DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.16

Different tools for handling Geographic Information Retrieval problems

Yisleidy Linares Zaila
PhD. Student Computer Science and Engineering
University of Bologna
Mura Anteo Zamboni 7, BO 40138
Italy
yisleidy.linares2@unibo.it

Usually in natural language several amounts of geographic references are found, this comes with the common necessity of giving a context with time and location details. Considering these locational semantics, user queries may be satisfied in a more accurate way. In this work a geo-ontology is built for identifying geographic terms. A toponym disambiguation algorithm is proposed for assigning to a place name its corresponding location. The importance of geographic terms in documents is determined by means of a weighting strategy, that is also used to compare geographic contents and to provide a ranking of results. It is also provided a technique for combining a standard textual ranking and the obtained geographic ranking. Final results are evaluated using GeoCLEF test collection and baseline techniques.

Keywords: Geographic information retrieval, Toponym Disambiguation, Spatial similarity measure

1. INTRODUCTION

It is estimated that more than 70% of all information in the world has some kind of geographic features (Jones et al. (2004)). Considering that users queries with geographic references are very natural, the development of search engines aware of geographical semantics has received lots of attention in both the academic and the industrial aspects.

According to Abdelmoty et al. (2005) one definition of Geographic Information Retrieval is the provision of facilities to retrieve and rank by relevance documents or other resources from an unstructured collection, on the basis of queries specifying both theme and geographic scope. This definition carries some main challenges such as:

i. Identification of geographic terms in documents and associating these terms with appropriate geographic locations: A common problem in GIR is that different locations may be named in the same way, this problem is called *toponym disambiguation*. Typical approaches to toponym disambiguation include using knowledge-based , map-based, data-driven methods, or a combination of the three (Buscaldi (2011)). Knowledge-based and data-driven approaches usually incorporate toponym-related information which is used to derive disambiguation rules (e.g. SPIRIT Jones et al. (2004)) and to train machine learning algorithms (Martins and Calado (2010)) respectively. Map-based methods assign geographic locations to toponyms by taking into account the spatial distribution among them (Leidner (2004, 2007)).

ii. Development of spatial similarity measures for comparing geographic information: The similarity between geographic terms is often approximated by geometric or geographic measurements, such as Euclidean distance, overlap or direction (Larson and Frontiera (2004)).

iii. Techniques to properly combine geographic and thematic relevance: A known technique combines documents from different rankings according to their scores, showing that outperforms other approaches in a IR context, as well as in a GIR context (Lee (1997); Palacio et al. (2010)).

In this work we proposed the GeoNW ontology, which is based on GeoNames, WordNet and Wikipedia resources. This stores all geographic knowledge that is used for extracting, analysing and comparing the geographic content present in documents and queries. The toponym disambiguation problem is handled by using the information

stored in GeoNW. It is based on the assumption that in a document, the geographic location with more neighbours also present has a higher probability to be the one actually referenced. A weighting strategy for quantifying the degree of influence of a geographic term over a document is also proposed. It is based on the frequency of the geographic term in the document and on its hierarchical and topological relation with the other geographic terms found in the text. Queries are similarly processed and a new spatial matching function to measure the geographic similarity between a document and a query is applied. It uses the results obtained from the weighting strategy. Moreover, we define a strategy for combining standard textual similarity measures with our geographical similarity measure. The proposed algorithm classifies as more relevant documents those with higher rank in both textual and geographic rankings. As a final result a unique ranked list of documents is obtained, expecting it better satisfies the user needs. Finally, the approach is evaluated using GeoCLEF (Mandl et al. (2009)) test collection and results are compared with baseline techniques.

2. GIR TOOLS

A GIR architecture can be seen as a model that separately analyses thematic and geographic information present in texts. The output is a ranked list of documents which are relevant to a specific user query (Figure 1).

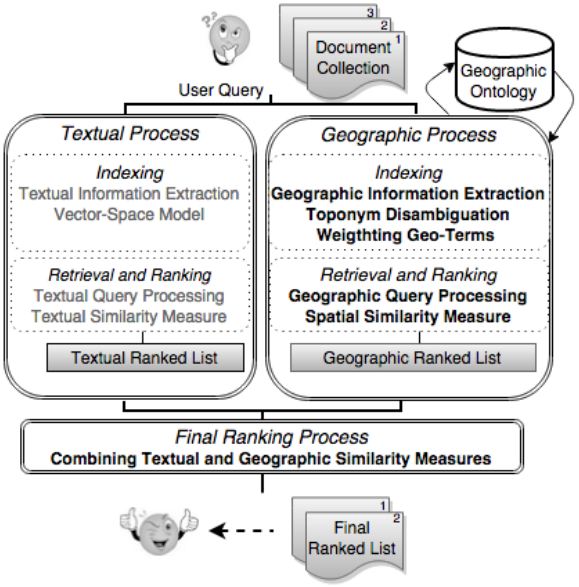

Figure 1: *GIR architecture*

In this work, tasks corresponding to *Textual Process* are achieved using standard IR techniques. The toponyms are extracted and disambiguated using GeoNW ontology. Then, to each geographic term in a document a weighting value that represents its importance in the document is assigned. The geographic information between a query and a document is compared through a spatial similarity measure obtaining a geographic ranking list of documents which is combined with a textual ranking list providing the final result that is showed to the user.

2.1. GeoNW

GeoNW is the geographic ontology, which is built using three different sources: GeoNames, WordNet and Wikipedia. One of the main reasons for building a new geographic knowledge source was that existing databases such as GeoNames refer to several places with the same geographic location (i.e. *Boston*) increasing the ambiguity problem. It means, there are many entries corresponding to the same place.

The structure of GeoNW is shown in Figure 2. Each physical or administrative place has associated a set of synonyms that were obtained from the alternateNames in GeoNames and from synsets in WordNet. This relation allows to recognise *capital of Spain* as *Madrid* or *the city of Light* as *Paris*. Also, each place has its geographic coordinates, which were provided from GeoNames. Furthermore, using Wikipedia a set of nationalities adjectives were added to administrative places, allowing to relate *Italian rivers* with *rivers in Italy*.

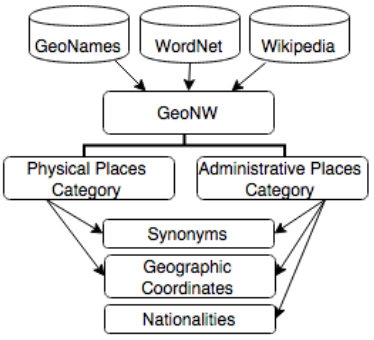

Figure 2: *GeoNW*

2.2. Toponym Disambiguation

According to Leidner (2007), toponym ambiguity can be classified in: i) morpho-syntactic ambiguity (e.g. *nice* the English adjective and the city in France); ii) feature type ambiguity (e.g. *Mississippi* the State of USA and *Mississippi* the river); iii) referential ambiguity (e.g. *London* the city in England, UK

disambiguation technique is based on this idea of ambiguity classification.

The goal of the disambiguation algorithm is to assign to a place name a unique geographic element. Thus, each step attempts to remove elements from the candidate list in order to take the one more related to the content. The technique uses the information stored in GeoNW and assumes that the more related a geographic entity is to other geographic terms in the document, the more probable to be referenced it will be.

If the place name can be a physical place (i.e. river, mountain) or an administrative place (i.e. country, city, town), a Feature Type Disambiguation strategy is applied. It is based on the candidate feature type[1] frequency in the document. The technique takes the geographic reference that corresponds to the more mentioned feature type.

The case when we have the same toponym associated to different geographic entities is solved through the Referential Disambiguation technique. In this case, ambiguity is only related to administrative places. The algorithm uses the hierarchical relationships among the ambiguous place name and the other geographic locations present in the document. This hierarchical relationship is obtained from GeoNW. The method chooses the place which its geographic location is closer, in the hierarchy, to all other places present in the document.

2.3. Document Geographic Focus Detection

The geographic focus is represented as a set of pairs $< w_i, g_i >$ where g_i is a geographic term and w_i is a numeric value that represents its influence over the document. It uses the relationships among all georeference in the document. It is also based on the principle that more relatives a geographic term has in a document, more the geographic term is associated to the document.

The weighting value assigned to each geographic term in a document d_i is computed by the expression:

$$\frac{freq(gt)}{maxG} * \left(TI(gt, d_i) + DI(gt, d_i) + \frac{1}{|dG_i|} \right)$$

where TI (Topologic Influence) and DI (Distance Influence) are functions that evaluate topological and metrical relationships respectively, $freq(gt)$ is the frequency of gt in d_i, $maxG$ is the maximum value reached by the expression above and $|dG_i|$ is the number of geographic elements in d_i. TI function is mainly based on the number of common ancestors between two geographic references, while DI is based on the geographical distance between two locations that have at least one common ancestor. If two geographic terms do not have any common ancestor, DI output is infinity. The latter avoids the negative effect of not related geographic terms that are too far from gt.

[1]Feature types correspond to physical places, such as rivers, mountains, lakes, etc.

2.4. Query Geographic Focus Detection

Due to, queries are usually composed by a short number of words, and usually have only one geographic term, the disambiguation process can not be the same used for documents. Given *cList*, defined as the list of all possible alternatives of an ambiguous place name *tp*. The list $cList_w$ is defined as:

$$cList_w = \{ < w_1, c_1 >, ..., < w_n, c_n > \}$$

where $c_i \in cList$ and w_i is the ratio of the number of times that *tp* matches with c_i and the total frequency of c_i in the collection D. Notice that $cList_w$ is built during the geographic indexing and it depends on the toponym disambiguation process explained above.

2.5. Spatial Similarity Measure

The proposed technique for assigning a score to a document according to a query is very intuitive. It is directly based on the results of processing the geographic focus of the query Q and documents $d_i \in D$. It can be seen as a combination of distance and topological methods, because it is strongly related to the relationships among the geographic terms and the geographical distance among them. Let QG and dG_i be the sets that represent the geographic focus of the query Q and the document d_i respectively, the spatial score function S_G is defined as:

$$S_G(QG, dG_i) = \sum_{n=1} \sum_{m=1} (w_n * w_m)_{c_n = g_m}$$

where c_n is a geographic entity present Q and g_m is a geographic entity present in d_i. Computing this expression for all documents in D, the geographic ranked list that corresponds to Q is obtained.

2.6. Combining Textual and Spatial Similarity Measure

Let R_T and R_G be the textual ranked list and geographic ranked list of the query Q in the collection D. R_T and R_G contains normalised values in a range of [0, 1], where 1 is the score of the most relevant document in D to the query Q. The CombTG function is defined as:

$$CombTG(R_T, R_G) = \sum_{i=1} \beta * (st_i + sg_i)$$

where st_{π_i} and sg_{φ_i} are the textual and geographic score of document d_i and β takes value 2 if d_i is in R_T and R_G, 1 if d_i is only in R_T and 0.5 if d_i is only in R_G. This strategy benefits documents retrieved in both lists and penalizes those that were retrieved only by their geographic information. It is based on the assumption that documents whose textual information is not relevant to the query will have less importance than those that do are.

3. EVALUATION

The algorithms are implemented as an extension of Terrier[2] tool. The evaluation process is based on GeoCLEF[3] test collection. There are 166 477 documents and 25 geographic topics. GeoCLEF documents were extracted from a set of articles from *The Los Angeles Times (1994)* and *The Glasgow Herald (1995)*.

A first evaluation of our approach was made by comparing with Terrier baseline approaches. The best results of Terrier baseline algorithms were obtained by using *DLH* 13 (Lu et al. (2013)) and *LGD* (Clinchant and Gaussier (2011)) models. Also we use the *BM* 25 approach as it is a well known standard technique for retrieval applications. In Table 1, the first three rows correspond to the results obtained by the application of standard IR techniques without geographic analysis, while the last three rows show the results using these techniques for textual analysis and the proposed strategy for geographic analysis. As a promising result we can see that our approach improves all *M AP* values using for textual processing any of the weighting models mentioned above.

Model	Recall	MAP
BM 25	0.905	0.378
DLH 13	0.908	**0.404**
LGD	**0.919**	0.394
CombTG_BM 25	0.902	0.465
CombTG_DLH 13	0.902	0.487
CombTG_LGD	**0.909**	**0.489**

Table 1: *Overall results*

On the other hand, results reported by GeoCLEF 2008 track overview Mandl et al. (2009) using MAP evaluation measure are around 0.3, thus our strategy reaches better results.

[2]http://terrier.org/
[3]http://www.uni-hildesheim.de/geoclef/

3.1. Toponym Disambiguation Strategy Evaluation

Toponym disambiguation is one of the problems we attempt to solve in this work. For evaluating the proposed technique we compare with a naive strategy that identifies as the geographic location of an ambiguous place name the one with the largest population. In Figure 3 *CombTG_BM* 25_*pop*, *CombTG_DLH* 13_*pop* and *CombTG_LGD_pop* correspond to the behaviour of our approach using for textual analysis *BM* 25, *DLH* 13 and *LGD* weighting models respectively and for disambiguating toponyms the geographic location with the largest population; while *CombTG_BM* 25, *CombTG_DLH* 13 and *CombTG_LGD* correspond to the same strategies but using the proposed disambiguation technique. As it is shown for all cases using our disambiguation algorithm, a slightly better result is obtained.

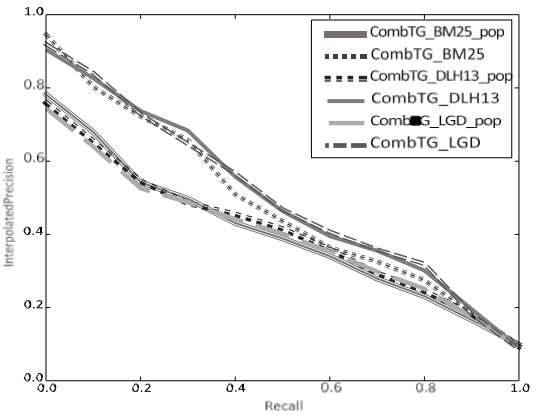

Figure 3: *Evaluating Toponym Disambiguation Strategy*

Although a more precise evaluation of the toponym disambiguation technique is needed in order to compare with other toponym disambiguation strategies, these preliminary results show that our approach produces a positive effect on the final result.

4. CONCLUSIONS

This paper describes a new approach for retrieving and ranking documents according to textual and geographic information. It uses standard IR techniques for processing textual information, while the geographic information is analysed using the geographic ontology GeoNW. A toponym disambiguation method is proposed, which according to the experiments, improves the naive technique. This paper also proposes a weighting strategy in order to quantify the influence of a geographic entity over a document. It is based on topology and distance rela-

test collection, the strategy is evaluated and results outperform baseline techniques. The best result is achieved by combining the *LGD* model with the proposed geographical analysis, obtaining a mean average precision of 0.489. Currently we are working on a more exhaustive evaluation of our strategy comparing the results with other approaches. We are also planning to include query expansion techniques in order to continue improving the overall results.

ACKNOWLEDGEMENTS

I thank my supervisor Danilo Montesi for his support in the development of this work. Special thanks also to ELIAS (Evaluating Information Access Systems) ESF Research Networking Programme for the awarded scholarship to participate in the 10th European Summer School in Information Retrieval (ESSIR 2015) and the FDIA Symposium.

REFERENCES

Abdelmoty, A. I., P. D. Smart, C. B. Jones, G. Fu, and D. Finch (2005). A critical evaluation of ontology languages for geographic information retrieval on the internet. In *Journal of Visual Languages and Computing*, Volume 16, pp. 331–358. Elsevier.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special 3*(2), 16–19.

Clinchant, S. and É. Gaussier (2011). Retrieval constraints and word frequency distributions a log-logistic model for ir. *Information retrieval 14*(1), 5–25.

Jones, C. B., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid (2004). The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science, Third International Conference, GIScience, Adelphi, MD, USA*, Volume 3234 of *Lecture Notes in Computer Science*, pp. 125–139. Springer.

Larson, R. R. and P. L. Frontiera (2004). Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Volume 3232 of *Lecture Notes in Computer Science*, pp. 45–56. Springer-Verlag.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *ACM SIGIR Forum*, Volume 31, pp. 267–276. ACM.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, Sheffield, UK*.

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph. D. thesis, Institute of Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Lu, S., B. He, and J. Xu (2013). Hyper-geometric model for information retrieval revisited. In *Information Retrieval Technology*, Volume 8281 of *Lecture Notes in Computer Science*, pp. 62–73. Springer Berlin Heidelberg.

Mandl, T., P. Carvalho, G. Di Nunzio, F. Gey, R. Larson, D. Santos, and C. Womser-Hacker (2009). Geoclef 2008: The clef 2008 cross-language geographic information retrieval track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access*, Volume 5706 of *Lecture Notes in Computer Science*, pp. 808–821. Springer Berlin Heidelberg.

Martins, B. and P. Calado (2010). Learning to rank for geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pp. 21. ACM.

Palacio, D., G. Cabanac, C. Sallaberry, and G. Hubert (2010). On the evaluation of geographic information retrieval systems. *International Journal on Digital Libraries 11*(2), 91–109.