Integrate Multilingual Web Search Results using Cross-Lingual Topic Models

Duo Ding
Shanghai Jiao Tong University, Shanghai, 200240, P.R. China
dingduo1@gmail.com

Abstract

With the thriving of the Internet, web users today have access to resources around the world in more than 200 different languages. How to effectively manage multilingual web search results has emerged as an essential problem. In this paper, we introduce the ongoing work of leveraging a Cross-Lingual Topic Model (CLTM) to integrate the multilingual search results. The CLTM detects the underlying topics of different language results and uses the topic distribution of each result to cluster them into topic-based classes. In CLTM, we unify distributions in topic level by direct translation, thus distinguishing from other multilingual topic models, which mainly concern the parallelism at document or sentence level (Mimno 2009; Ni, 2009). Experimental results suggest that our CLTM clustering method is effective and outperforms the 6 compared clustering approaches.

1 Introduction

The growing of the Internet has made the web multilingual. With the Internet, user can browse the web page written in any language, and search for results in any language in the world.

However, since users would have a large set of search results edited in many languages after multilingual search (shown as Figure 1), the redundancy issue became a problem. Here the "redundancy issue" stands for two problems. The first is that we would get duplicated results from different language search. This can be fixed by simply maintaining a set and throw away the duplicated results. The second problem is that the users will get so many search results after multilingual search that they cannot quickly find the results they want. To facilitate users' quick browsing, one effective solution might be post-retrieval document clustering, which had been shown by Hearst and Pedersen (1996) to produce superior results. So we can employ the Cross-Lingual Topic Models to cluster the numerous results into topic classes, each containing the results related to one specific topic, to solve the redundancy problem.

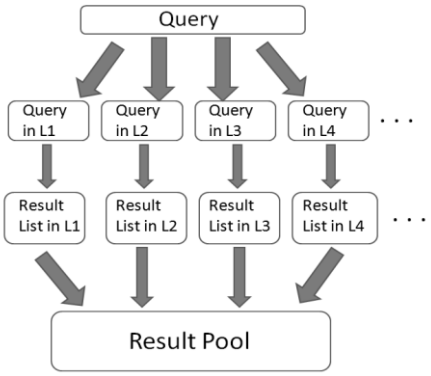

Figure 1: Multilingual Search

Our approach works in two steps. First we translate the topic documents into a unified language. Then, by conducting a clustering method derived from the Cross-Lingual Topic Model (CLTM), we cluster all the results into topic classes. We assume different "topics" exist among all the returned search results. (Blei 2003). Thus by detecting the underlying topics of search results, we give a topic distribution for each result and then cluster it into a particular class according to the distribution. Through experiments, the CLTM gives an impressive performance in clustering multilingual web search results.

2 Cross-Lingual Topic Models

Topic models have emerged as a very useful tool to detect underlying topics of text collections. They are probabilistic models for uncovering the underlying semantic structure of a document collection

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 20–24,
Chiang Mai, Thailand, November 8-12, 2011.

based on a hierarchical Bayesian analysis of the original texts (Blei et al. 2003). Having the method of assigning topic distributions to the terms and documents, this analysis of the context can be utilized on many applications. Meanwhile, the development of multilingual search is calling for useful cross-lingual tools to integrate the results in different languages. So we leverage Cross-Lingual Topic Models (CLTM) to accomplish the task of integrating multilingual web results.

Some similar methods have been proposed recently to define polylingual or multilingual topic models to find the topics aligned across multiple languages (Mimno 2009; Ni, 2009). The key difference between us is that the polylingual topic models assume that the documents in a tuple share the individual tuple-specific distribution over topics, while in the Cross-Lingual Topic Model, the distributions of tuples and different languages are identical. At the same time, our emphasis is to utilize the power of CLTM to solve the problem of clustering multilingual search results, which is different from other topic model tools.

2.1 Definition

Firstly we give the statistical assumptions and terminology in Cross-Lingual Topic Models (CLTM). The thought behind CLTM is that, for results within a specific language search result set, we model each result as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms in this language. In every language Li, Let K be a specified number of topics, V the size of the vocabulary, $\vec{\alpha}$ a positive K-vector, and η a scalar. We let DirV ($\vec{\alpha}$) denote a V-dimensional Dirichlet with vector parameter $\vec{\alpha}$ and DirK (η) denote a K dimensional symmetric Dirichlet with scalar parameter η.

There might be several topics underlying in the collection. We draw a distribution for each topic over words $\vec{\beta}_k \sim \mathrm{Dir}_V(\eta)$. And for each search result document, we draw a vector of topic proportions $\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$. Finally for each word, we firstly give a topic assignment $Z_{d,n} \sim \mathrm{Mult}(\vec{\theta}_d)$, where the range of $Z_{d,n}$ is 1 to K; then draw a word $W_{d,n} \sim \mathrm{Mult}(\vec{\beta}_{z_{d,n}})$, where the range of $W_{d,n}$ is from 1 to V.

From definition above we can see that the hidden topical structure of a collection is represented in the hidden random variables: the topics $\vec{\beta}_{1:K}$, the per-document topic proportions $\vec{\theta}_{1:D}$, and the per-word topic assignments $z_{1:D,1:N}$. This is similar to another kind of topic models, latent Dirichlet allocation (LDA).

We make central use of the Dirichlet distribution in CLTM, the exponential family distribution over the simplex of positive vectors that sum to one. Since we use distribution similar to latent Dirichlet allocation on each language result set, we give the Dirichlet density:

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \qquad (1)$$

The parameter $\vec{\alpha}$ is a positive K-vector, and Γ denotes the Gamma function, which can be thought of as a real-valued extension of the factorial function. Under the assumption that document collections (result sets) in different languages share a same topic distribution, we can describe the Cross-Lingual Topic Models in Figure 2.

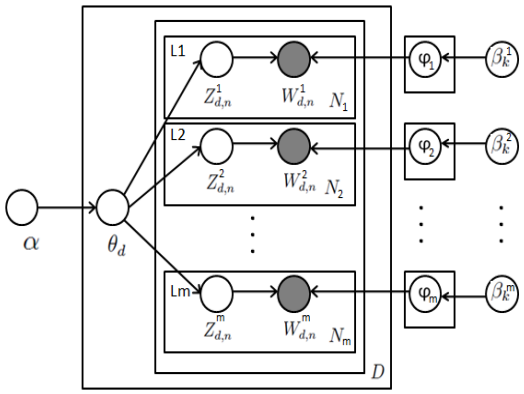

Figure 2: The graphical model presentation of the Cross-Lingual Topic Model (CLTM)

2.2 Clustering with CLTM

From the definition, we see that CLTM contains two Dirichlet random variables: the topic proportions $\vec{\theta}$ are distributions over topic indices {1, . . . , K}; the topics $\widehat{\beta}$ are distributions over the vocabulary. We use these variables to formulate our topic-detecting method.

Detecting Topics

In CLTM, exploring a corpus through a topic model typically begins with visualizing the posterior topics through their per-topic term probabilities $\widehat{\beta}$. In our method, we need to find several topics in the "Result Pool" of each query, thus making it possible to assign topic distributions to each result in the

set. To do so, we detect the topics in a result set by visualizing several posterior topics and use the following formula to calculate the word score:

$$\text{word-score}_{k,v} = \widehat{\beta}_{k,v} \log \left(\frac{\widehat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \widehat{\beta}_{j,v} \right)^{\frac{1}{K}}} \right) \quad (2)$$

We can see that the above formula is based on the TFIDF term score of vocabulary terms used in information retrieval (Baeza-Yates and Rbiero-Neto, 1999). We use this score to determine salient topics in a query's result set. The first part of it is similar to the term frequency (TF); the second part is similar to the document frequency (IDF).

Document Topic Distribution

When several topics are found in a result set, we would like to know the underlying topics contained in each result document so that we can cluster them into a particular class according to their topics. Since a result document may contain multiple topics and what we need is the most salient one, we can plot the posterior topic proportions and examine the most likely topic assigned to each word in this query to find the most salient topic. In our method, we sum up the distribution of every term in the document to form the final distribution of this doc.

$$\text{doc-score}_{k,v} = \sum_{i=1}^{N_v} \widehat{\beta}_{i,v} \quad (3)$$

This formula calculates the similarity of a document on the *K*th topic. N*v* denotes quantity of words that the *v*th result contains.

After the two-step processing, for each result document in a query's result list, we have *K* similarities which respectively denote the possibility for the document to be clustered to the *K*th topic class. We then conduct clustering on the result set based on this possibility to put them in different topic-based classes.

3 Experiments

In this section, we give experimental results on Cross-Lingual Topic Model clustering method, compared with 6 other clustering algorithms, to show that CLTM is a powerful tool in cross-lingual context analysis and multilingual topic-based clustering.

For this series of experiments we simply use the cluster results of two languages, English and Chinese to show the performance of different clustering methods (Because it is convenient to evaluate). However, due to the fact that the Cross-Lingual Topic Models are language independent, we believe that the method is also feasible in other languages.

3.1 Baseline Clustering Algorithms

In the first place, we apply 6 baseline clustering algorithms to the unified search results. We extract 20 frequently referred Chinese search queries and translate them into English. (Using Google Translate.) Then for each pair of queries we search them both in Chinese and English in the Google Search Engine, each recording top 40 returned results (including title, snippet and url). And then we regard English as the unified language and translate the 40 Chinese results into English, again using Google Translate, thus having totally 80 returned search results for each query.

In the next step, for each of the 80 results, we convert these 80 snippets into the vector-space format files. After that, we begin to cluster these result documents (snippets) into classes. In our definition, the cluster number is 5. The fixed-predefined clustering number is more effective for both baseline methods and CLTM method to conduct clustering and also drives it clearer to make comparisons.

The 6 baseline clustering algorithms we use are: repeated bisection (rb), refined repeated bisection (rbr), direct clustering (direct), agglomerative clustering (agglo), graph partitioning (graph), biased agglomerative (bagglo). We use a clustering tool, CLUTO, to implement baseline clustering.

The similarity function is chosen to be cosine function, and the clustering criterion function for the rb, rbr, and direct methods is

$$\text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)} \quad (4)$$

In this formula, *K* is the total number of clusters, *S* is the total objects to be clustered, *Si* is the set of objects assigned to the *i*th cluster, *ni* is the number of objects in the *i*th cluster, *v* and *u* represent two objects, and *sim(v, u)* is the similarity between two objects.

Clustering Algorithm	Parameter	Algorithm Description
Repeated Bisection	-rb	The desired k-way clustering solution is computed by performing a sequence of k-1 repeated bisections.
Refined Repeated Bisection	-rbr	Similar to the above method, but at the end, the overall solution is globally optimized.
Direct Clustering	-direct	In this method, the desired k-way clustering solution is computed by simultaneously finding all k clusters.
Agglomerative Clustering	-agglo	The k-way clustering solution is computed using the agglomerative paradigm whose goal is to locally optimize (min or max) a particular clustering criterion function.
Graph Partitioning	-grapg	The clustering solution is computed by first modeling the objects using a nearest-neighbor graph, and then splitting the graph into k-clusters using a min-cut graph partitioning algorithm
Biased Agglomerative	-bagglo	Similar to the agglo method, but the agglomeration process is biased by a partitional clustering solution that is initially computed on the dataset.

Table 1: Parameter and description of the 6 baseline clustering algorithms used in the experiment

For agglomerative and biased agglomerative clustering algorithm, we use the traditional UPGMA criterion function and for graph partitioning algorithm, we use cluster-weighted single-link criterion function. The parameters and explanations for each clustering algorithm are represented in Table 1.

3.2 Cross-Lingual Topic Model Clustering

In Cross-Lingual Topic Model based clustering, we firstly calculate the word score for each vocabulary by using formula (2) in Section 2. Thus for each query, there is a probability for each of its vocabulary word on 5 different topics. Then, we use formula (3) to calculate the probability of each document (each snippet) on 5 topics. Finally, we find the topic with highest probability in each document and assign the document into this topic class, which finishes the process of clustering.

In our evaluation process, we ask 7 evaluators to view the results of different clustering methods. Each of the evaluators is given the clustering results on 2 or 3 queries in 7 different methods (6 baseline methods plus CLTM). And they are asked to compare the results by giving two scores to each method. In the evaluation process, they are blind to the clustering method names of the assigned results. The first score is the "Internal Similarity", which accounts for the similarity of the results clustered into the same class. This score reveals the compactness of each topic class and the range of the score is from 1 to 10: 1 score means not good compactness and 10 scores means perfect compactness. The second score is called "External Distinctness", which shows whether the classes are distinct with each other. The range is also 1 to 10: 1 score represents poor quality and 10 represents the best performance. The results of evaluations are shown in Figure 3 and Figure 4.

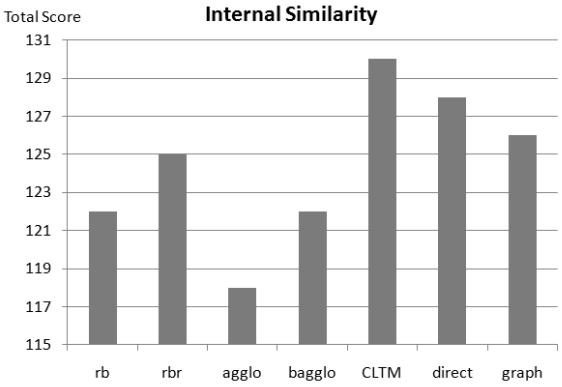

Figure 3: The Internal Similarity of 7 methods

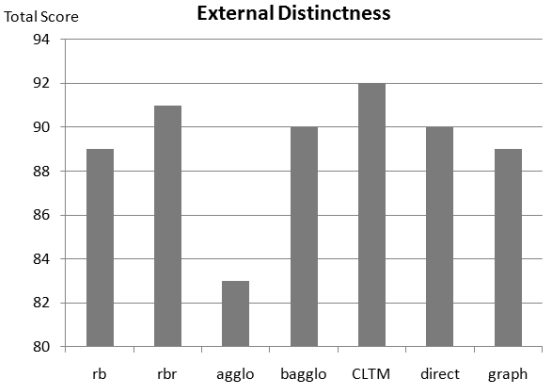

Figure 4: The External Distinctness of 7 methods

4 Conclusion

In this paper, we introduce the ongoing work of exploiting a kind of topic models, Cross-Lingual

Topic Models (CLTM), to solve the problem of integrating and clustering multilingual search results. The CLTM detects the underlying topics of the results and assign a distribution to each result. According to this distribution, we cluster each result to the topic class of which it is mainly about. We give each word a "word-score" which represents the distribution of topics on this word and sum all the term probabilities up in a result to obtain the topic distribution for each result document. To evaluate the effectiveness of Cross-Lingual Topic Models, we compare it with 6 baseline clustering algorithms on the same dataset. The experimental results of "Internal Similarity" and "External Distinctness" scores suggest that the Cross-Lingual Topic Model gives a better performance and provides more reasonable results for clustering multilingual web search documents.

Acknowledgments

The author would like to thank Matthew Scott of Microsoft Research Asia for helpful suggestions and comments. The author also thanks the anonymous reviewers for their insightful feedback.

References

Andreas Faatz: Enrichment Evaluation, technical report TR-AF-01-02 at Darmstadt University of Technology

A. V. Leouski and W. B. Croft. 1996. An evaluation of techniques for clustering search results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst.

Bernard J. Jansen, Amanda Spink,*, Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management 36 (2000).

Chi Lang Ngo and Hung Son Nguyen. 2004. A Tolerance Rough Set Approach to Clustering Web Search Results, PKDD 2004, LNAI 3202, pp. 515–517.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation, 3:993-1022.

D. R. Cutting, D. R. Karger, J. O. Pedersen and J. W. Tukey, Scatter/Gather. 1992. A cluster-based approach to browsing large document collections, In *Proceedings of the 15th International ACM SIGIR Conference (SIGIR'92),* pp 318-329.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith and Andrew McCallum. 2009. Polylingual Topic Models, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore.

He Xiaoning, Wang Peidong, Qi Haoliang, Yang Muyun, Lei Guohua, Xue Yong. 2008. Using Google Translation in Cross-Lingual Information Retrieval, *Proceedings of NTCIR-7 Workshop Meeting*, December 16–19, Tokyo, Japan

Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma and Jinwen Ma. 2004. Learning to Cluster Web Search Results. SIGIR04, Sheffield, South Yorkshire, UK.

Liddle, S., Embley, D., Scott, D., Yau, S. 2002. Extracting Data Behind Web. In *Proceedings of the Joint Workshop on Conceptual Modeling Approaches for E-business: A Web Service Perspective (eCOMO 2002)*, pp. 38–49 (October 2002)

Murata, M, Ma, Q, and Isahara, H. 2002. "Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval". *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan. NII, Tokyo.

McRoy, S. 1992. Using Multiple Knowledge Sources for Word Sense Discrimination, in Computational Linguistics, vol. 18, no. 1.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, volume 19, Issue 2.

P.S. Bradley, Usama Fayyad, and Cory Reina. 1998. Scaling Clustering Algorithms to Large Databases, From: KDD-98 Proceedings, AAAI (www.aaai.org).

Raghavan, S., Garcia-Molina, H. 2001. Crawling the Hidden Web. In: *Proceedings of the 27th International Conference on Very Large Data Bases*, pp.29–138.

W. B. Croft. 1978. Organizing and searching large files of documents, Ph.D. Thesis, University of Cambridge.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia, WWW 2009, Madrid, Spain.

Zamir O., Etzioni O. 1998. Web Document Clustering: A Feasibility Demonstration, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'98)*, 46-54.

Zamir O., Etzioni O. Grouper. 1999. A Dynamic Clustering Interface to Web Search Results. In *Proceedings of the Eighth International World Wide Web Conference (WWW8)*, Toronto, Canada.