geographical regions and places. The documents have a common structure: newspaper-specific information like date, page, issue, special filing numbers and usually one or more titles, a byline and the actual text. They were not geographically tagged and contained no semantic location-specific information (Mandl et al., 2007).

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

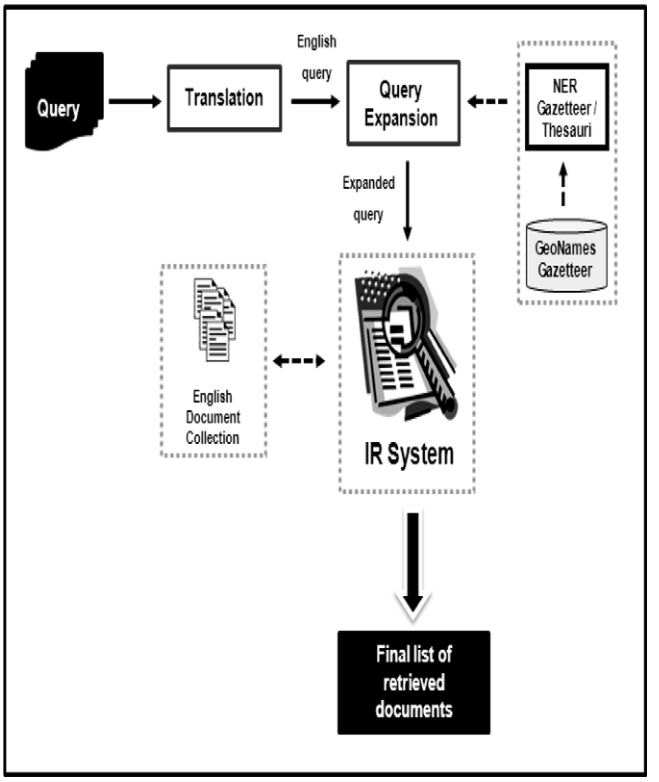

Fig. 3. Architecture of the query expansion.

– **The textual queries**. They are textual descriptions with three fields: *title* (T), *description* (D) and *narrative* (N). In some experiments we have used all fields (TDN), and other only TD labels. The format of these queries are showed in the Fig. 4.

3.1. Experimental method

The experiment framework has been set as follows:

(1) The document collection and the queries are preprocessed using the English *stop-words* and the Porter stemmer algorithm. Geo-entities are also extracted.
(2) The non-English queries are translated by means of the *Translator* subsystem.
(3) In order to generate the *Documents Index* (see Section 2.2), we have used Lemur. It is also used to retrieve the relevant documents for each query, applying traditional weighting methods such as TF · IDF, Okapi (Robertson & Walker, 1999) and the use of Pseudo-Relevant Feedback (PRF) (Salton & Buckley, 1990). We have used the *Okapi* + PRF weighting function because it always offers the best performance in all experiments.

```
<top>
 <num>10.2452/58-GC</num>
    <title>Travel problems at major airports near to London</title>
    <desc>To be relevant, documents must describe travel problems
    at one of the major airports close to London.</desc>
    <narr>Major airports to be listed include Heathrow, Gatwick,
    Luton, Stanstead and London City airport.</narr>
</top>
```

Fig. 4. Example of a query.

(4) For the query expansion experiments, we have used the GeoNames gazetteer in order to expand with geographic terms, and a thesaurus generated from the collection in order to expand with terms treated like synonyms. Both approaches are described in Section 2.3.

The final list of documents retrieved is evaluated using the relevance judgements provided by GeoCLEF (Gey et al., 2006; Mandl et al., 2007) and the TREC evaluation method. The evaluation has been accomplished by using the Mean Average Precision (MAP) (Harman, 1994) and the R-Precision. The MAP measure 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*" (Salton & Buckley, 1990). R-precision is the precision at the number of relevant documents in the collection for the query. If R-precision is 1.0, it means a perfect relevance ranking and a perfect recall. Recall is a measure of the extent to which relevant documents are found or retrieved. Recall is 1.0 when every relevant document is retrieved.

Table 1 shows all results per experiment using the GeoCLEF 2006 topics. Table 2 shows the same experiments but using the GeoCLEF 2007 topics. We have summarized them, showing only the best results. They have been achieved using all topic labels (TDN). The results obtained using TD labels are worse in general.

3.2. Experiments with query expansion

In this section we study the impact of the query expansion approach. We considered as baseline case for this approach not use any expansion method. We compared two types of query expansion:

- **Using geographic data**. We have used several heuristics or manual rules in order to expand locations using geographic data. This geo-information is extracted from the GeoNames gazetteer. Previously, the spatial relationship and the location included in the query have been detected. Examples of these heuristics are:
  - If the location type is *continent* and it has associated *in*, *of* or *from* as spatial relationships, we expand the query with the highest population countries of that continent. For example, if the query is "*Wine regions around rivers in Europe*", we expand it with *France*, *Germany*, *Russia* and *United Kingdom* as geo-terms, and therefore the preprocessed query would be "*wine region river europe france german russia united kingdom*".
  - If the location type is *country* and it has associated *in*, *of* or *from* as spatial relationships, we expand the query with the highest population cities of that country. For instance, if the query is "*Eta in France*", we expand it with *Paris*, *Marseille* and *Lyon* as geo-terms, and therefore the preprocessed query would be "*eta france paris marseille lyon*".

Table 1
Summary of results using the **GeoCLEF 2006 topics**.

Language	Experiment	R-Precision	MAP
English	Baseline	0.2934	0.3223
	Expansion using geographic data	0.2028	0.2295
	Expansion using thesaurus	0.2261	0.2610
	Filtering	0.2934	**0.3225**
German	Baseline	0.1818	0.1965
	Expansion using geographic data	0.1506	0.1838
	Expansion using thesaurus	0.1556	0.1864
	Filtering	0.1774	**0.2113**
Spanish	Baseline	0.2335	0.2572
	Expansion using geographic data	0.2341	0.2604
	Expansion using thesaurus	0.2334	0.2565
	Filtering	0.2260	**0.2634**

Table 2
Summary of results using the **GeoCLEF 2007 topics**.

Language	Experiment	R-Precision	MAP
English	Baseline	0.2578	0.2619
	Expansion using geographic data	0.2376	0.2380
	Filtering	0.2716	**0.2661**
German	Baseline	0.0666	0.0652
	Expansion using geographic data	0.0819	**0.0842**
	Filtering	0.0827	0.0763
Spanish	Baseline	0.2200	0.2361
	Expansion using geographic data	0.2060	0.2076
	Filtering	0.2369	**0.2441**

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

– If the location type is *city* and it has associated *near* as spatial relationship, we expand the query with the highest population cities closest to that location. In order to measure the *point-to-point* distance in kilometers between two locations, we have used the *Great-Circle* formula (see Section 2.2). For example, if the query is "*Travel problems at major airports near to London*", we expand it with *Luton*, *Reading* and *Maidstone* as geo-terms. The preprocessed query would be "*travel problem major airport london luton reading maidstone*".

- **Using thesaurus**. We have generated an own thesaurus from the GeoCLEF document collection, attending to a very high word co-location rate. We have used all the preprocessed words from queries, including the geographic terms. We expand the queries with the synonyms that have a similarity higher than a predefined threshold. The best results were obtained with a threshold value of 0.7. An example of query expansion using thesaurus is showed in the Fig. 5.

The analysis of results using the GeoCLEF 2006 topics shows that the use of geographical and thesaurus information does not improve the retrieval. The expansion method obtained poor results compared to the simple model in which we only use an IR system (baseline experiment). The use of the text from all topic labels improves the retrieval process, as opposed to consider the text from one or two topic labels. The query expansion with words from the thesaurus included noise but, in general, it worked better than the expansion using geographic knowledge. Two main reasons can explain the worse results obtained with the expansion of topics:

- For the information retrieval process, the inclusion of geographic knowledge as location entities introduced noise in the queries because the expansion heuristics were not correct.
- Sometimes, the NER module did not work well. For several queries the NER did not recognize some compound entities and locations (*New England*, *Middle East*, *Eastern Bloc*, etc.) and therefore, for these topics, the system was unable to realize the query expansion using geographical information.

This query expansion method has also been tested using the 2007 GeoCLEF topics. For these experiments we have expanded the queries using geographical information only and all the labels in the IR process. Table 2 shows these results. Figs. 6 and 7 show a graphic comparison of these experiments based on MAP and R-Precision measurements respectively.

3.3. Experiments with relevant documents filtering

In this section we study the impact of the relevant documents filtering. In 2007, we completely changed the approach of our system, because previous experiments demonstrate that the query expansion heuristics applied do not improve the re-

```
Original topic: Malaria in the tropics
Expanded topic: malaria tropic plasmodium heimlich
                neurosyphili timpone vivax
```

Fig. 5. Example of a query expansion using thesaurus.

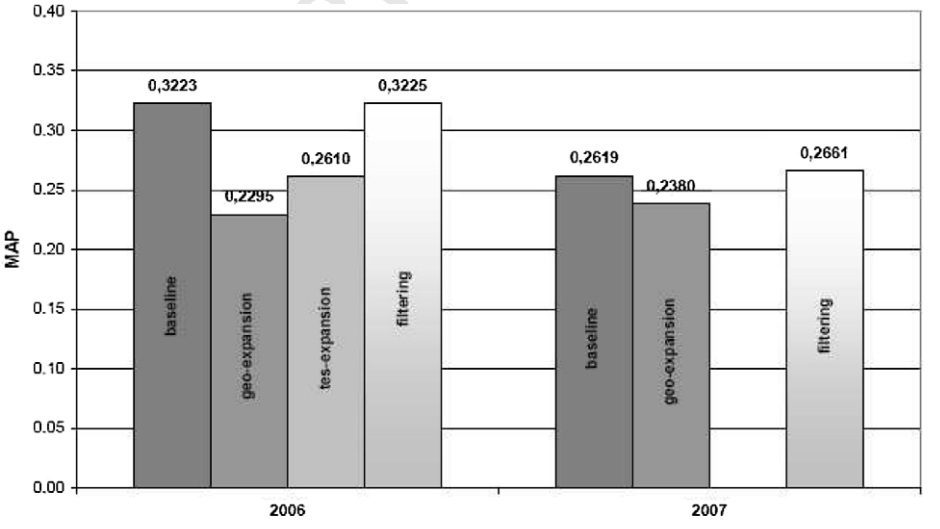

Fig. 6. Graphic comparison based on MAP measure using English topics.

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

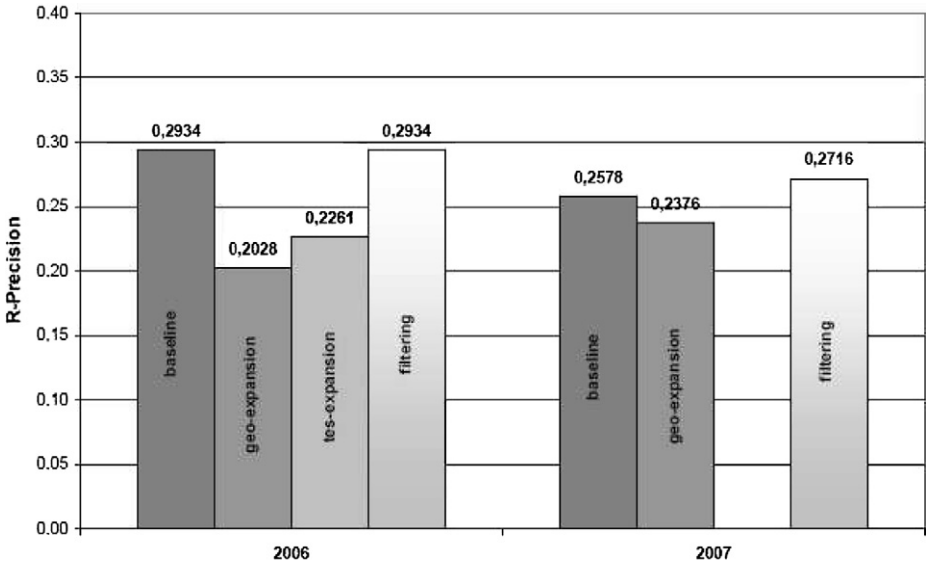

Fig. 7. Graphic comparison based on R-Precision measure using English topics.

sults. We introduced a very restrictive system: we removed those documents recovered by the IR system that did not satisfy certain filtering rules. Some of these manual rules have been explained in the Section 2.2.

For the filtering experiments we used all topic labels in the retrieval process. As baseline experiment we considered a simple information retrieval without applying any filtering process. We tried several weighting functions in the IR process (TF · IDF, Okapi and PRF), but in Tables 1 and 2 we only show the best results, using *Okapi + PRF*.

The main conclusion of this approach is that the filtering of the documents retrieved by the IR system works better than the expansion of queries. In addition, the results obtained by the filtering process improved the baseline cases (all experiments and all languages). Instead, the difference between the use of query expansion approach and the use of document filtering is important.

As we can see in Fig. 6, using 2006 English topics, document filtering has an improvement of 9.3% with regard to use query expansion with geographic information. This difference is also appreciated using 2006 German topics (2.7%) and using 2007 Spanish topics (3.6%). In other experiments, this difference is not very important, although document filtering gets always better results than query expansion, except using 2007 German topics, where the *Translator* subsystem did not work very well.

In some filtering experiments, the mean average precisions obtained with this approach and the baseline experiments are similar. The main reason is that the documents filtered are valid but their positions in the final ranking are similar to the original ones in the baseline result. The *Filtering and Re-ranking* subsystem should work better for these experiments in the future, maybe doing a thorough query analysis, identifying more accurately the geographic part in the topic. This would allow a more optimal re-ranking process for each relevant document.

4. Conclusions and future work

In this work we present a comparison between two approaches applied to the Geographic Information Retrieval task: query expansion vs. relevant documents filtering. We have used the GeoCLEF framework in order to carry out the experimentation. The results show that relevant documents filtering works better than query expansion in a basic GIR system, so the use of a filtering module is an important method for GIR systems.

However, the query expansion is an interesting method for some geographic queries. For instance, for those topics that contain a subregion of the world not clearly defined (*Northern Africa, Sub-Saharan Africa, Caribbean, Patagonia, Iberian Peninsula*, etc.). As future work, it would be interesting to consider the best way to make the geographical expansion of a topic, depending on the query type.

In order to improve the final results, we think that GIR systems require the combination between relevant documents retrieved by an IR system and the application of a geographical module. Usually, general approaches expand the queries with geographical information but in a wrong way. Experiments reported in this work show that a filtering process, instead of the query expansion, produces better results. In this case, an improvement could be the use of a higher number of documents recovered by the IR subsystem, so the probability of valid documents returned by the filtering and re-ranking process is higher.

On the other hand, the application of Pseudo-Relevant Feedback (PRF) and the traditional Okapi weighting method works well and good results are obtained by the IR subsystem.

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006

Our future work will include an improvement of the filtering and re-ranking process when the relevant documents obtained from the IR phase report a low score. The filtering method does not introduce new relevant documents that are not returned by the IR system, so we will focus on improving those queries that report a low IR score. Another query expansion module will be developed for these queries. In addition, since the Web contains valid geo-referenced information, we are interested in the development of a new module that validates the final relevant documents using the Web.

References

Andogah, G., & Bouma, G. (2007). Relevance measures using geographic scopes and types. In *CLEF* (pp. 794–801).
Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern information retrieval*. ACM Press/Addison-Wesley<http://www.ischool.berkeley.edu/hearst/irbook/glossary.html> .
Bolstad, P. (2005). *GIS fundamentals: A first text on geographic information systems* (2nd ed.). Eider Press.
Braschler, M., & Peters, C. (2004). Cross-language evaluation forum: Objectives, results, achievements. *Inf. Retr., 7*(1–2), 7–31<http://dblp.uni-trier.de/db/journals/ir/ir7.html> .
Buscaldi, D., & Rosso, P. (2007). On the relative importance of toponyms in geoclef. In *CLEF* (pp. 815–822).
Chang, K. (2007). *Introduction to geographic information system* (4th ed.). McGraw-Hill College.
Clementini, E., & DiFelice, P. (2000). Spatial operators. *SIGMOD Record, 29*(3), 31–38.
Clementini, E., Sharma, J., & Egenhofer, M. (1994). Modeling topological spatial relations: Strategies for query-processing. *Computers and Graphics, 18*(6), 815–822.
Q2 Coventry, K. R. (1998). Spatial prepositions, functional relations, and lexical specification (pp. 247–262).
Croft, B., & Yufeng, J. (1994). An association thesaurus for information retrieval. In Funck-Brentano, J.-L., & Seitz, F. (Eds.), *RIAO. CID* (pp. 146–161). <http://dblp.uni-trier.de/db/conf/riao/riao1994.html>.
FerrTs, D., & Rodrfguez, H. (2007). Talp at geoclef 2007: Results of a geographical knowledge filtering approach with terrier. In C. Peters, V. Jijkoun, T. Mandl, H. Mnller, D. W. Oard, & A. Peas, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 5152, pp. 830–833). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2007.html> .
GarcíaCumbreras, M., Ureña-López, L., MartínezSantiago, F., & PereaOrtega, J. (2007). Bruja system. the university of jaén at the spanish task of qa@clef 2006. In *Proceedings of the cross language evaluation forum (CLEF 2006)*.
Garrod, S., Ferrier, G., & Campbell, S. (1999). In and on: Investigating the functional geometry of spatial prepositions (Vol. 72).
Gey, F. C., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., et al (2006). Geoclef 2006: The clef 2006 cross-language geographic information retrieval track overview. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, & D. W. Oard, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 4730, pp. 852–876). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2006.html> .
Harman, D. (1994). Overview of the second text retrieval conference (trec-2). In *HLT*. Morgan Kaufmann<http://dblp.uni-trier.de/db/conf/naacl/naacl1994.html> .
Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of prepositions in English*. Cambridge University Press.
Hu, Y.-H., & Ge, L. (2006). The university of new south wales at geoclef 2006. In *CLEF* (pp. 905–912).
Jones, K. S. (1971). *Automatic keyword classification for information retrieval*. London: Butterworth.
Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science, 22*(3), 219–228<http://dblp.uni-trier.de/db/journals/gis/gis22.html> .
Larson, R. (1996). Geographic information retrieval and spatial browsing. In Smith, L., & Gluck, M. (Eds.), *Geographic information systems and libraries: Patrons, maps and spatial information* (pp. 81–124).
Larson, R. R. (2007). Cheshire at geoclef 2007: Retesting text retrieval baselines. In C. Peters, V. Jijkoun, T. Mandl, H. Mnller, D. W. Oard, & A. Peas, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 5152, pp. 811–814). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2007.html> .
Mandl, T., Gey, F. C., Nunzio, G. M. D., Ferro, N., Larson, R., Sanderson, M., et al (2007). Geoclef 2007: The clef 2007 cross-language geographic information retrieval track overview. In C. Peters, V. Jijkoun, T. Mandl, H. Mnller, D. W. Oard, & A. Peas, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 5152, pp. 745–772). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2007.html> .
McCurley, K. (2001). Geospatial mapping and navigation of the web. Association for Computing Machinery, Hong Kong, China. In *Tenth international world wide web conference*.
Perea-Ortega, J. M., Cumbreras, M. A. G., Vega, M. G., & Lpez, L. A. U. (2007). Filtering for improving the geographic information search. In C. Peters, V. Jijkoun, T. Mandl, H. Mnller, D. W. Oard, & A. Peas, et al. (Eds.), *CLEF. Lecture Notes in Computer Science* (Vol. 5152, pp. 823–829). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2007.html> .
Perea-Ortega, J. M., Cumbreras, M. A. G., Vega, M.G., & Lpez, L. A. U. (2008). Sinai-gir system. university of jaTn at geoclef 2008. In *Proceedings of the cross language evaluation forum (CLEF 2008)*.
Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.
Robertson, S. E., & Walker, S. (1999). Okapi/keenbow at trec-8. In *TREC*. <http://dblp.uni-trier.de/db/conf/trec/trec1999.html>.
Salton, G. (Ed.). (1971). *The SMART retrieval system. Experiments in automatic document processing*. Prentice-Hall: Englewood Cliffs.
Salton, G., & Buckley, G. (1990). Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences, 21*, 288–297.
Sinnott, R. (1984). Virtues of the haversine (Vol. 68).
van Kreveld, M. J., & Reinbacher, I. (2004). Good news: Partitioning a simple polygon by compass directions. *International Journal of Computational Geometry and Applications, 14*(4–5), 233–259<http://dblp.uni-trier.de/db/journals/ijcga/ijcga14.html> .
Vega, M. G., Cumbreras, M. A. G., Lpez, L. A. U., & Perea-Ortega, J. M. (2006). Geouja system. The first participation of the university of jaTn at geoclef 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, & D. W. Oard, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 4730, pp. 913–917). Springer<http://dblp.uni-trier.de/db/conf/clef/clef2006.html> .
Wise, S. (2002). *GIS basics*. CRC Press.

Please cite this article in press as: García-Cumbreras, M. Á..zz, et al. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.04.006