University of Pittsburgh at GeoCLEF 2008: Towards Effective Geographic Information Retrieval

Qiang Pu[1], Daqing He[2], Qi Li[2]
[1]School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, CHINA
puqiang@uestc.edu.cn

[2]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA
{dah44,qil14}@pitt.edu

Abstract. This paper reports University of Pittsburgh's participation in GeoCLEF 2008. As the first time participants, we only worked on the monolingual GeoCLEF task and submitted four runs under two different methods. Our GCEC method aims to test the effectiveness of our online geographic coordinate extraction and clustering algorithm, and our WIKIGEO method wants to examine the usefulness of using the geo-coordinate information in Wikipedia for identifying geo-locations. Our experiments results show that: 1) our online geographic coordinate extraction and clustering algorithm is useful for the type of locations that do not have clear corresponding coordinates; 2) the expansion based on the geo-locations generated by GCEC is effectiveness in improving Geographic retrievals. 3) Using Wikipedia we can find the coordinates for many geo-locations, but its usage for query expansion still need further studies. 4) query expansion based on title only obtained better results than using the combination of title and narrative parts, which are thought to contain more related geographic information. Further study is need for this part too.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

General Terms
Experimentation, Performance

Keywords
Geographic Coordinate Extraction and Clustering, Geographic Information Retrieval, Query Expansion

1 Introduction

Along with the rapidly developed Web technologies and services, Web users' queries increasingly contain geographic information. It is, therefore, important for Web search engines to be able to recognize the geographic information, and expand it with more concrete locations if the initial geographic information is inaccurate. This is the motivation that our research team participated in GeoCLEF 2008.

As the first time participants, we only worked on the monolingual GeoCLEF task and submitted four runs under two different methods for extracting geographic location information for query expansion. The first one is Geographic Information Retrieval with Geographic Coordinates Extraction and Clustering (GCEC). Its basic idea is that those locations in the same cluster with the original geographic location should be treated as the geographic approximations of the location which can be used for geographic query expansion. The second method is Wikipedia-based Geographic Information Retrieval (WIKIGEO). Geographic location names were mined from Wikipedia - the online encyclopedia which provides abundant types of knowledge. We also assume

that a query in our geographic information retrieval task can be segmented into a topic part, a geo part and the relation part that separate the topic part from the geo part.

In the remaining of the report, we will first present in details the two methods we developed, then we talk about the experiments and the initial results for GeoCLEF2008. We will conclude with some discussion about the methods.

2 Method 1: GeoIR with Geographic Coordinates Extraction and Clustering

2.1 Overview

The first method of our Geographic Information Retrieval (GeoIR) utilizes geographic coordinates and clustering methods. The system built based on this method, call GCEC system, consists of four main functional modules listed below.

- **Query Pattern Analysis (QPA)**: this module takes in charge of query parsing. It will split the original query into two parts: *geo-part* and the remaining, called *topic-part*, according to some rules.
- **Geo Part Process (GPP)**: this is the most important module in GCEC system. It is responsible for expanding geographic terms in the original query.
- **Topic Part Process (TPP)**: this module utilizes an online Chinese-English dictionary to extract the synonyms of the topic query terms in original query.
- **Candidate Term Selection (CTS)**: this module uses the global collection statistics to filter out some useless or noisy candidate terms that are obtained by the TPP and the GPP modules. Its output is the expanded query.

Figure 1 shows the architecture of GCEC system, and we will talk about the modules in details in Section 2.2.

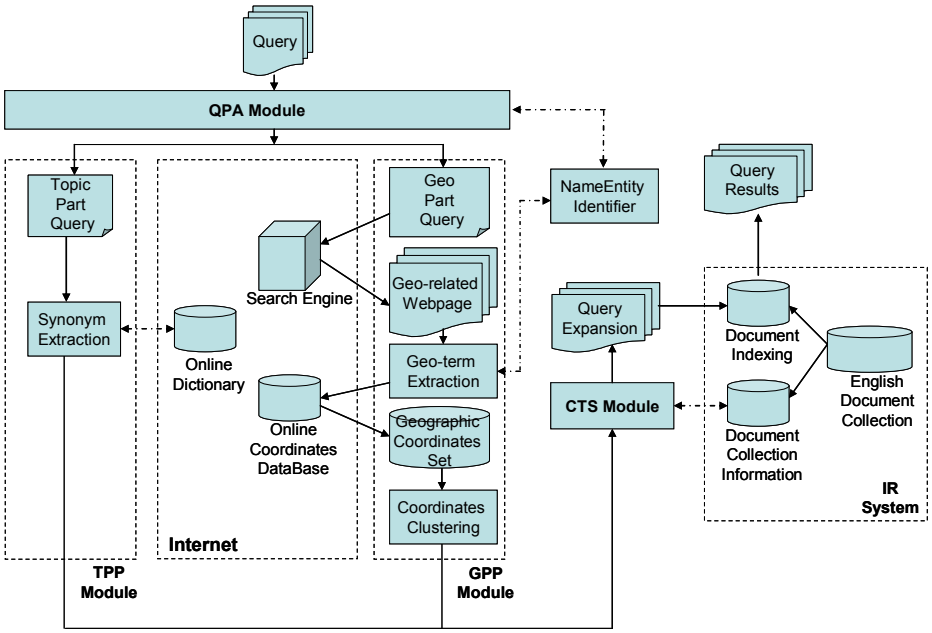

Figure 1. the Architecture of GCEC system

2.2 Main Module Description

2.2.1 QPA Module

In this paper, we are interested in the type of queries that have three clear segments of information. There is one segment called the topic part that indicates the topic that the user wants to search for. Then there is another segment explains the geographic place that a user wants to know, and it is called the geo part in this paper. Between these two parts, we assume that there always exists a certain kind of relationship between the topic part and the geo part. Often are prepositions, the relationship part acting as the boundary that separates the topic part and the geo part. Therefore, a location related query can always be clearly split into a topic part, a relationship part, and a geo part. For example, in query "Riots in South American prisons", preposition "in" splits the query into two parts, the left part "Riots" is the topic part, and the right part "South American prisons" is the geo part. This is similar to the segmentation of queries into "what", "relation-type", and "where" part in some related work [3] . The set of prepositions that can be used for the relation is obtained from [3].

We acknowledge that this view of queries could be too rigid, but it does help parsing queries easier. We hope that modules like GPP, TPP and CTS could compensate for the insufficient of query parsing.

Before the parsing of queries, it is necessary to perform a few preprocessing:

- All possessive geographic terms are transformed into non-possessive forms so that the corresponding geographic name entity can be identified correctly. For example, "American" is transformed into "America".
- Named entity identification tools, such as LingPipe[1], are used to markup the name entities, especially potentially locations in the queries.

To cope with the variety of queries, we also developed a set of heuristic rules for fine tuning the two parts:

- All terms after the preposition are the geo part. For example, "South America prisons" is the geo part for query "Riots in South American prisons". Term "prisons" is not excluded from this part because it could bring geographic prison-related information.
- The terms before the preposition could form the topic-part, but some terms in the geo part can be included in the topic-part too. For example, the term "prisons" in the query "Riots in South American prisons" is added into the topic part "riots" to form "prisons riots" as the final topic part. There could be two reasons for having this rule. Firstly, adding more terms to the topic part could improve the precision of the search. "Prisons riots" indicates riots that happen in prisons rather than in other places. Secondly, expanding terms in the topic part would get more synonymous content-bearing information in the query expansion stage.
- When geographic terms cannot be identified by name entity recognizing tools, the term tagged with <ORGANIZATION> will be used as a geographic term. For example, "OECD" in query "Unemployment in the OECD countries" is treated as the geo part. Another example is "Cities near active volcanoes" in GeoCLEF 2006 queries, the phrase "active volcanoes" will be the geo part.
- Anaphora is manually dealt with. For example, query "Most visited sights in the capital of France and its vicinity" has its "its vicinity" translated into "Paris' vicinity". Here, before this translation, "the capital of France" is automatically converted into "Paris" using our location database.
- If there are no geographic terms in the geo part and all words are low-cases in this part, geographic terms occurring in the topic part will be moved to the geo part. For example, "Portuguese" in the topic part in query "Portuguese immigrant communities in the world" will be moved to the geo part to form "the Portugal world". Though we notice this could be a little bit unreasonable transform, we still have done it for

[1] http://alias-i.com/lingpipe/, Natural language processing software for text analytics, text data mining and search

generalization reason of our QPA module.

2.2.2 TPP Module

This module is responsible for extracting synonyms for the terms in the topic part from an online dictionary[2]. Another purpose of this module is to find synonyms from the Web so that we do not rely on a synonym thesaurus.

To obtain synonyms for English words, we borrowed the back translation idea from CLIR [1]. By obtain an English-Chinese dictionary and a Chinese-English dictionary, a English word such as "prison" can use its translation "监狱" as the bridge to bring back the synonymous English words like "jail, jailhouse, Job's pound, penitentiary, quod, pokey, stir, lockup, calaboose, gaol, big house, sheriff's hotel, Bridewell, iron house, jail house" and so on. Of course, it is clear from the example that some noises would be introduced via this method. In Section 2.2.4, we will talk about how to select terms from a synonym set for query expansion.

2.2.3 GPP Module

GPP module is the most important module in the GCEC system. It is responsible for geographic query expansion by utilizing geographic coordinate extraction and clustering.

The basic idea here is that, if we can identify that some geographic locations are in the same cluster with the original geographic location in the geo part, these locations should be treated as the geographic approximations of the location, and thus they can be used as the terms for expanding the geo part.

Here we think that clustering cannot performed based on the co-occurrence information because some geographic terms such as "United States, Germany" are often appear together with unrelated locations in geographic sense such as "former Yugoslavia" if we look at articles from certain period [2]. Therefore, we propose to use the locations' own geographic coordinates for clustering.

Given geographic coordinates, clustering is reasonable straightforward if the geo part is a particular location with definite geographic coordinates. Figure 2a shows an example of this case. The particular geographic coordinates of a geo part is determined accurately, shown as the black rectangle that an arrow points to. Those points within the circle of the broken line can be viewed as the geographic approximations of the geo part. Because the geographic coordinates of the geo part is the cluster's center, there is only one cluster to be determined here.

However, not all geo locations have clear coordinates associated with them. One example is "South America". Because there is no accurate coordinates for it, it is hard to determine the cluster center. Some other examples include state/province names. Although their capitals can be used as the surrogates, if the location of capital is on border of a large state/province, the geographic information of the state/province can hardly be represented by the capital.

To cope with this problem, we assume that those candidate geo locations will form natural clusters that reflect their own geographic relationships. For example, the geo part "South America prisons", different places related to different prisons may scatter among the whole South America countries. Because no definite geographic coordinates is used as the center of cluster, several clusters rather than one are formed. In each cluster, the element with the shortest distance (d_{min}) to its cluster center is considered as the actual cluster center. If the ratio of other elements with its distance d_i to d_{min} is less than a threshold, which is 10 in this experiment, we treat such elements as geographic query expansion for the geo part. Figure2b illustrates this case clearly.

---

[2] http://dict.cn/, an online Chinese-English dictionary.

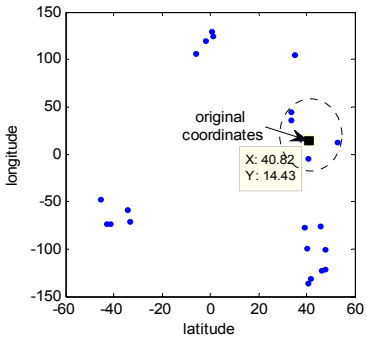

Figure 2a. cluster with a definite center

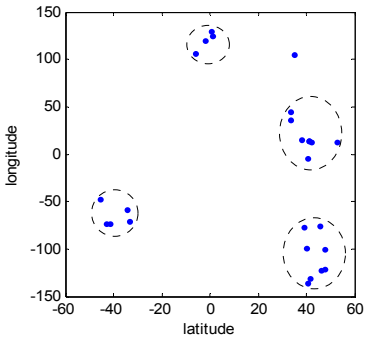

Figure 2b. clusters without a definite center

The process of geographic coordinate extraction and clustering is as follows:

- Google search engine is used to return top 10 retrieval results for the original geo part query. Hope that some related geographic location information is in these 10 Google results.
- By using a named entity identification tool like LingPipe, those marked as <LOCATION> are extracted as the candidates for expansion.
- Geographic coordinates are obtained online from sites like http://stable.toolserver.org/geohack/using external links of Wikipedia as entrance. A term without geographic coordinates is removed from the candidate list.
- After clustering these geo location candidates based on their geographic coordinates using K-Means algorithm, a set of geographic candidate terms will be selected for expanding the geo part.

2.2.4 CTS Module

The expansions in TPP for the topic part and GPP for the geo part can introduce noises. This module's task is to filter the noisy terms out from the candidate sets. The approach is to use the whole collection statistics.

According to term frequency-inverse document frequency model [6], we use the following formula to calculate the weight for a given term in the candidate sets:

$$w_{c,t} = \left(\frac{tf_{c,t}}{df_t} \right) \times \log \left(1 + \frac{cdf}{df_t} \right)$$

where $w_{c,t}$ is the weight of term t in collection, $tf_{c,t}$ is term occurrence frequency in collection, df_t is the number of documents containing term t, cdf is the number of all documents in collection. Note that here only uses collection information for fast filtering noisy terms without considering terms occurrence frequency in each documents. Our experiments show retrieval results can be improved if terms with weight, $1.6 < w_{c,t} < 3.5$, are selected as query expansion of both topic part and geo part.

3 Method 2: Wikipedia-based Geographic Information Retrieval

3.1 Using Wikipedia as Knowledge Base for Identifying Geographic Names – WIKIGEO System

As an open and online encyclopedia written collaboratively by volunteers around the world and organized the knowledge in encyclopedia style, Wikipedia is an invaluable online knowledge base about almost everything. In recent years, people started to extract knowledge from Wikipedia for various purposes, including word sense disambiguation [4], medical knowledge representation [5], and many other areas.

We are interested in mining geographic location names from the collection. Some types of knowledge can be clearly provided by encyclopedia structure, like name entry, category structure, and synonyms (i.e., redirect page) while some other knowledge is buried inside the article associated with the name entry, therefore the extraction needs deeper analysis and more powerful text analysis tools. Geographic location names belong to the second type of knowledge.

Because the articles of many geographic location related Wikipedia entries contain the complete set of latitude and longitude data, this motivated us to identify geographic names based on these data in the articles.

Our mining algorithm is as follow: for every Wikipedia page, we use pattern match method to find out all the pages satisfying our pre-define pattern that the page should contain keywords "coordinates" and contain digits nearby. This helps to extract potential geographic entries with their coordinates. Another important task in mining the geographic entries and their coordinates is to identify the connections between different mentions of the same location, and thus remove duplicates. Some locations actually refer to the same place but use different names or use different formats to representing the coordinates. We use the coordinates to calculate the distance between any two places. If the distance is less than 3 kilometers, we treat the two locations as the same place. The selection of 3 kilometers is an ad hoc decision. In total we extracted 370,787 geographic locations. They cover not only countries, states/provinces, counties/cities, but also continents, mountains, waterways and so on. The overall coverage is much larger than a world gazetteer we obtained which contains only 172,076 locations.

We use the geographic query parsing pilot study from last year (2007) to evaluate the mined results. The pilot study provides us geography related queries tagged with geographic names and their corresponding coordinates. Using those names and their coordinates as the ground truth, our mining results were evaluated as shown in Table 1.

Table 1: Geographic name identification evaluation. * means that this part includes mistakes in the ground truth

Total Query #	Correctly Identify geographic name and its coordination	Ambiguous geographic name	Fail to identify	Organization name	Special Geographic name
500	389	24	63 *	11	3

3.2 Query Expansion with Knowledge from Wikipedia

Once the extracted geographic names from the Wikipedia are available, we can use them to expand the locations in the queries. Our assumption is that if the location information in queries can be explicated explained, namely using more detailed location names to express where the query location covers, we can get more accurate search results.

We assume that the locations mentioned in the queries are the candidates for expansion. To cope with the long length of Wikipedia articles, the selection of the expansion terms were concentrated on geographic terms from the first paragraph of the Wikipedia article.

4 Experiments and Results

All our experiments were in the monolingual English environment. The English collection contains documents from Glasgow Herald that published in 1995 (GH95) and documents from Los Angeles Times

published in 1994 (LA94). Our retrieval system is Indri.

4.1 Runs and Results from GCEC system

We only submitted two runs to monolingual GeoCLEF for the time reason.

- PITTQP1: only the *title* elements of topic are used to generate query expansion of the run. All expansion is an automatic process according to methods mentioned in Section 2.2.
- PITTQP2: both the elements of *title* and *narrative* are used to generate query expansion of the run. Those geographic locations in narrative part are manually added to form expansion of *geo-part*, but the expansion of *topic part* is also automatic.

The *description* elements of topic are skipped as redundancy information in our experiments for its very similarity to the topic *title*. Some of its functions can be remedied by our synonym expansion.

Ours runs aim to:

- Without synonym dictionary and geographic knowledge base available, dynamically obtaining information of query expansion online will provide a substitutable and flexible method for improving effect of geographic information retrieval.
- To prove that geographic information from narrative part manually added as geographic query expansion will give better result as it uses more accurate geographic information.

Indri is selected as IR system on two preprocessed English collections (stop-words removing and Krovetz stemming). The experimental results are displayed in Table 2.

Table 2: Average Precision and R-Precision of our two runs

Experiments	Mean Average Precision	R-Precision
PITTQP1 (*Title Only*)	0.2624	0.2805
PITTQP2 (*Title+Narrative*)	0.2623	0.2706

The results show our first aim can be achieved when there is no synonym dictionary and geographic knowledge base for query expansion, but it's surprising that the *Title Only* based expansion achieved better results than combination of *Title and Narrative* both in MAP and RP.

4.2 Runs and Results from WIKIGEO system

For WIKIGEO system, we submitted two runs:

- PITTQi1: query expansion with keywords from title, description and narrative.
- PITTQi2: query expansion with keywords from title and narrative part.

Table 3: Average Precision and R-Precision of our two runs

Experiments	Mean Average Precision	R-Precision
PITTQi1 (*Title+Narratieve+Description*)	0.18570	0.19350
PITTQi2 (*Title+Narrative*)	0.17190	0.17990

As shown in Table 3, the WIKIGEO result is not good as GCEC method. The failure might come from the heuristic method of using the first paragraph to extract the geographic terms. This year's topics contain many high level geographic terms, like "South America". From Wikipedia, the obtained related geographic terms for "South America" are "America", "Pacific Ocean", or "Atlantic Ocean", which are not useful, and maybe hurting,

expansion terms in this case.

Therefore, it seems that it is not always desirable to expand the queries for any given topic. There might be some topics whose queries should not be expanded using our method. If we treat this as a classification problem, maybe we can build a classifier to automatically learn which types of topics should be expanded, and which are not. During the building of the classifier, we also need to figure out a set of features that are useful to represent the queries.

Table 4: Average Precision and R-Precision of baseline and experiment system

	Mean Average Precision	R-Precision
Baseline	0.2444	0.2618
WIKIGEO	0.2405	0.2579

We used the search topics from 2005 to 2007 (total 75) to build up the classification test collection. For comparison, we built a baseline system that is a plain geographic search engine using the title and description parts of the topic statements as the queries. The experiment WIKIGEO system used the same inputs.. stopword removal and stemming (snowball) were applied to the queries in both systems.

As shown in Table 4, WIKIGEO's results are similar to or slightly worse than that of the baseline. We divided the results into three groups: those topics that the experiment system performed worse than the baseline (negative effect), the two systems performed the same (zero effect), and the experiment system performed better (positive effect). As shown in Table 5 that WIKIGEO has no effect for most topics (about 35/75=47%), and within the rest, there are more negative effect results than positive ones. Therefore, our classifier is still sub-optimal in identifying the topics for expansion. We need further study on this topic.

Table 5: Effect of WIKIGEO to search results (according to MAP value)

	Negative Effect	Zero Effect	Positive Effect
WIKIGEO	25	35	15

5 Conclusions

In this paper, we present our participation in GeoCLEF 2008. As the first time participants, we only worked on the monolingual GeoCLEF evaluation and submitted four runs under two different methods. Our GCEC method aims to test the effectiveness of our online geographic coordinate extraction and clustering algorithm, and our WIKIGEO method wants to examine the usefulness of using the geo-coordinate information in Wikipedia for identifying geo-locations.

Our experiments results show that: 1) our online geographic coordinate extraction and clustering algorithm is useful for the type of locations that do not have clear corresponding coordinates; 2) the expansion based on the geo-locations generated by GCEC is effectiveness in improving Geographic retrievals. 3) Using Wikipedia we can find the coordinates for many geo-locations, but its usage for query expansion still need further studies. 4) query expansion based on title only obtained better results than using the combination of title and narrative parts, which are thought to contain more related geographic information. Further study is need for this part too.

Our future work can move in several directions, which include better method for locating related geo-location information, determining when it is appropriate to perform query expansion, and the parameters for effectiveness query expansion.

Acknowledgements

This work was partially supported by China Scholarship Council and the University of Pittsburgh.

References

[1] He, D., Oard, D. W., Wang, J., Luo, J., Demner-Fushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M. and Leuski, A. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. *ACM Transactions on Asian Language Information Processing*, 2(3): 219-244.2003

[2] Li, Z. S., Wang, C., Xie, X. and Ma, W. Y. MSRA Columbus at GeoCLEF 2006. In *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006.* 2006.

[3] Li, Z. S., Wang, C., Xie, X. and Ma, W. Y. Query Parsing Task for GeoCLEF2007 Report. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007.* 2007.

[4] Mihalcea, R. Using Wikipedia for Automatic Word Sense Disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics.* pages 196-203. 2007.

[5] Pedro, V., Niculescu, S. and Lita, L. Okinet: Automatic Extraction of a Medical Ontology From Wikipedia. In *WiKiAI08: a workshop of AAAI2008.* 2008.

[6] Salton, G., Wong, A. and Yang, C. S. A Vector Space Model for Automatic Indexing. *Communication of the ACM*, 18(11): 613-620.1975