MSRA Columbus at GeoCLEF2007

Zhisheng Li[1], Chong Wang[2], Xing Xie[2], Wei-Ying Ma[2]

[1]Department of Computer Science, University of Sci. & Tech. of China, Hefei, Anhui, 230026, P.R. China
zsli@mail.ustc.edu.cn

[2]Microsoft Research Asia, 4F, Sigma Center, No.49, Zhichun Road, Beijing, 100080, P.R. China
{chwang, xingx, wyma}@microsoft.com

Abstract

This paper describes the participation of Columbus Project of Microsoft Research Asia (MSRA) in GeoCLEF2007 (a cross-language geographical retrieval track which is part of Cross Language Evaluation Forum). This is the second time we participate in this event. Since the queries in GeoCLEF2007 are similar to those in GeoCLEF2006, we leverage most of the methods that we used in GeoCLEF2006, including MSRAWhitelist, MSRAExpansion, MSRALocation and MSRAText approaches. The difference is that MSRAManual approach is not included in GeoCLEF2007 this time, and we use MSRALDA instead. In MSRALDA, we combine the Latent Dirichlet Allocation (LDA) model with the text retrieval model. The results show that the application of LDA model in GeoCLEF monolingual English task needs to be further explored.

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; H.2.3 [**Database Management**]: Languages – Query Languages

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic information retrieval, System design, Latent Dirichlet Allocation, Evaluation

1. Introduction

In general web search and mining, location information is usually discarded. However, people need to deal with locations all the time, such as dining, traveling and shopping. GeoCLEF [7] aims at providing necessary platform for evaluating the geographic information retrieval. We also participated in GeoCLEF2006, and this is second time we participate this GeoCLEF event. The same to GeoCLEF2006, we only participated in the Monolingual GeoCLEF evaluation (EN-EN) and submitted five runs based on different methods.

2. Geographic Information Retrieval System

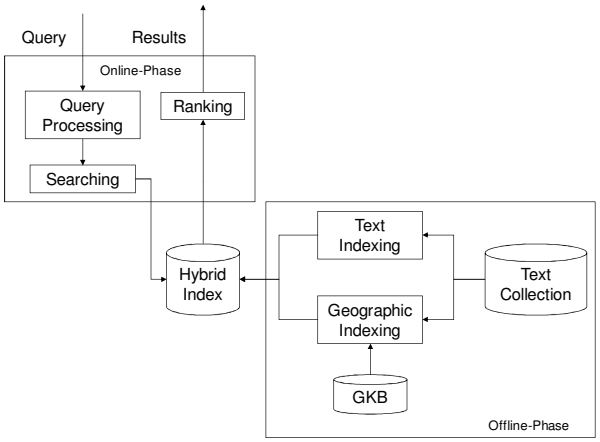

Figure 1. Architecture of our GIR system

Figure 1 is the system flow of our GIR system used in GeoCLEF2007. Our geographic information retrieval system is mainly composed of geo-knowledge base, location extraction module, geo-focus detection module, query-processing module, geo-indexing module and geo-ranking module. We will describe these modules in the following sections.

2.1 Geographic Knowledge Base

The Geographic Knowledge Base (GKB) we use is the same as that used in the last year. We use an internal geographic database as our basic gazetteer. This gazetteer contains basic information about locations all over the world, including location name, location type, location importance and hierarchical relationship between locations. We utilize this gazetteer to extract locations, to disambiguate locations and detect focuses of documents. Besides this gazetteer, we also use some other resources to improve the performance, including stop word list, person name list, white list and location indicator list. The method to generate the stop word list can be found in our report last year [12]. The white list and location indicator list is maintained manually, while the person name list is downloaded from the Internet. Finally we integrated all these resources as a GKB.

2.2 Location Extraction Module

Location Extraction module aims to extract locations and also disambiguate them from unstructured text. It is used in the query processing module and geo-indexing module. We manually composed the rules to address this task. It includes several parts: text parsing, geo-parsing, geo-disambiguating and geo-coding. For more details, please see our reports last year [12].

2.3 Geographic Focus Detection Module

When the locations' exact positions are determined, we want to get the focus of the documents. We adopted the algorithm described in [5]. Its main idea is to accumulate the score of each node in the hierarchical tree from bottom to up, and to sort all the nodes whose score are not equal to zero. Then we can get a list of focuses about the documents. The location with the biggest score is the most possible focus. For example, a document, mainly talking about Redmond economics, also has mentioned Seattle economics and the focus of the document is Redmond.

2.4 Query Processing Module

GeoCLEF2007 topics are structured topics, in which they contain topic numbers, topic-titles, topic-descriptions and topic-narratives. They don't provide explicit locations and relationships, so we need to parse the queries first and identify the geographic references, e.g. the textual terms, spatial relationships and the locations, from the different parts of the topics. But some topics are hard to be parsed. For example, "Lakes with monsters", "Rivers with floods", they don't contain explicit locations in the topics. For other examples, "Sport events in the French speaking part of Switzerland", "F1 circuits where Ayrton Senna competed in 1994", these human-language style queries are too difficult for machines to understand. Therefore, we designed three schemes to process the topics: automatic extraction, pseudo feedback and man-made whitelist.

1. Automatic extraction. We use the location extraction module to extract locations and get coordinates. To identify the relationships, e.g. "in", "near", we design a simple relationship matching program by adopting a rule-based approach. Except locations and relationships, we regard the left parts in the query-title as the text keyword. In such a way, we can handle topics containing explicit locations, e.g. "Damage from acid rain in northern Europe", "OSCE meetings in Eastern Europe".
2. Pseudo Feedback. For topics which don't contain explicit locations, we use pseudo feedback technique to expand the queries. We do this in the following steps. First, we search the topic title in our search engine to get the top-N documents (here we set N = 100), then we use the location extraction module to extract the locations from these documents and select the most frequent ones (the top 10 ones in our experiments). Finally, we use the selected ones as the locations for the queries.
3. Manual expansion. For the topics like "Whisky making in the Scottish Islands", "Water quality along coastlines of the Mediterranean Sea", though they contain location names in the titles, it is still difficult to identify the precise locations from these imprecise names, e.g. the coastlines of the Mediterranean Sea, the Scottish Islands. We expand them to exact locations manually by looking up in our geographic base as the location whitelist.

After processing the topics, we obtain the textual terms, spatial relationship and locations of the topics and send them to our GIR system.

2.5 Geo-Indexing Module

We use a hybrid indexing schema in our GIR system, which contains two parts: text index and geo-index. In our system, explicit locations and implicit locations [9] are indexed together and different geo-confidence scores are assigned to them. The advantage of this mechanism is that no query expansion is necessary and implicit location information can be computed offline for fast retrieval. In our system, we adopt two types of geo-indexes: one is called focus-index, which utilizes the inverted index to store all the explicit, and implicit locations of documents; the other is called grid-index, which divides the surface of the Earth into 1000 × 2000 grids. The documents will be indexed by these grids according to their focuses. For more details, please see our reports last year [12].

2.6 Geo-Ranking module

For the ranking module, we adopt IREngine, developed by MSRA, as our basic search engine. Then we integrated the geo-ranking module into it. To test the effectiveness of different methods, we totally designed three kinds of ranking algorithms: 1) pure textual ranking. Its basic ranking function is BM25; 2) linearly combining the text relevance and the geo-relevance; 3) linearly combing the text relevance and the LDA relevance.

2.6.1 Geo-based Model

In the second scheme, we retrieve a document list with geo-relevance from the geo-index by looking up the geographic terms. That is, for the focus-index, the matched docID list can be retrieved by looking up the locationID in the inverted index. For the grid-index, we can get the docID list by looking up the grids that the query location covers. We first retrieve two lists of documents relevant to the textual terms and the geographical terms respectively, and then merge them to get the final results. For re-ranking, we used a combined ranking function $R_{combined} = R_{text} \times \lambda + R_{geo} \times (1 - \lambda)$, where R_{text} is the textual relevance score and R_{geo} is the geo-relevance score. Experiments show that textual relevance scores should be weighted higher than geo-relevance scores ($\lambda = 0.8$ In our experiments).

2.6.2 LDA-based Model

For the third scheme, we explored the Latent Dirichlet Allocation model in our GeoCLEF2007 experiments. Latent Dirichlet Allocation (LDA) model [10] is a semantically consistent topic model. In LDA, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents. The graphical model of LDA is shown in Figure 2.

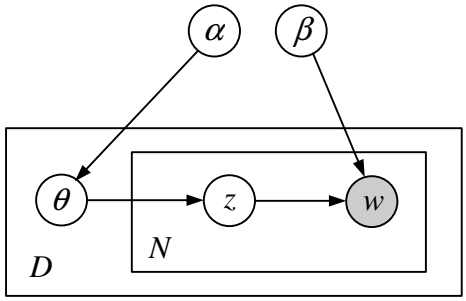

Figure 2. Graphical representation of LDA model.

The generative process for LDA can be stated as follows:

For each text document d:

1. Choose $\theta \sim Dirichlet(\alpha)$.
2. For each word w_n in document d:
    a. Choose a topic $z_n \sim Multinomial(\theta)$.
    b. Choose a word $w_n \sim p(w_n|z_n, \beta)$, which is a topic-specific multinomial probability distribution.

Thus, the likelihood of generating a corpus $D = \{d_1, d_2, \ldots, d_M\}$ is:

$$P(D|\alpha, \beta) = \prod_{m=1}^{M} \int p(\theta_m|\alpha) \left(\prod_{n=1}^{N_m} \sum_{z_{mn}} p(z_{mn}|\theta_m) p(w_{mn}|z_{mn}, \beta) \right) d\theta_m$$

The LDA model is represented as a probabilistic graphical model. Compared to the probabilistic Latent Semantic Indexing model (pLSI) [11], LDA processes fully consistent generative semantics by treating the topic mixture dis-

tribution as a K-parameter hidden variable rather than a large set of individual parameters which are explicitly linked to the training set. Thus LDA overcomes the overfitting problem and has the fully generative process for new documents.

In [13], Xing et al. discussed the application of LDA in ad hoc retrieval. We use the similar approach for our geographic information retrieval task in GeoCLEF2007, which allows us to compute a probability a query given a document using LDA model. That is each document is scored by the likelihood of its model generating a query Q,

$$P(Q|d) = \prod_{q \in Q} P(q|d)$$

where d is a document model, Q is the query and q is a query term in Q. $P(Q|d)$ is the likelihood of the document model generating the query terms under the "bag-of-words" assumption that terms are independent given the documents. In our experiment, we use LDA model as the document model.

After we computed the $P_{lda}(Q|d)$, we selected the top 1000 documents with the highest $P_{lda}(Q|d)$ for each query. We also use our text search engine to retrieve top 1000 documents respectively. Then we merged these two document-lists. If one document in both of the list, we used a combined score function $R_{combined} = R_{text} \times \lambda + P_{lda} \times (1 - \lambda)$, where R_{text} is the textual relevance score and P_{lda} is the LDA model probability (here we set λ= 0.5). Both scores are normalized. Otherwise, we computed a new score for the document by multiplying a decay factor 0.5. Finally we re-ranked all these documents by the new scores and selected the top 1000 ones as result.

3. Monolingual GeoCLEF Experiments (English - English)

In Table 1, we show all the five runs submitted to GeoCLEF. When the topic field is "Title", we just use the title element of the topic to generate the query of the run. When the topic field is "Title + Description", this means that the title and desc are both used in the run. When the topic field is "Title + Description + Narrative", this means that title, desc and narr are all used. And the "Description" field in Table 1 gives a simple explanation of the methods used in the runs. Priorities are assigned by us, where priority 1 is the highest and 5 the lowest.

Table 1. Run information

Run-ID	Topic Fields	Description	Priority
MSRALDA	Title	Combining LDA Model	1
MSRAWhiteList	Title + Description	using geo knowledge base and manual query construction	2
MSRAExpansion	Title + Description	using query expansion	3
MSRALocation	Title	without geo knowledge base and query expansion	4
MSRAText	Title + Description + Narrative	using pure text	5

In MSRALDA, we used the title elements to generate the queries. Then we used the LDA-based model described in section 2.6.2 to select 1000 documents for each query. In MSRAWhiteList, we used the Title and Desc elements of the topics to generate the queries. For some special queries, e.g. "Scottish Islands", "coastlines of the Mediterranean Sea", we cannot get the exact locations directly from our gazetteer, so we utilized the GKB to get the corresponding geo-entities. Then we can make a whitelist manually for the geo-terms of these queries. In MSRAExpansion, we generated the queries with title and desc elements of the topics. Different from MSRAWhiteList, the queries were automatically expanded based on the pseudo-feedback technique. First we used the original queries to search the corpus. Then we extracted the locations from the returned documents and calculated the times each location appears in the documents. Finally we got the top 10 most frequent location names and combined them with the original geo-terms in the queries. In MSRALocation, we used the title elements of the topics to generate the queries. And we do not use geo knowledge base or query expansion method to expand the query locations. We just utilize our location extraction module to extract the locations automatically from the queries. In MSRAText, we generated the queries with title, desc and narr elements of the topics. We just utilized our pure text search engine "IREngine" to process the queries.

4. Results and Discussion

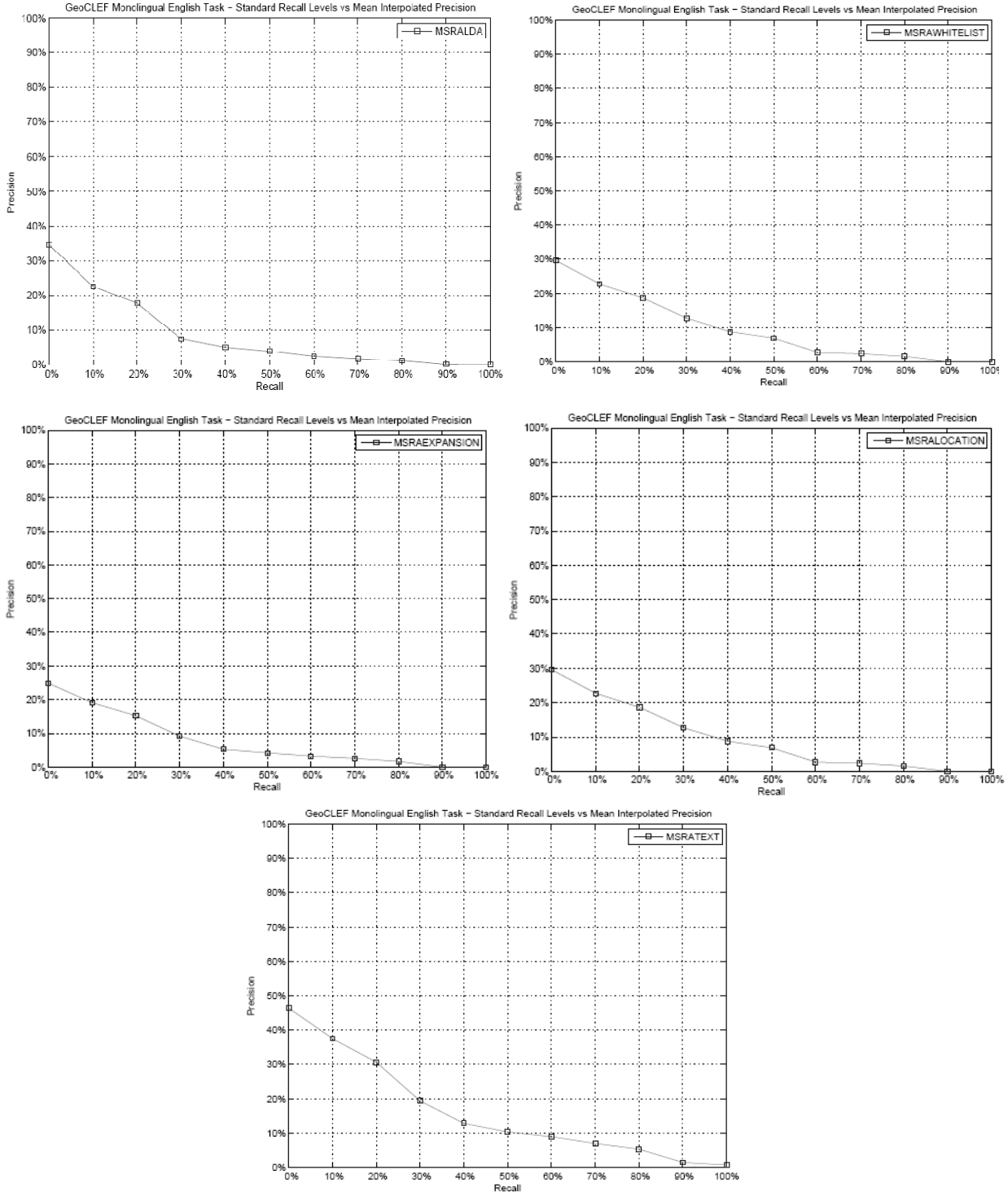

Figure 3. Standard recall levels vs mean interpolated precision for all five runs

Figure 3 and Table 2 show the results of MSRA Columbus on the GeoCLEF monolingual English task. MSRAText run achieves the best precision in our results. The precision of MSRALDA run decreases after combing the LDA model with the pure text model. As the same as GeoCLEF2006, the performance of MSRAExpansion is the lowest

among the five runs, because many unrelated locations are added to new topics after pseudo feedback for some topics.

From Table 2, we can see that MSRALDA drops the performance significantly compared with MSRAText by about 7.6% in MAP. This indicates that linearly combining LDA model with text model does not work well. The reason may be that we haven't tune the parameter to be the best or linear combination is not a good choice.

Though the MAP of MSRALDA is lower than MSRAText, it still outperforms the latter one in some cases. For example, for the 10.2452/53-GC "Scientific research at east coast Scottish Universities", MSRAText just retrieves 39 relevant documents, while MSRALDA retrieves 43 relevant ones (The number of relevant documents is 64). For 10.2452/65-GC "Free elections in Africa", MSRAText retrieves 59 relevant documents and MSRALDA retrieves 74 (The number of relevant documents is 93). And we can see that the standard deviation of MSRALDA is just 0.09, lower than MSRAText. This indicates that MSRAText performs badly in some cases while MSRALDA performs more stably.

MSRAWhiteList and MSRALocation achieve similar MAP with each other, about 8.6%. Their MAPs are much lower than MSRAText by about 6.5% and just a little better than MSRAExpansion. Different from the results of GeoCLEF2006, automatic location extraction and manual expansion don't bring improvements.

Table 2. MAP & Standard Deviation for five runs

RUN-ID	MAP	Standard Deviation
MSRALDA	7.51%	0.090
MSRAWhiteList	8.61%	0.145
MSRAExpansion	7.01%	0.134
MSRALocation	8.61%	0.145
MSRAText	15.19%	0.197

5. Conclusions

We conclude that the application of LDA model in GeoCLEF monolingual English task needs to be further explored. Another conclusion is that automatic location extraction from the topics does not improve the retrieval performance, even decrease it sometimes. The third conclusion is the same as last year. That is automatic query expansion by pseudo feedback weakens the performance because the topics are too hard to be handled and many unrelated locations are added to new topics. Obviously, we still need to improve the system in many aspects, such as query processing, geo-indexing and geo-ranking.

6. Reference:

[1] E. Amitay, N. Har'El, R. Sivan and A. Soffer. Web-a-where: Geotagging Web Content. SIGIR 2004.

[2] Y.Y. Chen, T. Suel and A. Markowitz. Efficient Query Processing in Geographical Web Search Engines. SIGMOD'06, Chicago, IL, USA.

[3] B. Martins, M. J. Silva and L. Andrade. Indexing and Ranking in Geo-IR Systems. GIR'05, Bremen, Germany.

[4] A.T. Chen. Cross-Language Retrieval Experiments at CLEF 2002. Lecture Notes in Computer Science 2785, Springer 2003.

[5] C. Wang, X. Xie, L. Wang, Y.S. Lu and W.Y. Ma. Detecting Geographical Locations from Web Resources. GIR'05, Bremen, Germany.

[6] M. Sanderson and J. Kohler. Analyzing Geographical Queries. GIR'04, Sheffield, UK.

[7] GeoCLEF2007. http://ir.shef.ac.uk/geoclef/

[8] C.B. Jones, A.I. Abdelmoty, D. Finch, G. Fu and S. Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Lecture Notes in Computer Science 3234, 2004.

[9] Z. S. Li, C. Wang, X. Xie, X. F. Wang and W.Y. Ma. Indexing implicit locations for geographic information retrieval. GIR'06, Seattle, USA.

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research. 3:993-1022. Jan. 2003.

[11] Hofmann, T. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[12] Z.S. Li, C. Wang, X. Xie and W.Y. Ma, MSRA Columbus at GeoCLEF 2006, working note, GeoCLEF 2006, Alicante, Spain, Sep. 2006

[13] Wei, X. and Croft, W.B. LDA-based Document Models for Ad-hoc Retrieval. In the Proceedings of SIGIR '06, 178-185, 2006.