SINAI at NTCIR-9 GeoTime: a filtering and reranking approach based solely on geographical entities

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega
and L. Alfonso Ureña-López
SINAI research group,Computer Science Department, University of Jaén
Escuela Politécnica Superior, A-3, Paraje Las Lagunillas
23071, Jaén, Spain
{jmperea,magc,mgarcia,laurena}@ujaen.es

ABSTRACT

Geographic Information Retrieval (GIR) is an active and growing research area that focuses on the retrieval of textual documents according to a geographical criteria of relevance. In recent years, the IR research community has paid particular attention in IR systems that take into account temporal constraints. Temporal Information Retrieval (TIR) is a recent field which addresses the combination of usual IR techniques with new ones for addressing temporal dimension of relevance. The NTCIR-GeoTime task was created to evaluate IR systems that combine geographical and temporal constraints. In this work, we propose a filtering and reranking function for these type of systems based on the retrieval status value calculated by the IR engine and the geographical similarity between the document and the query. Due to we have only considered the geographical criteria, the obtained results show that the proposed function does not improve the baseline experiment applying solely an IR approach. Therefore, it is necessary to improve the proposed reranking function taking into account the temporal entities found in the document collection and topics.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval - retrieval models, search process

General Terms

Experimentation, Performance, Measurement

Keywords

Geotemporal Information Retrieval, Geographical Information Retrieval, Information Retrieval

`[SINAIUJAEN]`
`[English GeoTime]`
`[Terrier][Geonames][Wikipedia][Illinois NE Tagger]`

1. INTRODUCTION

Many of the searches performed to Information Retrieval (IR) engines in the web contain both geographical and time elements. Nowadays, this combination adds a real challenge to the community responsible for improving or evaluating these systems, since they have to take into account not only the geographical constraint motivated by the geographical scope detected in the query ("*in Africa*", "*South American country*"), but also the complexity of temporal constraints such as "*in 1978*" or "*between 1990 and 1992*". In this way, the GeoTime evaluation task emerged in 2010 as a track of the NTCIR Workshop [4] with the aim of evaluating retrieval techniques focusing in geographical and temporal constraints.

In the field of Geographical Information Retrieval (GIR), a geographical query is structured as a triplet of

$$<theme><spatial\ relationship><location>$$

where *theme* is the main subject of the query, *location* represents the geographical scope of the query and *spatial relationship* determines the relationship between the subject and the geographical scope. For example, the triplet for the geographical query "*airplane crashes close to Russian cities*" would be

$$<airplane\ crashes><close\ to><Russian\ cities>$$

GIR is concerned with improving the quality of geographically specific information retrieval with a focus on access to unstructured documents [7]. 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 within Spain. The use of geographical constraints in queries has been previously explored in the GeoCLEF track [6, 9] between 2005 and 2008, as a part of the Cross Language Evaluation Forum (CLEF) campaigns[1].

On the other hand, Temporal Information Retrieval (TIR) is a recent field which addresses the combination of usual IR techniques with new ones for addressing temporal dimension of relevance [8]. As is well known, temporal elements are present everywhere and, therefore, every temporal context contains a fragment of information with an important value for IR purposes. It has been shown that taking into account the value of temporal information, it can improve retrieval systems [1]. In brief, the major challenge in TIR is to understand the meaning of the temporal expressions present in the text with the aim of improving the results of the search process.

This is the first participation of the SINAI[2] research group in the NTCIR-GeoTime evaluation task. Previously, related to the GIR field, we have participated in the GeoCLEF track for three years [3, 10, 13]. The main objective of this work is to analyze the behavior of our GIR system used in GeoCLEF

[1] `http://clef-campaign.org`
[2] `http://sinai.ujaen.es`

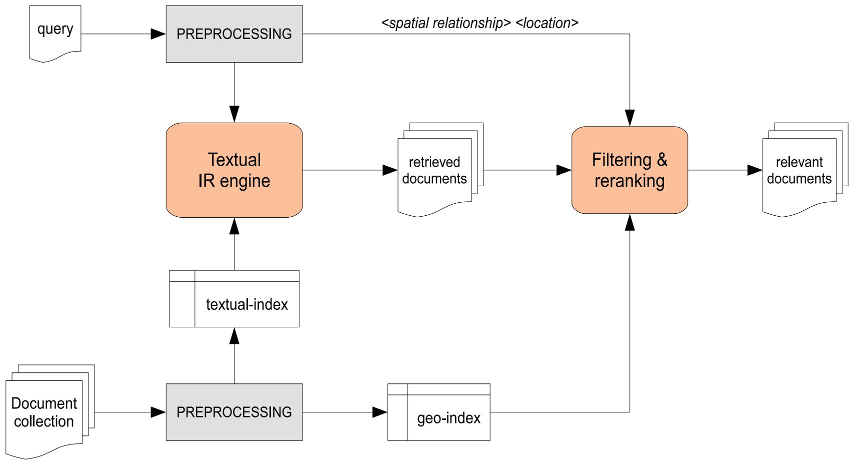

Figure 1: Overview of the GIR system used by SINAI team in the NTCIR-9 GeoTime task

for the GeoTime task. Therefore, in this first approach, the temporal information is not taken into account, since we have only focused on geographical entities detected in documents and queries. In Section 2, we describe the main components of our system. Then, in Section 3, the experiments carried out and the results along with a brief analysis are shown. Finally, in Section 4, the conclusions and future work are expounded.

2. SYSTEM OVERVIEW

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 system presented 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 geographical scope and the spatial relationship that may contain. On the other hand, the document collection is also preprocessed, detecting all the geographical 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 geographical entities) is run against the search engine in order to obtain 1,000 relevant documents. Finally, such documents are filtered and reranked, setting in the last positions of the final list, those documents that do not match with the geographical scope detected in the query. By contrast, those documents that fit the geographical scope detected, are set in the first positions. The final ranking of each relevant document will depend on the number of geographical entities that it contains and fit the geographical constraint of the query, increasing the Retrieval Status Value (RSV) assigned by the IR engine. For example, if the geographical entity detected in the query is "*Africa*" and the spatial relationship is "*in*", then those documents that contain a country or city that belongs to Africa are setting in the first positions of the final list, while those documents that do not contain any geographical entity belonging to Africa are filtered and placed in the last positions of the ranking. Next, we explain in more detail the main processes followed in our system.

2.1 Document and query preprocessing

The preprocessing carried out with the queries was mainly based on detecting their geographical entities. This also involves specifying the triplet explained in Section 1, which will be used later during the filtering and reranking process. To detect such triplet, we have used a Part Of Speech tagger (POS tagger) like TreeTagger[3] [14], taking into account some lexical syntactic rules such as *preposition + proper noun*. Moreover, the stop words were removed and the Snowball stemmer[4] was applied to each word of the query, except for the geographical entities. For example, the triplet obtained for the query "*When and where did anti-government demonstrations occur in Uzbekistan?*" was

<anti-government demonstrations><in><Uzbekistan>

On the other hand, a similar offline preprocessing was carried out with the document collection. During this process, two textual indexes were generated:

- a **geographical index**, which contains the locations detected in each document. We have used Geo-NER[12] to recognize geographical entities in the collection and queries. Geo-NER is a Named Entity Recognizer (NER) for geographical entities based on Wikipedia and GeoNames[5].
- a **textual index**, which contains the preprocessed text (*stemmer* and *stopper*) of each document, including the geographical entities in their original form, i.e, without applying stemmer to them.

[3] TreeTagger v.3.2 for Linux. Available in http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[4] Available in http://snowball.tartarus.org
[5] http://www.geonames.org

It is important to note that we consider several geographical scopes for a document, as many as geographical entities have been detected in it. Therefore, in the geographical index we can have different locations for the same document.

2.2 Indexing and retrieval

In 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 [11], 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 with Laplace after-effect and normalization two. In addition, an automatic query expansion process called *Bose-Einstein (Bo1)* has been applied during the retrieval phase. Such function is also included in the Terrier tool by default. *Bo1* assign a score to each candidate expansion term, evaluating its importance by calculating the divergence of its distribution in a pseudo-relevance document set from a random distribution [2].

2.3 Filtering and reranking

The documents retrieved by the IR engine are used as a input in the filtering and reranking process, which is reponsible for modifying the RSV score of each document depending on the geographical similarity with the query. The geographical similarity between a document and a query is calculated using the following formula:

$$sim_{geo}(Q,D) = \frac{\sum\limits_{i \in geoEnts(D)} match(i,GS,SR) \cdot freq(i,D)}{|geoEnts(D)|} \tag{1}$$

where the function $match(i, GS, SR)$ returns 1 if the geographical entity i satisfies the geographical scope GS for the spatial relationship SR and 0 otherwise. $freq(i, D)$ means frequency of the geographical entity i in document D, and $|geoEnts(D)|$ represents the total number of geographical entities detected in the document D.

To explain the performance of the *match* function, we can use the following query of the GeoTime 2011 task: "*When and where did Hurricane Katrina make landfall in the United States?*". In this case, it is a geographical 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 geographical entity i (for example, *New York*) in a retrieved document (D) which belongs to United States, then $match(NewYork, UnitedStates, in) = 1$. If the geographical entity did not belong to the geographical 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 geographical entity i of the document, the geographical scope (GS) of the query and the spatial relationship (SR) recognized 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.

Regarding the reranking procedure, if $sim_{geo}(Q, D) = 0$, then the document D is filtered or discarded and keeps the RSV score assigned by the IR engine. On the other hand, for those documents that their geographical similarity is greater than zero, the filtering and reranking process modifies the RSV of the retrieved document (*RSV(D)*) taking into account its geographical similarity ($sim_{geo}(Q, D)$) and its previous RSV using the following formula:

$$RSV'_D = RSV_D + log(RSV_D) + sim_{geo}(Q,D) \tag{2}$$

3. EXPERIMENTS AND RESULTS

Although the collections provided by the GeoTime 2011 organizers were the same as those used in the previous year (Japanese and English collections), they were expanded significantly by adding documents from earlier NTCIR workshops. This expansion was carried out both in size and time coverage. More detailed information about the data collection and the topic development can be found in [5].

In order to evaluate the proposed approach we have only used the English collection of the NTCIR9-GeoTime task. Our main goal in this work was to compare the behavior of the filtering and reranking process regarding the baseline case in which no filtering or reranking is applied. Therefore, our baseline experiment can be seen as a simple IR approach, without taking into account any extra processing, except the typical ones in IR approaches (stemming and removing stopwords). The experiments proposed in this work are the following:

- **SINAIUJAEN-EN-01-D**: baseline IR with Terrier, without applying any filtering or reranking process, and only using the content of the label "*description*" (DESC) from the topics.
- **SINAIUJAEN-EN-02-DN**: the same configuration as the previous experiment but using the content of the labels DESC and "*narrative*" (NARR) from the topics.
- **SINAIUJAEN-EN-03-D**: application of the filtering and reranking process over the list of documents retrieved by the SINAIUJAEN-EN-01-D experiment.
- **SINAIUJAEN-EN-04-DN**: application of the filtering and reranking process over the list of documents retrieved by the SINAIUJAEN-EN-02-DN experiment.

The results obtained for each proposed experiment are shown in Table 1. The Mean Average Precision (MAP) has been used as main evaluation metric. We also show the results applying other evaluation metrics such as *Q-measure* and *normalized Discounted Cumulative Gain* (nDCG@1000). More information about both metrics can be found in [**?**].

Experiment	MAP	Q	nDCG@1000
01-D	0.4341	0.4564	0.6587
02-DN	**0.4759**	0.4983	0.6941
03-D	0.4266	0.4514	0.6505
04-DN	0.4611	0.4898	0.6824

Table 1: Experiments and results of the SINAI research group in the NTCIR9-GeoTime English task

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

As can be observed in Table 1, the results obtained applying the filtering and reranking process do not improve in any case those obtained using the baseline experiments regarding the main evaluation metric (MAP). The percentage of difference is similar between using the DESC and NARR labels or only using the DESC label from topics. The application of the filtering and reranking process does not reach the results obtained by the baseline experiments with a difference of -3.21% when we use the DESC and NARR labels and -1.76% when only the DESC label is used. The main reason of this behaviour is due to the fact of not taking into account the temporal expressions in the filtering and reranking process. Therefore, many documents that have not been filtered, are reranked in top positions because they contain some geographical entities that match with the geographical scope detected in the query but probably do not fit with the temporal constraint recognized in the query. In this sense, the application of a temporal entity recognizer in the preprocessing of document collection and topics, may improve the results obtained by the filtering and reranking process.

4. CONCLUSIONS AND FUTURE WORK

In this work we propose a simple filtering and reranking process in order to improve the baseline IR approach followed to solve the NTCIR9-GeoTime English task. This process is based on the RSV score calculated for the IR engine and the geographical similarity between the document and the geographical scope detected in the query. The comparison of the obtained results shows that the proposed reranking approach does not improve the results obtained with the baseline experiments. The main reason of this behaviour is due to the fact of not taking into account the temporal expressions in the reranking process, althought considering solely the geographical entities in the document collection, the results are not so bad.

Therefore, for future work, we will try to improve the reranking function considering the temporal expressions in both, document collection and topics, and analyzing the type of the geographical constraint in the query, since for some queries the geographical scope is more bounded than for others.

5. ACKNOWLEDGMENTS

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), projects TEXT-COOL 2.0 (TIN2009-13391-C04-02) and Red Temática TIMM: Tratamiento de Información Multimodal y Multilingüe (TIN2009-06135-E) from the Spanish Government, a grant from the Andalusian Government, project GeOasis (P08-TIC-41999) and Geocaching Urbano research project (RFC/IEG2010).

6. REFERENCES

[1] O. Alonso, M. Gertz, and R. A. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

[2] G. Amati. *Probabilistic models for information retrieval based on divergence from randomness*. PhD thesis, University of Glasgow, 2003.

[3] M. García-Vega, M. A. García-Cumbreras, L. A. Ureña-López, and J. M. Perea-Ortega. GEOUJA System. The First Participation of the University of Jaén at GEOCLEF 2006. In *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 913–917. Springer, 2006.

[4] F. C. Gey, R. R. Larson, N. Kando, J. Machado, and T. Sakai. NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *Proceedings of NTCIR-8, Tokyo, Japan*, pages 147–153, 2010.

[5] F. C. Gey, R. R. Larson, J. Machado, and M. Yoshioka. NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2. In *Proceedings of NTCIR-9, Tokyo, Japan*, 2011.

[6] F. C. Gey, R. R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 908–919. Springer, 2005.

[7] C. B. Jones and R. S. Purves. Geographical Information Retrieval. In *Encyclopedia of Database Systems*, pages 1227–1231. Springer US, 2009.

[8] J. Machado, J. Borbinha, and B. Martins. Experiments with Geo-Temporal Expressions filtering and query expansion at document and phrase context. In *Proceedings of NTCIR-8, Tokyo, Japan*, pages 159–166, 2010.

[9] T. Mandl, P. Carvalho, G. M. D. Nunzio, F. C. Gey, R. R. Larson, D. Santos, and C. Womser-Hacker. GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF*, volume 5706 of *Lecture Notes in Computer Science*, pages 808–821. Springer, 2008.

[10] J. M. Perea-Ortega, M. A. García-Cumbreras, M. García-Vega, and L. A. Ureña-López. Filtering for improving the geographic information search. In *CLEF*, volume 5152 of *Lecture Notes in Computer Science*, pages 823–829. Springer, 2007.

[11] J. M. Perea-Ortega, M. A. García-Cumbreras, M. García-Vega, and L. A. Ureña-López. Comparing several textual information retrieval systems for the geographical information retrieval task. In *NLDB*, volume 5039 of *Lecture Notes in Computer Science*, pages 142–147. Springer, 2008.

[12] J. M. Perea-Ortega, F. Martínez-Santiago, A. Montejo-Ráez, and L. A. Ureña-López. 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.

[13] J. M. Perea-Ortega, L. A. Ureña-López, M. García-Vega, and M. A. García-Cumbreras. Using query reformulation and keywords in the geographic information retrieval task. In *CLEF*, volume 5706 of *Lecture Notes in Computer Science*, pages 855–862. Springer, 2008.

[14] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.