# INAOE at GeoCLEF 2008: A Ranking Approach based on Sample Documents

Esaú Villatoro-Tello, Manuel Montes-y-Gómez and Luis Villaseor-Pineda
Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), México
{villatoroe, mmontesg, villasen}@ccc.inaoep.mx

**Abstract**

This paper describes the system developed by the Language Technologies Laboratory of INAOE for the Geographical Information Retrieval task of CLEF 2008. The presented system focuses on the problem of ranking documents in accordance to their geographical relevance. It is mainly based on the following hypotheses: *(i)* current IR machines are able to retrieve relevant documents for geographic queries, but they can not generate a pertinent ranking; and *(ii)* complete documents provide more and better elements for the ranking process than isolated query terms. Based on these hypotheses, our participation at GeoCLEF 2008 aimed to demonstrate that using some query-related sample texts it is possible to improve the final ranking of the retrieved documents. Experimental results indicated that our approach could improve the MAP of some sets of retrieved documents using only an average of two sample texts. These results also showed that the proposed approach is very sensitive to the presence of irrelevant sample texts as well as to the ambiguity of geographical terms.

### Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

### General Terms

Measurement, Performance, Experimentation

### Keywords

Information Retrieval, Geographical Information Retrieval, Geographic Entity Recognition, Gazetteer

## 1 Introduction

Geographic Information Retrieval (GIR) is aimed at the retrieval of documents based not only on conceptual keywords, but also on spatial information (i.e. geographical references) [1, 2].

Recent development of GIR systems [3, 4, 5] evidence, on the one hand, that traditional IR machines are able to retrieve the majority of the relevant documents for most queries, but, on the other hand, that they have severe difficulties to generate a pertinent ranking of them. Based on these facts, we designed a new GIR method that aims to improve the ranking of retrieved documents by considering information from some query-related sample texts.

The proposed method was evaluated in the Monolingual English exercise of the 2008 GeoCLEF task. In particular, the purposes of our experiments were twofold: first, to confirm that traditional

IR machines can achieve high recall levels, and second, to probe that using some query-related sample texts allow improving the original ranking of the retrieved documents.

The following section describes the general architecture of our GIR system as well as the main functionality of each one of its components. Then, Section 3 shows the results from a subset of the official submitted runs to GeoCLEF 2008. Finally, Section 4 presents some conclusions and describes ongoing work.

## 2 System Description

Figure 1 shows the general architecture of the proposed system. It is divided in two main stages: the *retrieval stage* and the *ranking stage*. The goal of the first stage is to retrieve as many as possible relevant documents for a given query, whereas, the function of the second stage is to improve the ranking of the retrieved documents. The following section describes in detail the main components of both stages.

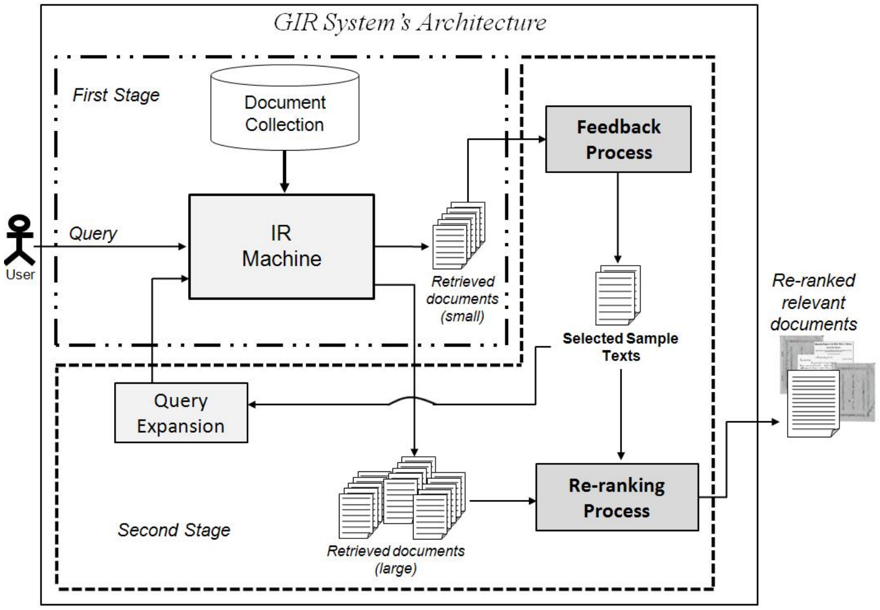

Figure 1: General architecture of the proposed GIR system

### 2.1 Modules Description

#### 2.1.1 Information Retrieval Machine

The core module of the proposed system is the information retrieval (IR) machine. In our architecture, this module is used two times: in a first iteration, it retrieves a set of relevant documents using the original given query, then, in a second iteration, it retrieves a larger set of relevant documents considering an expanded query.

For the experiments, the IR machine was implemented using the Lemur open source search engine[1], and the documents and query were preprocessed by applying stopword removal and suffix stripping.

[1] http://www.lemurproject.org/

### 2.1.2 Feedback Module

The objective of this module is to select some relevant items from the set of retrieved documents. We call these items *sample texts*, and use them for two different purposes. On the one hand, to modify the original query and perform a second IR process, and on the other hand, to re-rank the set of retrieved documents.

The implementation of this module was based on the blind relevance feedback (BRF) technique [6], which consists of selecting as relevant items the first $N$ top-ranked retrieved documents.

### 2.1.3 Query Expansion

This module takes as input the set of sample texts and extracts from them a set of relevant terms. Then, it uses these terms to expand the original query. The idea of is to provide more information to the IR machine in order to improve its performance, in particular, to improve its recall rate.

Similar to the previous module, in this case we also implemented this module following the suggestions by Chen [6]. That is, we selected as relevant terms the K most frequent terms from the sample texts.

The expanded query is sent to the IR machine, and a new set of documents is retrieved. Finally, this new set of documents is analyzed by the re-ranking process and the output of the system is generated.

### 2.1.4 Re-ranking Module

This module is the main contribution of our system. Its goal is to re-rank the set of *retrieved documents* using the information (thematic and geographic) contained in the query-related *sample texts*. Figure 2 shows the main processes of this module.

Before describing each process it is important to mention that, in theory, all sample texts are relevant to the query; nevertheless, in practice, this condition not always happens. Because of that, the final process of this module considers the integration of evidence from different sample texts.

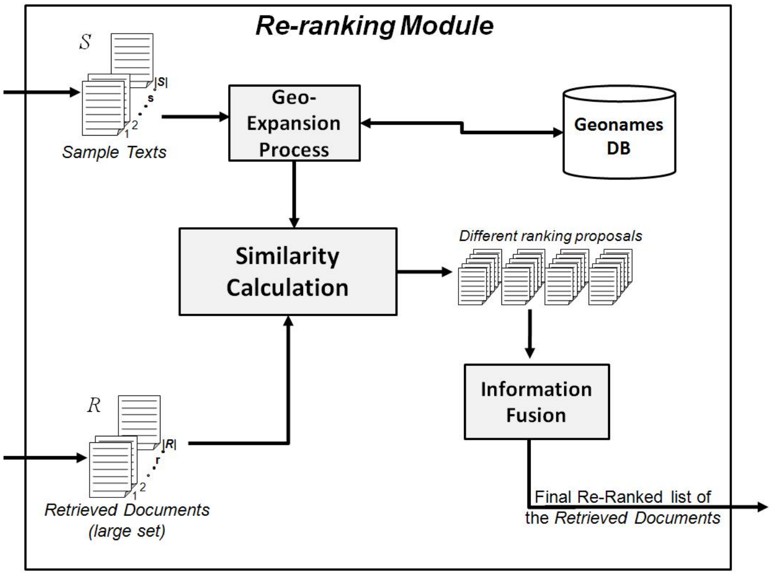

Figure 2: Zoom in to the Re-ranking Module

**Geo-Expansion.** The objective of this process is to expand the geographical terms (*geo-terms*) contained in the sample texts. Mainly, it adds to each geo-term its two nearest ancestors.

For instance, if the term *Madrid* occurs in a sample text, then the document is expanded using the terms *Spain* and *Europe*. By doing this geographic expansion, it is expected to identify a major number of query-related documents.

In order to perform the geographic expansion we employed the Geonames database[2].

**Similarity Calculation.** In this process the set of retrieved documents is compared against each one of the sample texts, generating, in this way, several different rankings of the former documents. The comparison of the documents considers both, their *thematic* and *geographic* information. In particular, the similarity between two documents is computed as indicated by formula 1[3].

$$SQ(s,r) = (\lambda \times SQ_{thematic}(s,r)) + ((1-\lambda) \times SQ_{geographic}(s,r)) \tag{1}$$

where $s$ represents a *sample text*, $r$ represents a document from the set of *retrieved documents*, and $\lambda$ is a weighting value.

In order to be able to compute the thematic and geographic similarities, it was necessary to identify all geographical named entities from the given documents. For this purpose, in our experiments we used the LingPipe tool[4]. It is also important to mention that both similarities were computed using the *cosine* formula and that the $\lambda$ coefficient was set to 0.7 in order to give more importance to the thematic similarity. This decision was based on evidence from previous GeoCLEF tracks [3, 4, 5] that indicate that the thematic part of queries and documents is more useful for the IR process.

**Information Fusion** The goal of this process is to merge, into one single list, all the information contained in the different ranking proposals. In order to generate this single ranking we employed techniques of *information fusion* [7, 8]. In particular we used the well-known Round Robin technique.

## 3 Experiments and Results

This section describes the results from a subset of our experiments evaluated at GeoCLEF 2008. For all of them, we preprocessed the documents and queries as described in Section 2.1.1, and used the following evaluation measures: Mean Average Precision (MAP), R-Precision, Precision at the first 5 documents (P@5) and Overall Recall.

Table 1 shows the results from our four baseline runs. The first two rows correspond to the results of the first IR iteration (refer to Figure 1). In this case, the run inaoe-BASELINE1 employed the title and description fields, whereas the run inaoe-BASELINE2 used all available information: title, description and narrative. As it can be seen, adding the information from the narrative did not help the IR process, on the contrary, the best performance was achieved by exclusively using the title and description fields.

The third and fourth rows of Table 1 indicate the results achieved in the second IR iteration. For these two experiments, we expanded the original query using the K most frequent terms from the top N documents retrieved by the inaoe-BASELINE1 run. We named these experiments as inaoe-BRF-N-K. As expected, the query expansion process allows both configurations to obtain better results than the BASELINE1, especially for the case of the recall rate.

Table 2 shows the results achieved by the proposed method. The results correspond to the following experiment configurations: if the name of the experiment contains the tag **RRBF**, it means that we re-ranked the retrieved documents making no distinction between the thematic and geographic parts; if the name contains the tag **RRGeo**, it means that we re-ranked the retrieved documents making a distinction between the *thematic* and *geographic* parts, but without applying

[2] http://www.geonames.org/

[3] Some research groups have used similar ways for computing the similarities among documents [9, 10]

[4] A HMM-based Named Entity Recognition System (http://alias-i.com/lingpipe/)

Table 1: Baseline results

| Experiment ID | MAP | R-Prec | P@5 | Recall |
|---|---|---|---|---|
| inaoe-BASELINE1 | 0.234 | 0.261 | 0.384 | 0.835 |
| inaoe-BASELINE2 | 0.201 | 0.226 | 0.272 | 0.815 |
| inaoe-BRF-5-2 | 0.258 | 0.267 | 0.344 | 0.863 |
| inaoe-BRF-5-5 | 0.246 | 0.264 | 0.328 | 0.863 |

any geographic expansion process; finally, if the name contains the tag **RRGeoExp**, it means that we distinguished between the *thematic* and *geographic* parts, and also that we performed a geographic expansion of the sample texts (refer to Section 2.1.4).

From the results of Table 2, we can observe that the distinction between the thematic and geographic parts allowed to obtain a better performance. It is also possible to notice that the MAP differences between the experiment tagged as RRGeo and the one tagged as RRGeoExp was not very significant. We believe that this performance happened due to the noise introduced by our nave geo-expansion process as well as by the errors from the incorrect selection of the sample texts.

Table 2: Results of the proposed approach

| Experiment ID | MAP | R-Prec | P@5 |
|---|---|---|---|
| inaoe-RRBF-5-5 | 0.241 | 0.268 | 0.384 |
| inaoe-RRGeo-5-5 | 0.244 | 0.266 | 0.384 |
| inaoe-RRGeoExp-5-5 | 0.246 | 0.270 | 0.384 |

### 3.1 Additional experiments

The goal of these additional experiments was to evaluate the effect of using only relevant *sample texts* in the re-ranking process. In order to accomplish this evaluation, these experiments considered the manual selection of the *sample texts*. In particular, we defined as *sample texts* only those relevant documents among the first five positions from the first IR iteration (i.e., from the first five documents of the inaoe-BASELINE1 run). Table 3 shows the results from these experiments. It is important to point out that in all cases the average number of *sample texts* was of two. Therefore, Table 3 can be read as:“By taking only two sample texts, and making no distinction between thematic and geographic parts, we could reach a MAP of 0.306”. These results show that the proposed method works well, but also indicate that it is very sensitive to the presence of incorrect sample texts.

Table 3: Using only relevant sample texts

| Experiment ID | MAP | R-Prec | P@5 |
|---|---|---|---|
| inaoe-RRBF | 0.306 | 0.304 | 0.496 |
| inaoe-RRGeo | 0.315 | 0.307 | 0.520 |
| inaoe-RRGeoExp | 0.318 | 0.310 | 0.536 |

## 4 Conclusions

The results from our participation at GeoCLEF 2008 showed that the use of query-related sample texts allows improving the original ranking of the retrieved documents. Nevertheless, they also showed that the proposed method is very sensitive to the presence of incorrect sample texts, and that it is also affected by the incorrect expansion of the geographical terms.

Our current work is mainly focused on tackling these drawbacks. In particular, we are working in a new sample-text selection method which is based on the idea of keeping only those documents that have a high density levels among its *geographic* and *thematic* terms, and in a new strategy for geographic expansion that considers on the one hand the inclusion of a more precise disambiguation strategy, and on the other hand, including the use of geographical coordinates.

**Acknowledgements.** This work was done under partial support of CONACyT (scholarship 165545), SNI and INAOE.

## References

[1] R. Purves, C. Jones editos : SIGIR2004 Workshop on Geographic Information Retrieval, Sheffield, UK, 2004.

[2] Andreas Henrich, and Volker Lüdecke. Characteristics of Geographic Information needs. In *Proceedings of Workshop on Geographic Information Retrieval, GIR'07*, Lisbon, Portugal, 2007. ACM Press.

[3] F. Gaey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *the 6th Workshop of the Cross-Language Evaluation Forum*, CLEF 2005, Viena, Austria, 2005.

[4] F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, and P. Rocha. GeoCLEF 2002: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In *the 7th Workshop of the Cross-Language Evaluation Forum*, CLEF 2006, Alicante, Spain, 2006.

[5] T. Mandl, F. Gey, G. Di Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In *the 8th Workshop of the Cross-Language Evaluation Forum*, CLEF 2007, Budapest, Hungary, 2007.

[6] A. Chen. Cross-Language Retrieval Experiments at CLEF 2002, In *Springer,LNCS No.2785*, pages 28-48, 2003.

[7] J. Lee. Analysis of multiple evidence combination. In *20th anual ACM SIGIR Conference*, 1997.

[8] W. C. Lin, and H. H. Chen. Merging mechanisms in multilingual information retrieval. In *Working Notes CLEF 2002*, Rome, Italy, 2002.

[9] Geoffrey Andhogah, and Gosse Bouma. University of Groningen at GeoCLEF 2007. In *Working notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007

[10] B. Martinis, and N. Cardoso, and M. Silveira-Chavez, and L. Andrade, and M. J. Silva. The University of Lisbon at GeoCLEF 2006. In *Working notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.