Cross-Lingual Topic Discovery From Multilingual Search Engine Query Log

DI JIANG, SKLSDE Lab, Beihang University, China and Baidu, China
YONGXIN TONG, SKLSDE Lab, Beihang University, China
YUANFENG SONG, Baidu, China

Today, major commercial search engines are operating in a multinational fashion to provide web search services for millions of users who compose search queries by different languages. Hence, the search engine query log, which serves as the backbone of many search engine applications, records millions of users' search history in a wide spectrum of human languages and demonstrates a strong multilingual phenomenon. However, with its salience, the multilingual nature of a search engine query log is usually ignored by existing works, which usually consider query log entries of different languages as being orthogonal and independent. This kind of oversimplified assumption heavily distorts the underlying structure of web search data. In this article, we pioneer in recognition of the multilingual nature of a query log and make the first attempt to cross the language barrier in query logs. We propose a novel model named *Cross-Lingual Query Log Topic Model* (CL-QLTM) to analyze query logs from a cross-lingual perspective and derive the latent topics of web search data. The CL-QLTM comprehensively integrates web search data in different languages by collectively utilizing cross-lingual dictionaries, as well as the co-occurrence relations in the query log. In order to relieve the efficiency bottleneck of applying the CL-QLTM on voluminous query logs, we propose an efficient parameter inference algorithm based on the MapReduce computing paradigm. Both qualitative and quantitative experimental results show that the CL-QLTM is able to effectively derive cross-lingual topics from multilingual query logs and spawn a wide spectrum of new search engine applications.

CCS Concepts: • **Information systems** → **Information retrieval**; *Retrieval models and ranking*; *Novelty in information retrieval*

Additional Key Words and Phrases: Search engine, query log, probabilistic topic model

ACM Reference Format:
Di Jiang, Yongxin Tong, and Yuanfeng Song. 2016. Cross-lingual topic discovery from multilingual search engine query log. ACM Trans. Inf. Syst. 35, 2, Article 9 (September 2016), 28 pages.
DOI: http://dx.doi.org/10.1145/2956235

This work is supported in part by the National Grand Fundamental Research 973 Program of China under Grant No. 2014CB340304, National Science Foundation of China (NSFC) under Grant No. 61502021, 61328202, and 61532004, State Key Laboratory of Software Development Environment under Grant No. SKLSDE-2016ZX-13, the Hong Kong RGC Project N_HKUST637/13, NSFC Guang Dong Grant No. U1301253, and Microsoft Research Asia Fellowship 2012.
The work was done during Di Jiang's visit to the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, China.
Authors' addresses: D. Jiang, the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, China and Baidu Campus, No. 10 Shangdi 10th Street, Haidian District, Beijing, China; email: jiangdi@baidu.com; Y. Tong (corresponding author), the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering and International Research Institute for Multidisciplinary Science, Beihang University, China; email: yxtong@buaa.edu.cn; and Y. Song, Baidu Campus, No. 10 Shangdi 10th Street, Haidian District, Beijing, China; email: songyuanfeng@baidu.com.

DOI: http://dx.doi.org/10.1145/2956235

1. INTRODUCTION

Major commercial search engines such as Google, Bing, and Yahoo are currently operating in multiple regions and nations. Due to the lingual diversity of the search engine users, the search engine query log usually records millions of users' web search histories in a variety of human languages [Grimes et al. 2007]. For example, in a public query log[1] released by Yahoo, there exists web search data in nine different languages, covering Chinese, English, French, German, Italian, Japanese, Korean, Portuguese, and Spanish. However, in existing research work, the query log entries of different languages are usually considered as being orthogonal and independent. This kind of oversimplified assumption is not aligned with the reality of web search.

Consider the example shown in Table I. q_1, q_2 and q_3 are English search queries, while q_4 and q_5 are Chinese ones. It is easy to see that query log entries in different languages are not independent due to two reasons: (1) Some query words in English and Chinese queries have exactly the same meaning and they can be translated into each other by utilizing a cross-lingual dictionary. For example, the English word "microsoft" is translated as "微软" in Chinese and the word 'Disney" is translated to the Chinese word "迪士尼". Hence, some search queries in different languages can be related to each other by the words with the same meanings. (2) The clicked URLs are essentially cross-lingual. Search queries in different languages can result in the same URL clickthrough. For instance, we observe that there exists a Chinese query and an English query, both of which are about the Disneyland Resort and they result in the clickthrough on the official websites of the Disneyland Resort in Orlando.

The above example illustrates that the query log entries in different languages are coupled rather than being independent. In this article, we recognize the multilingual nature of query logs and study the problem of capturing the latent structure of query logs by discovering cross-lingual topics. Finding cross-lingual topics is effective to improve the performance of search engines since a significant portion of search queries and web pages are multilingual. In later sections, we will show that the cross-lingual topics are able to improve the performance of real-life applications such as query/URL recommendation and information retrieval. Modeling their semantics in a cross-lingual approach provides better understanding of the latent semantics. Although some probabilistic topic models, such as those presented by Carman et al. [2010] and Jiang et al. [2012], have been proposed for query log analysis, they are designed to work only for monolingual query logs and would not work for deriving topics from query logs in multilingual scenarios. The deficiency comes from the following fact: all these topic models rely on the co-occurrences of query words to compose a topic, but the query words in different languages generally do not co-occur with each other in each web search task. There do exist some cross-lingual topic models such as the Multilingual Topic Model (ML-LDA) [Ni et al. 2009] and Polylingual Topic Model (PLTM) [Mimno et al. 2009]. These cross-lingual topic models are only applicable to document pairs from parallel/comparable corpus, which is clearly not the case of a query log. Moreover, a query log is quite different from the documents or articles studied in the conventional setting of topic modeling. A query log contains words, URLs, sessions, and timestamps which cannot be modeled by existing cross-lingual topic models. Thus, with the existing probabilistic topic models, we can not effectively capture the multi-lingual nature of query logs. Deriving multilingual topics from query logs is nontrivial and the challenges are essentially threefold:

—Since web search is essentially dynamic, how to cross the language barrier in a dynamic fashion is still an open problem.

[1]http://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

Table I. Search Engine Query Log Example

ID	Query	Clicked URL
q_1	disney orlando	disneyworld.disney.go.com
q_2	thallium poisoning	www.youtube.com/ watch?v=qOCf0dU0AAg
q_3	microsoft	www.microsoft.com
q_4	迪士尼	disneyworld.disney.go.com
q_5	微软	www.microsoft.com

—Since a multilingual search engine query log contains many types of information, such as the textual search queries, URLs, and timestamps, how to effectively integrate all this information in a principled way is still a challenging issue and has not been solved in literature.
—Since a multilingual search engine query log is voluminous and applying probabilistic topic models on such big data is time-consuming, how to relieve the efficiency bottleneck of applying the technique of topic modeling to multilingual query log has rarely been explored before.

In order to handle the aforementioned challenges, we propose a novel probabilistic topic model, called *Cross-Lingual Query Log Topic Model* (CL-QLTM), which captures the underlying structure of multilingual query logs by discovering cross-lingual topics. The CL-QLTM takes a multilingual query log and a cross-lingual dictionary as the input and outputs a set of latent topics which contain words in different languages. Based on a commercial search engine query log, we gauge the performance of the CL-QLTM by a wide range of standard metrics and search engine applications. The CL-QLTM demonstrates superior performance against strong baselines and shows significantly improved performance in different applications.

The contributions of this article are summarized as follows:

—We recognize the multilingual nature of search engine query logs and systematically study the issue of discovering cross-lingual topics from a multilingual search engine query log.
—We develop the CL-QLTM, which effectively crosses the language barrier in a multilingual query log.
—In order to effectively train the CL-QLTM on a massive query log, we further develop an efficient parameter inference method based on the MapReduce paradigm, in order to significantly relieve the efficiency bottleneck.
—Extensive experiments are conducted to compare the CL-QLTM with several strong baselines. The results show that the CL-QLTM outperforms several strong baselines with respect to a variety of metrics.

The rest of the article is organized as follows. We review the related work in Section 2. In Section 3, we discuss the data format of the multilingual search engine query log utilized in this article. In Section 4, we discuss the CL-QLTM and the strategy of latent parameter inference. In Section 5, we discuss the practical issues of inferring the latent parameters of the CL-QLTM on a voluminous query log. We present the experimental evaluations in Section 6 and conclude the article in Section 7.

2. RELATED WORK

In this section, we review close related work from three categories, probabilistic topic modeling, cross-lingual text mining, and monolingual query log mining.

2.1. Probabilistic Topic Modeling

In recent years, probabilistic topic modeling is recognized as an effective approach to explore the knowledge within data. Blei et al. [2003] proposed Latent Dirichlet Allocation (LDA) to analyze electronic archives. Griffiths and Steyvers [2004] applied LDA to scientific articles and studied its effectiveness in finding scientific topics. By extending LDA, Wang and McCallum [2006] and Blei and Lafferty [2006] presented topic models that capture both the latent structure of data and how the structure changes over time. Recently, topic models have been successfully applied in query intent mining [Jiang et al. 2016], app search [Jiang et al. 2013], semantic information retrieval [Jiang et al. 2015a], and analysis of short text streams [Ren et al. 2013, 2014; Ren and de Rijke 2015; Diao et al. 2012]. Efficient parameter inference is a challenging issue in applying topic modeling. There exists some work about probabilistic topic modeling with parallelized parameter inference algorithms [Newman et al. 2009; Zhai et al. 2012].

There is work focusing on cross-lingual topic modeling for parallel/comparable corpus. Vulić et al. [2011] studied the bilingual LDA model, which was to discover translations of terms in comparable corpora without using any linguistic resources. In particular, the proposed model utilized knowledge from word-topic distributions to obtain the better translation results than that of traditional similarity-based approaches. A model integrating the relevance modeling framework into the topic modeling was proposed in Vulić and Moens [2013] to address monolingual and cross-lingual ad-hoc retrieval. Furthermore, Vulic and Moens [2014] proposed a probabilistic method to modeling Cross-lingual Semantic Similarity (CLSS) in context of comparable data. Vulić et al. [2011] projected words and sets of words into a shared latent semantic space generated by language-pair independent latent semantic concepts. Note that the aforementioned method cannot be straightforwardly transferred to cross-lingual by utilizing a cross-lingual dictionary since language translation is a many-to-many mapping. Simply translating the word in one language to another will distort the original meaning of the source document. Vulić et al. [2015] provides a comprehensive survey on probabilistic topic models that work with parallel and comparable texts and the survey is a overview of representative multilingual modeling from high-level assumptions to mathematical foundations. Mimno et al. [2009] introduced a polylingual topic model that discovers topics aligned across multiple languages. Fukumasu et al. [2012] proposed Symmetric Correspondence LDA that incorporates a hidden variable to control a pivot language in order to support cross-lingual analysis of bilingual text.

2.2. Cross-Lingual Text Mining

Recently, many works have been conducted for cross-lingual text mining. Lavrenko et al. [2002] proposed a model of cross-lingual information retrieval that does not rely on query translation or document translation. Jagarlamudi and Daumé III [2010] proposed a generative model using a bilingual dictionary to mine multilingual topics from an unaligned corpus. Zhang et al. [2010] proposed a way to incorporate a bilingual dictionary into a topic model so that it would enable the model to extract shared latent topics in text data of different languages. Ni et al. [2009] leveraged Wikipedia to help analyze and organize Web information in different languages. In Wang et al. [2008], a method was proposed for classifying non-English queries against an English taxonomy and classifier using widely available machine translation systems. The model in Ambati and Rohini [2006] addressed the problem of building cross-lingual information retrieval systems for language pairs in which the source language is a minority language and the target language is a majority language with existing search engines.

Table II. Examples of Bilingual Search Engine Query Log

QueryID	UserID	Query	Time	Rank	URL
q_1	25014	disney orlando	2012-03-13 11:37:28	1	http://www.innsofcal.com/
q_2	25014	disney resort	2012-03-13 11:38:33		
q_3	45419	microsoft	2012-05-31 21:31:25	2	http://www.microsoft.com/
q_4	51162	迪士尼(disney)	2012-03-08 23:10:31	3	http://http://www.360.cn/...
q_5	63245	微软(microsoft)	2012-04-14 02:21:12	1	www.microsoft.com
q_6	64732	银行信用卡 (bank credit card)	2012-06-11 22:11:35	3	http://http://cc.cmbchina.com/

2.3. Monolingual Query Log Mining

In recent years, some studies have addressed the issue of monolingual query log analysis using probabilistic topic models. Jiang et al. [2012, 2015] proposed a general framework, called Geographical Web Search Topic Discovery (G-WSTD), to discover latent geographic search topics. Particularly, G-WSTD consisted of two core topic models which aim to capture the semantic commonalities across discrete geographic locations and discover web search topics in a specific region, respectively. Furthermore, Jiang et al. [2014] designed a topic model to recommend personalized query suggestions based on diversity awareness of users. Moreover, Jiang et al. [2014] proposed a fast topic discovery algorithm for monolingual query log streams. Jiang et al. [2016] also adopted a query log of users to understand the users' latent intents behind the search queries. Although the aforementioned research focused on the topic of query log mining, they were primarily designed to work only for a monolingual query log and cannot be utilized for deriving topics from a query log in the multilingual scenario. To the best of our knowledge, the present work is the first one that recognizes and takes full advantage of the multilingual salience of a search engine query log. We will show that the proposed CL-QLTM can not only derive cross-lingual topics from a multilingual query log but can also support a wide spectrum of downstream search engine applications.

3. MULTILINGUAL QUERY LOG

In this section, we discuss the format of the multilingual query log utilized in the article. Table III presents notations that we will use throughout this article. Without loss of generality, we utilize a bilingual query log that contains English and Chinese entries to showcase our ideas. As shown in Table II, each entry of the multilingual query log contains the query identifier, the user identifier, the search query, the timestamp, the rank of the clicked URL (if any), and the clicked URL (if any). The techniques proposed in this article can be easily transferred to the scenarios where the query log contains entries in other languages.

There is a subtle linguistic difference between English and Chinese search queries: there exists no space between the Chinese words that work as the basic linguistic units in Chinese. Hence, the Chinese words should be segmented from the search queries. In order to achieve this goal, we apply the state-of-the-art method, called Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) [Zhang et al. 2003], to conduct Chinese word segmentation. For example, the Chinese search query "银行信用卡" is segmented into two Chinese words (i.e., "银行" (bank) and "信用卡" (credit card)) by ICTCLAS.

A common and important phenomenon in web search is search session. Formally, search session is defined as a series of consecutively submitted queries that satisfy the same information need. For example, in Table II, q_1 and q_2 form a session because they are all about Disneyland and they are in temporal proximity. Deriving search sessions from a raw query log is well studied in literature, therefore, we utilize the method in Huang and Efthimiadis [2009] to discover search sessions in a query log.

Table III. Glossary

Notations	Description
F^T	inverse word frequency of translation bipartite
F^C	inverse word frequency of clickthrough bipartite
$N^{L_1}(w)$	the number of words in L_1 that are connected with w
$N^{L_2}(w)$	the number of words in L_2 that are connected with w
$N(u_j)$	the number of query words that are connected with the URL u_j
$\|\mathbf{V_1}\|$	the number of words in L_1
$\|\mathbf{V_2}\|$	the number of words in L_2
$\mathcal{W}(w_i^{L_1}, w_j^{L_2})$	the edge weight between $w_i^{L_1}$ and $w_j^{L_2}$
$\mathcal{W}(w_i, u_j)$	the edge weight between w_i and u_j
θ	topic distribution
D_l	the number of documents in language l
S_d	the number of sentences in document d
z_k	the kth topic
d	a document
w	a word
u	a URL
t	a timestamp
$\mathbf{w}_s$	the words in sentence s
$\mathbf{u}_s$	the URLs in sentence s
$\mathbf{t}_s$	the timestamps in sentence s
L	the language set
R^T	cross-lingual constraint
R^C	clickthrough constraint

We employ two complimentary resources to cross the language barrier in a multilingual query log. These two cross-lingual resources bridge the gaps between different languages from two orthogonal perspectives. The first cross-lingual resource is the dictionary that translates query words from one language into another. The second cross-lingual resource is the query-URL relation, which implicitly captures the semantic similarity between query words in different languages. The cross-lingual dictionary provides static and principled information about the translative relations between query words in different languages. The second cross-lingual resource provides more updated information about the semantic relations between query words in different languages. Therefore, the two resources are complementary and collectively enhance the performance of CL-QLTM.

We represent the information of the first cross-lingual resource via the *Translation Bipartite* (e.g., Figure 1), which captures the translative relations between query words in language L_1 and those in language L_2. A bilingual dictionary built upon the languages L_1 and L_2 represents a many-to-many mapping between the words in the two languages. With the many-to-many mapping, we construct a bipartite $G^T = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{E}^T)$, where $\mathbf{V}_1$ is the vocabulary of the language L_1, $\mathbf{V}_2$ is the vocabulary of L_2, and $\mathbf{E}^T$ is the set of the edges in G^T. An important issue of constructing G^T is to determine the weights for the edges. The weighting scheme of the edges is critical since a word often has multiple meanings in translation. For example, in Figure 1, the word "play" refers to both a verb (e.g., "to play games") or to a noun (e.g., "Shakespeare play"). Therefore, it is more reasonable to weigh the edges of G^T_{ij} differently by considering their distinguishing capabilities. It is intuitive that a word w with a higher relation frequency is less discriminative. This observation motivates us to propose an important concept, referred to as the *inverse word frequency of translation bipartite* (F^T), to

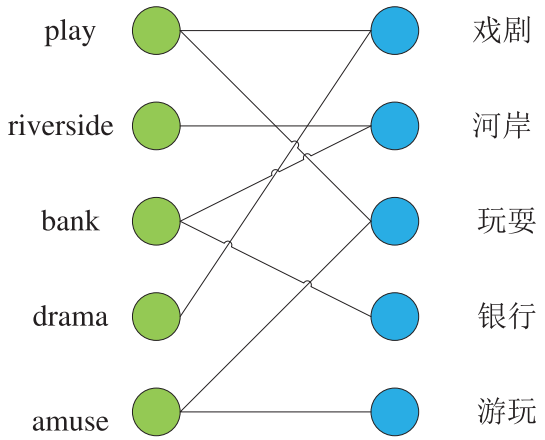


Fig. 1. Translation bipartite graph.

measure the discriminative ability (whose utility is quite like the Inverse Document Frequency (IDF) in information retrieval) of the words in G^T, where G^T_{ij} stands for the edge connecting node i and node j. For instance, suppose $|\mathbf{V}_1|$ is the total number of words in L_1, then the F^T value for the word w in L_2 is defined as:

$$F^T(w) = \log|\mathbf{V}_1| - \log n^{L_1}(w) = \log\frac{|\mathbf{V}_1|}{N^{L_1}(w)}, \tag{1}$$

where $N^{L_1}(w)$ is the total number of words in L_1 that are connected with w and it can be calculated by $N^{L_1}(w) = \sum_{w' \in L_1} 1(w', w)$. Similarly, the F^T value of a word $\hat{w}$ in L_1 can be calculated as:

$$F^T(\hat{w}) = \log|\mathbf{V}_2| - \log n^{L_2}(\hat{w}) = \log\frac{|\mathbf{V}_2|}{N^{L_2}(\hat{w})}. \tag{2}$$

We then calculate the weight of the edge $(w_i^{L1}, w_j^{L2}) \in \mathbf{E}^T$ by multiplying the F^T values of the two words with their co-occurrence frequencies c in a unified way, namely,

$$\mathcal{W}\left(w_i^{L_1}, w_j^{L_2}\right) = c_{ij} \cdot F^T\left(w_i^{L_1}\right) \cdot F^T\left(w_j^{L_2}\right), \tag{3}$$

where the co-occurrence frequency c_{ij} of two words i and j is empirically estimated through counting how many times i is translated into j in a parallel corpus.

The *Translation Bipartite* explicitly captures the translative relations of query words in different languages. We proceed to describe the multilingual *Clickthrough Bipartite* (i.e., Figure 2), which implicitly captures the semantic relations between query words in different languages via the URLs. We denote the bipartite as $G^C = (\mathbf{V}, \mathbf{U}, \mathbf{E}^C)$, where $\mathbf{V}$ is the aggregated vocabulary of query words in different languages, $\mathbf{U}$ is the vocabulary of URLs, and $\mathbf{E}^C$ are the edges of G^C. Through the universal URLs, this bipartite bridges the gap between the query words in different languages. It is worth noting that this bipartite is essentially complementary to the information that is modeled in G^T.

In the G^C, it is intuitive that a heavily clicked URL with a high query word frequency is less discriminative. Hence, we propose the *inverse word frequency of clickthrough bipartite* (F^C) to measure the discriminative ability of the URLs. The F^C of the URL

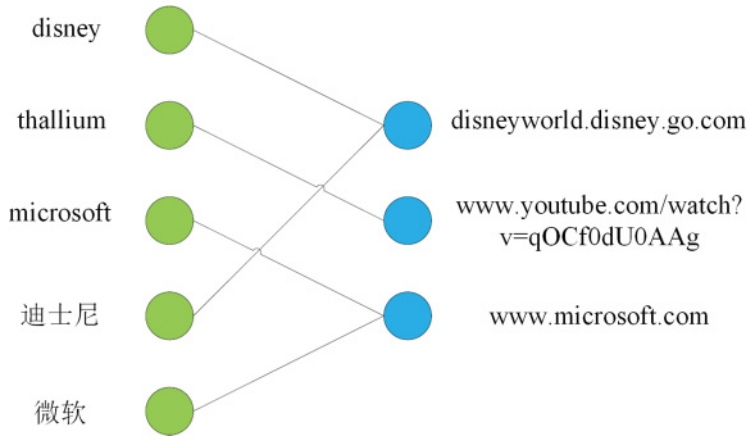


Fig. 2. Clickthrough bipartite graph.

u_j is defined as follows:

$$F^C(u_j) = \log(|\mathbf{V}|) - \log N(u_j) = \log \frac{|\mathbf{V}|}{N(u_j)}, \tag{4}$$

where $N(u_j)$ is the total number of query words that are connected with the URL u_j and it can be calculated by $N(u_j) = \sum_{w_i \in \mathbf{V}} 1(w_i, u_j)$. In G^C, we calculate the weights of the edges by multiplying the inverse query frequencies F^C with the raw frequencies c in a unified way, namely,

$$\mathcal{W}(w_i, u_j) = c_{ij} \cdot F^C(u_j). \tag{5}$$

The intuition behind the above edge weighing mechanism is that different edges are treated differently so that the common relations with less frequent, yet more specific URLs are of greater value than the common relations on frequent URLs.

4. CROSS-LINGUAL QUERY LOG TOPIC MODEL

In this section, we present the details of CL-QLTM. In Section 4.1, we discuss the generative assumptions. In Section 4.2, we discuss the parameter inference.

4.1. Generative Assumptions of CL-QLTM

We have discussed the working horses that are utilized to cross the language barrier in the multilingual query log. We proceed to discuss the generative assumptions of CL-QLTM by considering some unique features of web search data. We consider the query log entries from each user as a document and organize each document as a bag of search sessions. In summary, each user is associated with a document, each document contains several search sessions, and each search session contains query words, URLs (if any), and timestamps.

The generative process of CL-QLTM is presented in Algorithm 1 and its graphical model is shown in Figure 3. We assume that each document is generated by first drawing a document-specific mix θ over K topics. We further assume that query words and URLs in the same session share the same topic. This assumption aligns with the reality: the information in the same session is to satisfy the same information need and, thus, is semantically coherent enough to form a topic. As we constrain that the query words and URLs within a session share the same topic, we utilize a search session as the basic unit for topic assignment. Thus, a session-specific topic z is drawn from θ.

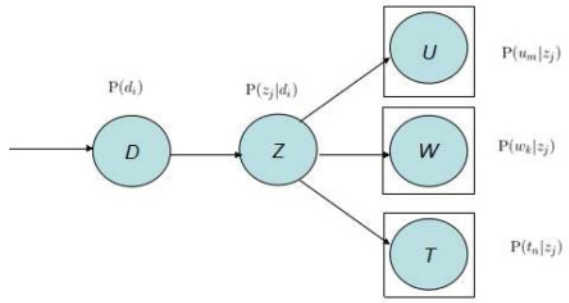


Fig. 3. Graphical model of CL-QLTM.

ALGORITHM 1: Generative Procedure of CL-QLTM

for *each document* $d \in 1, \ldots, D$ **do**
  for *each search session* s *in* d **do**
    choose a topic $z \sim$ Multinomial(θ_d);
    generate query words $w \sim$ Multinomial(ϕ_z);
    generate URLs $u \sim$ Multinomial(Ω_z);
    generate the temporal information $t \sim p(t|z)$;
  end
end

Within the session, some query words are drawn from a Multinomial distribution based on the topic z. The URLs are drawn from another Multinomial distribution based on the topic z. An important but tricky issue lies in the usage of the timestamps in query logs. It is well known that there exist different types of topics in terms of their temporal prominence. The temporal patterns of web search topics can be broadly classified into three types: periodic, background, and bursty. A periodic topic is one that repeats in regular intervals, a background topic is one covered uniformly over the entire period, and a bursty topic is a transient topic that is intensively covered only in a certain time period. The definitions of the three types of temporal prominence are defined as follows:

$$\begin{cases} p^1(t|z) = \dfrac{1}{t_e - t_s}, & (6) \\ p^2(t|z) = \dfrac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z)^2}{\sigma_z^2}}, & (7) \\ p^3(t|z) = \displaystyle\sum_n p(t|z,n)p(n). & (8) \end{cases}$$

The background temporal pattern $p^1(t|z)$ is modeled by a uniform distribution. In $p^1(t|z)$, t_e and t_s are the newest and the oldest timestamps in the query log. The bursty temporal pattern $p^2(t|z)$ is modeled by a Gaussian distribution. The periodic temporal pattern $p^3(t|z)$ is modeled as a mixture of Gaussian distributions. In $p^3(t|z)$, n is the period id, $p(t|z,n) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_z} e^{-\frac{(t-\hat{\mu}_z - nT)^2}{\hat{\sigma}_z^2}}$, and $p(n)$ is uniform in terms of n.

Based on the assumptions discussed above, we can easily see that the topic assignment of a session is subject to the query words, the URLs, and the timestamps within the session. Ultimately, each query word w is picked in proportion to how much the enclosing document prefers the topic z and how much the search topic prefers w. Each URL u is picked in proportion to how much the enclosing document prefers the topic z

and how much the topic prefers u. Each timestamp t is generated in proportion to how much the enclosing document prefers the topic z and how much the topic prefers t.

4.2. Latent Parameter Inference

We proceed to discuss how to conduct a latent parameter inference for CL-QLTM. Assume that a set of K cross-lingual topics need to be discovered from a query log with $|L|$ languages. The log-likelihood of the observed query words, URLs, and timestamps is as follows:

$$\mathcal{L} = \sum_{l=1}^{|L|} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} \log \left[\sum_{k=1}^{K} \prod_{w \in s} P(w|z_k)^{n_{w,s}} \prod_{u \in s} P(u|z_k)^{n_{u,s}} \prod_{t \in s} P(t|z_k) P(z_k|d) \right], \tag{9}$$

where l is a language, d is a document, s is a session, w is a query word, u is a URL, and t is a timestamp. D_l is the number of the documents in language l, S_d is the number of sentences in document d, and K is the number of topics. $n_{w,s}$ is the number of w in s and $n_{u,s}$ is the number of u in s. We utilize a maximum likelihood estimator to estimate the latent parameters by Expectation Maximization (EM). In the maximization step, the objective $E[\mathcal{L}]$ is given as follows:

$$E[\mathcal{L}] = \sum_{l=1}^{|L|} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} \sum_{k=1}^{K} P(z_k|\mathbf{w}_s, \mathbf{u}_s, \mathbf{t}_s) \log \left[\prod_{w \in s} P(w|z_k)^{n_{w,s}} \prod_{u \in s} P(u|z_k)^{n_{u,s}} \prod_{t \in s} P(t|z_k) P(z_k|d) \right], \tag{10}$$

where $P(z_k|\mathbf{w}_s, \mathbf{u}_s, \mathbf{t}_s)$ is obtained from the previous expectation step, i.e.,

$$P(z_k|\mathbf{w}_s, \mathbf{u}_s, \mathbf{t}_s) = \frac{\prod_{w \in s} P(w|z_k)^{n_{w,s}} \prod_{u \in s} P(u|z_k)^{n_{u,s}} \prod_{t \in s} P(t|z_k)^{n_{t,s}} P(z_k|d)}{\sum_{k=1}^{K} \prod_{w \in s} P(w|z_k)^{n_{w,s}} \prod_{u \in s} P(u|z_k)^{n_{u,s}} \prod_{t \in s} P(t|z_k)^{n_{t,s}} P(z_k|d)}. \tag{11}$$

Since CL-QLTM assumes that document-topic relation, topic-word relation, and topic-URL relation are multinomial, the constraints are as follows:

$$\sum_{j=1}^{M_w} p(w_j|z_k) = 1; \sum_{j=1}^{M_u} p(u_j|z_k) = 1; \sum_{k=1}^{K} p(z_k|d_i) = 1. \tag{12}$$

We are faced with an optimization problem with constraints. The corresponding Lagrange function is obtained as follows:

$$\begin{aligned} \mathcal{H} \;=\; & E[\mathcal{L}] + \sum_{k=1}^{K} \tau_k \left(1 - \sum_{j=1}^{M_w} p(w_j|z_k) \right) \\ & + \sum_{k=1}^{K} \tau_k \left(1 - \sum_{j=1}^{M_u} p(u_j|z_k) \right) + \sum_{i=1}^{N} \rho_i \left(1 - \sum_{k=1}^{K} p(z_k|d_i) \right). \end{aligned} \tag{13}$$

We then calculate derivatives with respect to $p(z_k|d_i)$, $p(w_j|z_k)$, and $p(u_j|z_k)$, set the derivatives to zero, and get the following update formulas:

$$P(z_k|d_i) = \frac{\sum_{l=1}^{|L|} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}{S_d}. \tag{14}$$

$$P(w_j|z_k) = \frac{\sum_{l=1}^{|L|}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d} n_{w_j,s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}{\sum_{l=1}^{|L|}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{m=1}^{M} n_{w_m,s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{15}$$

$$P(u_j|z_k) = \frac{\sum_{l=1}^{|L|}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d} n_{u_j,s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}{\sum_{l=1}^{|L|}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{m=1}^{M} n_{u_m,s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{16}$$

After each iteration of the expectation-maximization procedure, we need to update the temporal parameters for the bursty and periodic topics. For a bursty topic z, the parameter μ_z and δ_z are updated accordingly as follows:

$$\mu_z = \frac{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s}) t_t}{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{17}$$

$$\delta_z^2 = \frac{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})(t_t - \mu_z)^2}{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{18}$$

For the periodic topic z, we partition the time line into intervals of length T and assume that each document is only related to its corresponding interval. In other words, $p(t_s|z)$ is set as 0 if the session s is not in the k-th interval. μ_z and δ_z for periodic topic z can be updated according to the following two formulas:

$$\mu_z = \frac{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})(t_t - I_d T)}{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{19}$$

$$\delta_z^2 = \frac{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})(t_t - \mu_z - I_d T)^2}{\sum_{l=1}^{L}\sum_{d=1}^{D_l}\sum_{s=1}^{S_d}\sum_{t}^{T_s} P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{t_s})}. \tag{20}$$

The key idea of CL-QLTM is to compose a cross-lingual topic from different languages by forcing a topic distribution to assign similar probabilities to query words and URLs that are in proximity in G^T and G^D. We achieve this by adding such preferences formally to the likelihood function of CL-QLTM as soft constraints, so that when we estimate the latent parameters of CL-QLTM, we would try to not only fit the web search features well but also try to fit the cross-lingual resources in the *Translation Bipartite* and the *Clickthrough Bipartite* well. Below, we present how we implement this strategy in detail.

4.3. Incorporating Cross-Lingual Resources

Based on G^T, we add a constraint to the likelihood function of CL-QLTM in order to smooth the query word distribution of topics. In this way, we encourage the query words that are connected in G^T to share the same topic. The main extension is to add a cross-lingual constraint R^T to incorporate the knowledge in G^T. R^T is formally defined as follows:

$$R^T = \frac{1}{2}\sum_{(w,w')\in\mathbf{E}^T} \mathcal{W}(w,w') \sum_{j=1}^{K}\left(\frac{p(w|z_j)}{Deg(w)} - \frac{p(w'|z_j)}{Deg(w')}\right)^2, \tag{21}$$

where $\mathcal{W}(w,w')$ is the weight of the edge between w and w' in G^T and $Deg(w)$ is the degree of word w, i.e., the sum of the weights of all the edges ending with w. R^T measures the difference between $p(w|\theta_j)$ and $p(w'|\theta_j)$ for each pair (w,w') in a bilingual

dictionary; the more they differ, the larger R^T would be. So it can be regarded as a "loss function" to help us assess how well the word distributions in multiple languages are correlated semantically. Clearly, we would like the extracted topics to have a small R^T. We choose this specific form of loss function because it would make it convenient to solve the optimization problem of maximizing the corresponding regularized maximum likelihood. Similarly, we introduce the clickthrough constraint R^C to capture the knowledge modeled in G^C. R^C is defined as follows:

$$R^C = \frac{1}{2} \sum_{(w,u)\in \mathbf{E}^C} \mathcal{W}(w,u) \sum_{j=1}^{K} \left(\frac{p(w|z_j)}{Deg(w)} - \frac{p(u|z_j)}{Deg(u)} \right)^2, \tag{22}$$

where $\mathcal{W}(w,u)$ is the weight of the edge between the query word w and the URL u in G^C.

Putting $\mathcal{L}$, R^T, and R^C together, we maximize the following objective function $\mathcal{O}$ as:

$$\mathcal{O} = \mathcal{L} - \alpha R^T - \beta R^C, \tag{23}$$

where α and β are parameters to balance the likelihood and the influence of two cross-lingual resources. We will search for a set of values for all the latent parameters that can maximize $\mathcal{O}$. After incorporating the cross-lingual information in G^T and G^C, there is no closed form solution in the maximization step for the whole objective function. Hence, the traditional EM algorithm cannot be applied. We now discuss how to solve this problem by the Generalized Expectation-Maximization (GEM) algorithm [Hebert and Leahy 1989]. The major difference between EM and GEM lies in the maximization step. Instead of finding the globally optimal solution ψ, which maximizes the expected complete data log-likelihood $\mathcal{O}(\psi)$ in the maximization step of EM algorithm, GEM only needs to find a better ψ in each new iteration. Let ψ_n denote the parameter values of the previous iteration and ψ_{n+1} denote the parameter values of the current iteration. The convergence of the GEM algorithm only requires $\mathcal{O}(\psi_{n+1}) \geq \mathcal{O}(\psi_n)$. Hence, our method is to maximize $\mathcal{L}$ and then gradually decrease R^T and R^C by the Newton-Raphson method. If there is no ψ_{n+1} subject to $\mathcal{O}(\psi_{n+1}) > \mathcal{O}(\psi_n)$, then we consider ψ_n to be the local maximum point of the objective function. The algorithm for optimizing parameters via R^T and R^C is formalized in Algorithms 2 and 3. We sequentially apply the two algorithms to update the latent parameters in CL-QLTM until convergence is achieved.

ALGORITHM 2: Optimizing Parameters By R^T

Input: Parameters ψ_{n+1}; Newton step parameter γ^T; $p(w|z_j)_{n+1}$

Output: $p'(w|z_j)_{n+1} p(w|z_j)^{(1)}_{n+1} \leftarrow p(w|z_j)_{n+1}; p(w|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^T) p(w|z_j)^{(1)}_{n+1} +$
$\gamma^T \frac{\sum_{(w,w')\in E^T} \mathcal{W}(w,w') p(w'|z_j)^{(1)}_{n+1}}{\sum_{(w,w')\in E^T} \mathcal{W}(w,w')}$;

for $\mathcal{O}(\psi^{(2)}_{n+1}) \geq \mathcal{O}(\psi^{(1)}_{n+1})$ **do**

  $p(w|z_j)^{(1)}_{n+1} \leftarrow p(w|z_j)^{(2)}_{n+1}; p(w|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^T) p(w|z_j)^{(1)}_{n+1} +$
  $\gamma^T \frac{\sum_{(w,w')\in E^T} \mathcal{W}(w,w') p(w'|z_j)^{(1)}_{n+1}}{\sum_{(w,w')\in E^T} \mathcal{W}(w,w')}$;

end

if $\mathcal{O}(\psi^{(1)}_{n+1}) \geq \mathcal{O}(\psi_n)$ **then**

  $p'(w|z_j)_{(n+1)} \leftarrow p(w|z_j)^{(1)}_{n+1}$;

end

Return $p'(w|z_j)_{n+1}$;

ALGORITHM 3: Optimizing Parameters By R^C

Input: Parameters ψ_{n+1}; Newton step parameter γ^C;
$p(w|z_j)_{n+1}$; $p(u|z_j)_{n+1}$
Output: $p'(w|z_j)_{n+1}$; $p'(u|z_j)_{n+1}$
$p(w|z_j)^{(1)}_{n+1} \leftarrow p(w|z_j)_{n+1}$; $p(u|z_j)^{(1)}_{n+1} \leftarrow p(u|z_j)_{n+1}$;
$p(w|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^C)p(w|z_j)^{(1)}_{n+1} + \gamma^C \frac{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)p(u|z_j)^{(1)}_{n+1}}{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)}$;
$p(u|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^C)p(u|z_j)^{(1)}_{n+1} + \gamma^C \frac{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)p(w|z_j)^{(1)}_{n+1}}{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)}$;
for $\mathcal{O}(\psi^{(2)}_{n+1}) \geq \mathcal{O}(\psi^{(1)}_{n+1})$ **do**
  $p(w|z_j)^{(1)}_{n+1} \leftarrow p(w|z_j)^{(2)}_{n+1}$; $p(u|z_j)^{(1)}_{n+1} \leftarrow p(u|z_j)^{(2)}_{n+1}$; $p(w|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^C)p(w|z_j)^{(1)}_{n+1} +$
  $\gamma^C \frac{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)p(u|z_j)^{(1)}_{n+1}}{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)}$; $p(u|z_j)^{(2)}_{n+1} \leftarrow (1-\gamma^C)p(u|z_j)^{(1)}_{n+1} +$
  $\gamma^C \frac{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)p(w|z_j)^{(1)}_{n+1}}{\sum_{(w,u)\in E^C} \mathcal{W}(w,u)}$;
end
if $\mathcal{O}(\psi^{(1)}_{n+1}) \geq \mathcal{O}(\psi_n)$ **then**
  $p'(w|z_j)_{(n+1)} \leftarrow p(w|z_j)^{(1)}_{n+1}$; $p'(u|z_j)_{(n+1)} \leftarrow p(u|z_j)^{(1)}_{n+1}$;
end
Return $p'(w|z_j)_{n+1}$ and $p'(u|z_j)_{n+1}$

4.4. Complexity Analysis

In this section, we systematically investigate the complexity of the CL-QLTM parameter inference. For each iteration of the EM algorithm, all the K topics need to be scanned for each word, URL, and timestamp. Hence, the complexity of each iteration is $O(K(W + U + T))$, where W is the number of words and U is the number of URLs and T is the number of timestamps.

4.5. Moving from Bilingual to Multilingual

Although most of the above discussion is described in a bilingual scenario, the proposed techniques can be easily transferred to multilingual scenarios. Note that Equation (9) itself supports $|L|$ languages. The flexibility of CL-QLTM lies in its design: the likelihood of Equation (9) models the domain knowledge (such as query words, URLs, sessions, etc) in web search while the cross-lingual resources are further introduced in the fashion of regularization. Hence, in order to make the CL-QLTM support more than two languages, we only need to introduce more cross-lingual regularization in a fashion analogous to that described in Section 4.3.

5. EFFICIENCY ISSUES

A multilingual query log is typically voluminous and the efficiency of training is a critical issue when the CL-QLTM is applied in real-life scenarios. When we apply the proposed model to large data sets, we find that efficiency is the obstacle for applying the CL-QLTM in real-life scenarios. To address the efficiency bottleneck, we discuss how to deploy the latent parameter learning on a distributed system under the MapReduce programming paradigm. MapReduce is a programming paradigm for distributed processing of large data sets [Dean and Ghemawat 2008]. In the map stage, each process node receives a subset of data as input and produces a set of intermediate key/value pairs. In the reduce stage, each process node merges all intermediate values associated with the same intermediate key and outputs the final computation results.

We partition the training data into subsets and distribute each subset to a process node. For the CL-QLTM, the parameter inference is a pipeline containing three

Table IV. The Key/Value Pairs Emitted at the Map Stage of the First MapReduce Job

Key	Value
(z_k, w_j)	$Value_1 = \sum_{l=1}^{L} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} n_{w_j,s} p_{sz_k}$ $Value_2 = \sum_{l=1}^{L} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} \sum_{m=1}^{M} n_{w_m,s} p_{sz_k}$
(z_k, u_j)	$Value_3 = \sum_{l=1}^{L} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} n_{u_j,s} p_{sz_k}$ $Value_4 = \sum_{l=1}^{L} \sum_{d=1}^{D_l} \sum_{s=1}^{S_d} \sum_{m=1}^{M} n_{u_m,s} p_{sz_k}$
z_k	t_{z_k}

Table V. The Key/Value Pairs Emitted at the Map Stage of the Second MapReduce Job

Key	Value
(z_k, w_j)	$Value = (1-\gamma) p(w_j \mid z_k)$
(z_k, w')	$Value = \gamma \frac{\mathcal{W}(w_j, w')}{Deg(w')} p(w_j \mid z_k)$

Table VI. The Key/Value Pairs Emitted at the Map Stage of Phase III

Key	Value
(z_k, w_j)	$Value = (1-\gamma) p(w_j \mid z_k)$
(z_k, u)	$Value = \gamma \frac{\mathcal{W}(w_j, u)}{Deg(u)} p(w_j \mid z_k)$
(z_k, u_j)	$Value = (1-\gamma) p(u_j \mid z_k)$
(z_k, w)	$Value = \gamma \frac{\mathcal{W}(u_j, w)}{Deg(w)} p(u_j \mid z_k)$

consecutive MapReduce jobs. In the first job, we update the parameters according to Equations (14) to (20). The second and the third jobs correspond to Algorithm 2 and Algorithm 3, respectively. A controlling process spawns the MapReduce jobs and keeps track of the number of iterations and convergence criteria. Model parameters, which are static for the duration of the MapReduce jobs, are loaded by each mapper from HDFS. In the map stage, each process node scans the assigned subset of training data once.

For each search session s and each topic k, the process node infers the posterior probability $p_{sk} = P(z_k|\mathbf{w_s}, \mathbf{u_s}, \mathbf{u_s})$ by Equation (11) for each session and emits the key/value pairs as shown in Table IV. In the reduce stage, each process node collects all values for an intermediate key. For example, suppose the intermediate key (z_k, w_j) is assigned to process node n. Then, n receives a list of values $(<Value_1>; <Value_2>)$ and derives $P(w_j|z_k)$ by $\frac{\sum_i Value_{i,1}}{\sum_i Value_{i,2}}$. Similarly, we derive $P(u_j|z_k)$ by $\frac{\sum_i Value_{i,3}}{\sum_i Value_{i,4}}$. The temporal parameters are updated according to Equations (17) to (20) based on the values $<t_{z_k}>$.

As shown in Table V, in the second MapReduce job, each word w_j emits two types of entries with respect to a topic k. The first one is the $(1-\gamma)p(w_j|z_k)$. The second is a set of the neighboring words of w_j, with respect to the topic k. In the reduce stage, each process node simply collects all values for an intermediate key, adds them together, and obtains the updated $p'(w|z_j)_{n+1}$. Similarly, we can implement the third MapReduce job according to the key/value pairs defined in Table VI.

6. EXPERIMENTS

In this section, we evaluate the performance of the proposed CL-QLTM. In Section 6.1, we describe the experimental setup. In Section 6.2, we present some topic examples. In Section 6.3, we quantitatively evaluate the CL-QLTM with several metrics. In Section 6.4, we gauge the efficiency of the parallel training procedure of the CL-QLTM.

Table VII. Examples of Cross-Lingual Topics Derived by PLSA

Olympic		Software		Finance		Legislation	
湖(lake)	0.006204	下载(download)	0.145538	查询(query)	0.040177	court	0.007289
Olympic	0.005862	官方版(official version)	0.020653	快递(express package)	0.016831	case	0.006140
公园(park)	0.004451	360	0.015811	信用卡(credit card)	0.010818	charges	0.004310
Athens	0.004280	免费(free)	0.012987	中心(center)	0.009374	legal	0.004274
race	0.004023	浏览器(browser)	0.011484	加盟(join)	0.007890	trial	0.003987
running	0.003510	安全(security)	0.008293	交通(trafic)	0.007539	law	0.003808
gold	0.003254	完整(complete)	0.007266	单号(serial number)	0.006836	drugs	0.003198
world	0.002783	ppt	0.006642	圆通(flexible express)	0.006172	guilty	0.003162
新区(new district)	0.002741	模板(template)	0.005102	公积金(public accumulation fund)	0.005899	investigation	0.002372
Greek	0.002741	剧场版(theatrical version)	0.004552	工商(industrial and commercial)	0.005665	action	0.002372
European	0.002612	中文版(chinese version)	0.003708	个人(personal)	0.004250	lawyer	0.002336

Table VIII. Examples of Cross-Lingual Topics Derived by LDA

Internet		Software		Education		Copyright	
qq	0.050669	下载(download)	0.144243	大学(university)	0.036280	piracy	0.006275
云(cloud)	0.048862	迅雷(Thunder)	0.035245	管理(management)	0.015565	files	0.005001
邮箱(email box)	0.024222	官方(official)	0.020552	科技(technology)	0.015150	legal	0.004470
地图(map)	0.017840	软件(software)	0.016988	学院(school)	0.014804	P2P	0.004364
视频(video)	0.015111	免费(free)	0.015315	信息(information)	0.013663	BitTorrent	0.004258
优酷(youku)	0.013881	破解(crack)	0.013060	技术(technology)	0.010067	file-sharing	0.004151
产品(product)	0.013501	完整(complete)	0.008187	上海(shanghai)	0.009271	copyright	0.003727
登录(login)	0.010806	中文版(chinese version)	0.006550	工程(engineering)	0.008891	peer-to-peer	0.003196
空间(space)	0.010191	mac	0.005277	电子(electronics)	0.007742	industry	0.003196
翻译(translate)	0.009845	itunes	0.004477	网络(network)	0.007715	music	0.002771
歌曲(song)	0.009845	榜(list)	0.004477	研究生(graduate student)	0.007024	networks	0.002665

Finally, we demonstrate three downstream applications of the cross-lingual topics discovered by the CL-QLTM in Section 6.5.

6.1. Experiment Setup

The data set we utilized is a bilingual query log, which is obtained from a major commercial search engine. The query log contains both English and Chinese search queries submitted by 1 million users within 3 months. The data set contains about 2.4 million English search queries, 11.6 million Chinese search queries, and 5.1 million search sessions.

To process the Chinese search queries, we use the ICTCLAS [Zhang et al. 2003] to segment search queries into Chinese phrases. Both Chinese and English stopwords are removed from our data. The cross-lingual dictionary we utilized is from Chinese-English dictionary (CEDICT)[2]. For each Chinese phrase, if it has several English meanings, we add an edge between it and each of its English translations in G^T. If one English translation is an English phrase, we add an edge between the Chinese phrase and each English word in the phrase in G^T.

6.2. Topic Examples

Some topic examples are presented in Tables VII, VIII, IX, and X. To enhance the readability, we add an English translation to each Chinese word/phrase in our results. We observe that PLSA and LDA primarily derive monolingual topics, since they are not designed with cross-lingual capability. For example, in Table VIII, the content about the topic "Software" derived by LDA is mainly composed by Chinese words such as "下载" (download), "官方" (official), and so on. However, occasionally, their topics contain words from different languages due to the phenomenon of cross-lingual occurrence in data. For example, the same topic "Software" derived by LDA contains English words such

[2]http://cgibin.erols.com/mandarintools/cedict.html. Note that it may be further enhanced by some advanced semantic networks, such as BabelNet (www.babelnet.org).

Table IX. Examples of Cross-Lingual Topics Derived by PCLSA

Education		Computer		Movie		Game	
大学(university)	0.024319	电脑(PC)	0.027761	电影(movie)	0.057807	游戏(game)	0.044689
上海(Shanghai)	0.017279	设置(configuration)	0.015960	picture	0.021074	qq	0.035999
2016	0.013023	install	0.011738	导演(director)	0.011790	邮箱(email box)	0.017300
北京(Beijing)	0.012931	system	0.011306	美国(US)	0.010632	video	0.011745
招聘(recruit)	0.011965	win10	0.010342	韩国(Korea)	0.008237	login	0.007730
center	0.011620	密码(password)	0.008015	cartoon	0.006784	163	0.007482
college	0.011218	笔记本(notebook)	0.007117	拍摄(record)	0.005988	空间(space)	0.007180
科技(technology)	0.010102	硬盘(hard disk)	0.006984	排行榜(list)	0.005619	网易(Netease)	0.006743
information	0.009204	driver	0.006984	women	0.005385	account	0.006274
服务(service)	0.005823	desktop	0.004491	brand	0.004793	活动(activity)	0.005702

Table X. Examples of Cross-Lingual Topics Derived by CL-QLTM

Education		NBA		Election		Real Estate	
university	0.041848	basketball	0.011204	election	0.060606	estate	0.032281
high	0.035745	players	0.007469	consulate	0.045454	real	0.032281
院系(department)	0.030514	NBA	0.005602	politics	0.015151	区域(district)	0.008608
city	0.010462	76er	0.005602	vote	0.015151	hotel	0.006456
tuition	0.010462	明星(star)	0.004668	总统(president)	0.015151	sale	0.006456
school	0.006974	队伍(team)	0.003734	corruption	0.005988	price	0.004098
college	0.005231	Kobe	0.002801	共和党(republican)	0.005988	dealers	0.004098
图书馆(library)	0.004359	championship	0.002801	money	0.004491	保险(insurance)	0.002732
grant	0.004359	final	0.001867	governor	0.004491	apartment	0.023224
注册(enrollment)	0.003487	offseason	0.001867	委员会(council)	0.002994	经纪人(brokers)	0.002049

Table XI. Topical Coherence

Model	top-5 coherence	top-10 coherence	top-15 coherence
PLSA	2.16	1.96	1.32
LDA	2.17	1.96	1.33
PCLSA	2.17	1.95	1.32
CL-QLTM	2.19	1.97	1.33

as “itunes” and “mac.” Compared with the baselines, the CL-QLTM can not only find coherent topics from the cross-lingual corpus, but it can also show the content about one topic from two languages. For example, in **Education**, which is about “university” and “department,” the Chinese search queries mention a lot about “department” and “library” while the English search queries discuss more on topics such as “tuition” and “city.” Similarly, in **NBA**, the Chinese search queries mention some teams and stars in the NBA, while the English search queries mention a lot about “offseason.” Similar results can also be observed from the topics **Election** and **Real Estate**. The empirical results discussed above showcase the output of the CL-QLTM and reveal some basic features of the discovered cross-lingual topics.

In order to further quantify the topical coherence, we hired five human experts to help label the coherence of the discovered topics. The labels range from 0 to 3, where 0 indicates the words have no coherence at all and 3 indicates the topical coherence is high. From each topic, the human experts are required to gauge the coherence of its top-5, top-10, and top-15 words. The results listed in Table XI are the top-5, top-10, and top-15 coherences averaged on 1,000 randomly chosen topics. We find that the topical coherence of PLSA and LDA is quite high. The reason is relatively straightforward: although LDA has no cross-lingual capability, it can still find coherent topics within each single language. Another insightful observation comes from comparing the CL-QLTM and PCLSA, which derive similar cross-lingual topics. We find that the CL-QLTM always outperforms PCLSA, showing that the CL-QLTM is more effective for the scenario of analyzing web search data.

6.3. Quantitative Evaluation

In this section, we utilize several standard metrics to evaluate the CL-QLTM. As a baseline method, we apply the PCLSA [Zhang et al. 2010], which is a pioneering cross-lingual topic model for multilingual corpus.

We perform significance testing using the 10-fold cross-validated paired t-test on a data set that contains 10,000 search queries to evaluate the models' capability of predicting unseen data. The differences between perplexity are considered statistically significant for p-values lower than 0.05. Perplexity is a standard measure of evaluating the generalization performance of a probabilistic model [Rosen-Zvi et al. 2004]. It is monotonically decreasing in the likelihood of the held-out data. Therefore, a lower perplexity indicates better generalization performance. Specifically, perplexity is calculated according to the following equation:

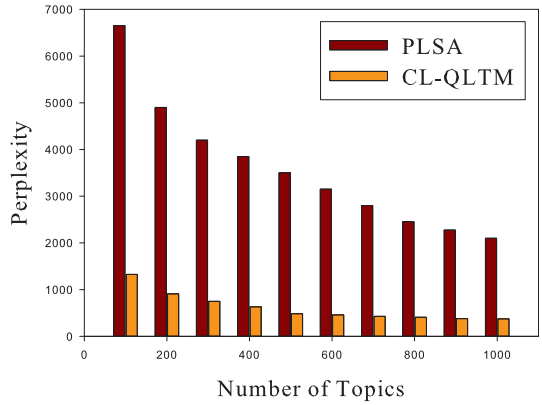
$$Perplexity_{heldout}(\mathcal{M}) = \left(\prod_{d=1}^{D} \prod_{i=1}^{N_d} p(w_i|\mathcal{M}) \right)^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}, \tag{24}$$

where $\mathcal{M}$ is the model learned from the training process. The differences between the perplexity of two models are statistically significant ($p < 0.01$), and the best results are given in Figure 4, from which we observe that the proposed model demonstrates much better capability in predicting unseen data compared with the baselines, such as LDA and PCLSA. For example, when the number of search topics is set to 1,000, the perplexity of PCLSA is 1,052. The CL-QLTM significantly reduces the perplexity and achieves a perplexity of 378. The result shows that the CL-QLTM is more suitable for analyzing multilingual web search data.

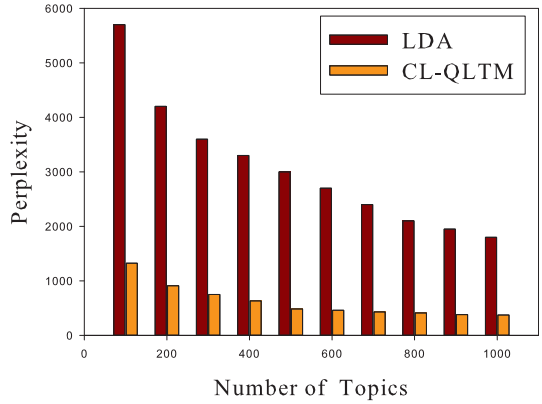
Another metric is defined for gauging how effective the proposed models are in predicting the remaining query terms after observing a portion of the user's search history. Suppose we observe the query terms $w_{1:P}$ from a user's query log; we are interested in finding which model provides a better predictive distribution $p(w|w_{1:P})$ of the remaining query terms. We use Equation (25) to calculate the perplexity of the testing data. The comparison results are presented in Figure 5. We also use the paired t-test, and the differences between this metric are considered statistically significant for p-values lower than 0.05. We observe that the proposed models, again, significantly outperform the three baselines ($p < 0.01$). When the observed data is set to 90%, PCLSA demonstrates a perplexity of 767 and the CL-QLTM shows a perplexity of 185. The result suggests that the CL-QLTM has better capability to predicting the user's future web search given the user's search history.

$$Perplexity_{portion}(\mathcal{M}) = \left(\prod_{d=1}^{D} \prod_{i=P+1}^{N_d} p(w_i|\mathcal{M}, w_{a:P}) \right)^{\frac{-1}{\sum_{d=1}^{D}(N_d-P)}}. \tag{25}$$

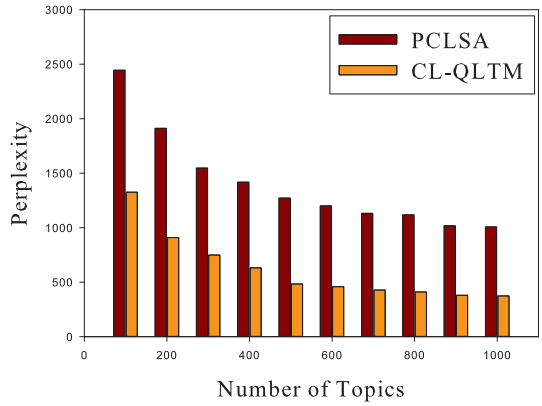
The third metric that we use for evaluation is the Kullback–Leibler divergence (KL-Divergence) between discovered search topics. Similar to Yin et al. [2011], we utilize KL-divergence to evaluate the distinctiveness of discovered search topics. The larger the average KL-divergence is, the more distinct the topics are. We perform significance testing using the paired t-test, and we show the average distance of term distributions of all pairs of search topics measured by KL-divergence in Figure 6(a). The differences between this metric are considered statistically significant for p-values lower than 0.05. For this metric, we only utilize PCLSA as the baseline. PLSA and LDA only find topics in a single language, hence, the KL-divergence between topics in different languages exaggerates their performance in discovering distinct topics. Based on the experimental results, we find that the KL-divergence of CL-QLTM is much higher



(a) Perplexity Comparison with PLSA

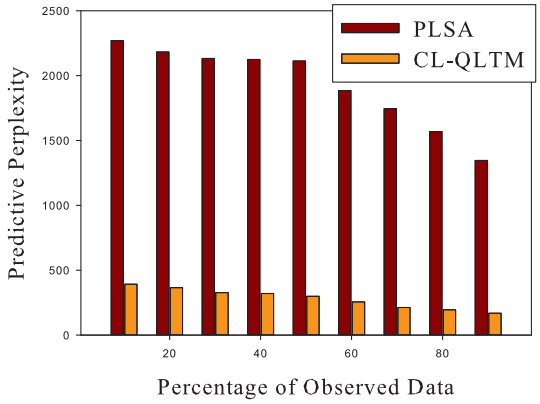


(b) Perplexity Comparison with LDA

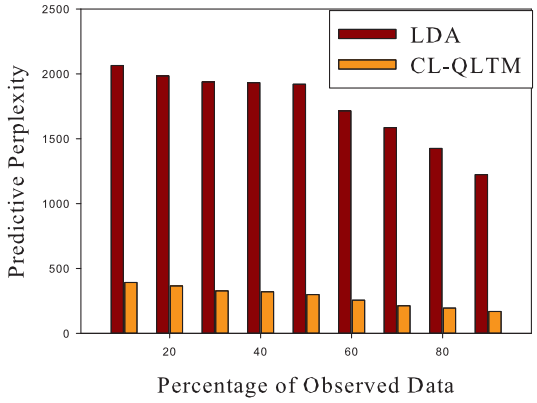


(c) Perplexity Comparison with PCLSA

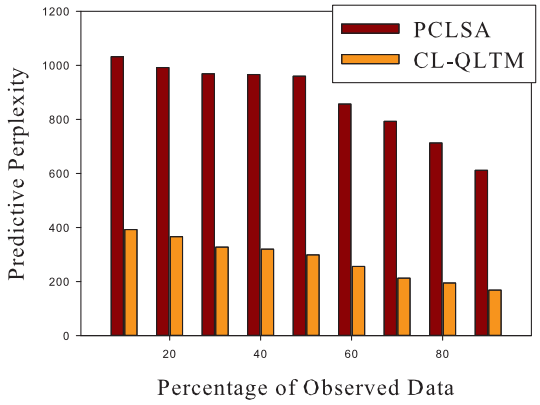
Fig. 4. Perplexity evaluation.



(a) Predictive Perplexity Comparison with PLSA

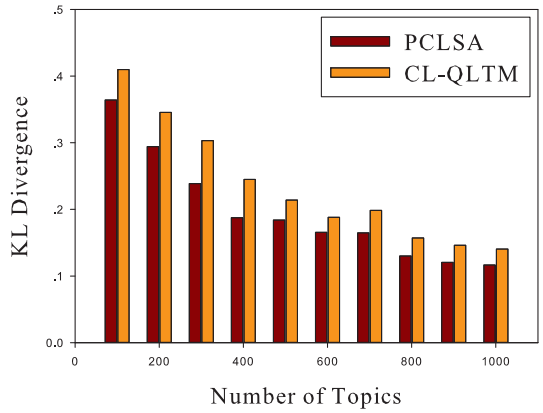


(b) Predictive Perplexity Comparison with LDA

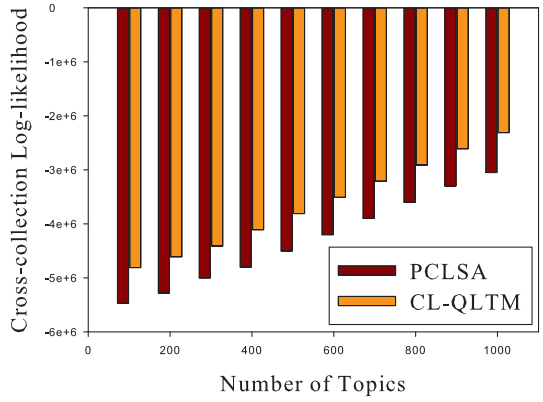


(c) Predictive Perplexity Comparison with PCLSA

Fig. 5. Predictive perplexity for partially observed data.



(a) KL Divergence



(b) Cross-collection likelihood

Fig. 6. Quantitative evaluation of KL divergence and cross-collection likelihood.

than the PCLSA. The word distributions in the search topics discovered by CL-QLTM are more distinctive than those obtained by the baseline. The result indicates that CL-QLTM is effective to find different facets of the location commonalities.

We further evaluate how well CL-QLTM can discover common topics among corpus in different languages. We utilize the "cross-collection" likelihood measure [Zhang et al. 2010] for this purpose. To make this article self-explained, we present the basic idea as follows: suppose we got k cross-lingual topics from the multilingual query log; then, for each topic, we split the topic into two separate sets of topics—English topics and Chinese topics. Then, we use the word distribution of the Chinese topics (translating the words into English) to fit the English search queries and use the word distribution of the English topics (translating the words into Chinese) to fit the Chinese search queries. If the topics are common topics in the whole cross-lingual query log, then the "cross-collection" likelihood should be larger than those topics which are not commonly shared by the English and the Chinese search queries. To calculate the likelihood of fitness, we use the folding-in method proposed in Hofmann [2001]. To translate topics from one language to another, e.g., Chinese to English, we look up the bilingual dictionary and do word-to-word translation. If one Chinese word has several English translations,

we simply distribute its probability mass equally to each English translation. For comparison, we use the standard PCLSA model as the baseline. Basically, suppose PCLSA mined k topics in the Chinese search queries and k topics in the English search queries. Then, we also use the "cross-collection" likelihood measure to see how well the k Chinese topics fit the English search queries and those k semantic English topics fit the Chinese search queries. We also use the paired t-test, and the differences between this metric are considered statistically significant for p-values lower than 0.05. The experimental results are shown in Figure 6(b). From the results, we can see that CL-QLTM has a higher "cross-collection" likelihood, meaning that it significantly ($p < 0.01$) finds better common topics compared to the baseline PCLSA.

The aforementioned metrics systematically demonstrate the superiority of CL-QLTM in analyzing a cross-lingual query log through different dimensions. The experimental results univocally show that the CL-QLTM provides a better fit for web search data and it is more suitable for the scenarios of multilingual query log analysis.

6.4. Efficiency Evaluation

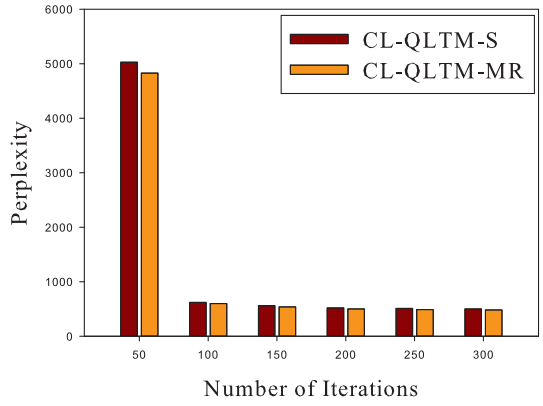
To systematically evaluate the parallel parameter interference algorithms discussed in Section 5, we measure their performance using held-out dataset perplexity like Newman et al. [2009]. The topic distribution of each document is learned using the training part and perplexity is computed using this distribution and words from the testing part. We compared CL-QLTM-S (i.e., CL-QLTM trained on a single processor) with CL-QLTM-MR (i.e., CL-QLTM trained on MapReduce). Figure 7(a) shows that, for a fixed number of topics (i.e., $k = 500$), the perplexity of converged results is quite close no matter whether we use CL-QLTM-S or CL-QLTM-MR (the number of computing nodes $|P|$ is 10). For example, when 50 iterations are conducted, CL-QLTM-S and CL-QLTM-MR achieve slightly different perplexity. When the iteration reaches 300, both CL-QLTM-S and CL-QLTM-MR converge and achieve the perplexity of about 480. Similar results can also be observed when varying the number of topics k and the number of computing nodes $|P|$ for CL-QLTM-MR. It is worth emphasizing that, despite no theoretical convergence guarantees, CL-QLTM-MR converges to a good solution in every run we did.

To properly determine the utility of the distributed algorithms, it is necessary to check whether the CL-QLTM-MR is systematically converging more slowly than their single processor counterparts. If this were the case, it would mitigate the computational gains of parallelization. In fact, our experimental results in Figure 7(b) consistently show that the convergence rate for the distributed algorithms is much faster than the single processor cases. For example, when the number of computing nodes $|P|$ is 10, CL-QLTM-MR converges after 150 minutes, while CL-QLTM-S consumes roughly 900 minutes, demonstrating a speedup of about 6. The major overhead of CL-QLTM-MR involves the data IO after each MapReduce iteration. In summary, CL-QLTM-MR is able to learn models whose predictive performance is similar to that of CL-QLTM-S. CL-QLTM-MR yields significant speedup in practice, making it scalable on massive multilingual query logs.

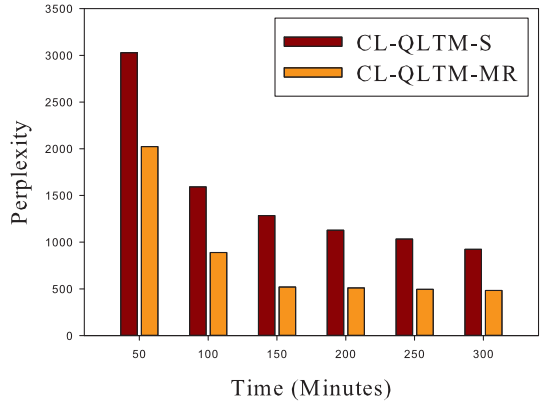
6.5. Applications of CL-QLTM

In this section, we discuss how to apply the CL-QLTM to search engine applications such as cross-lingual query recommendation, cross-lingual URL recommendation, and cross-lingual information retrieval, which showcase the potential utility of the CL-QLTM.

6.5.1. Cross-Lingual Query Recommendation. Cross-lingual query recommendation is an important application of web search engines [Gao et al. 2007]. The logic behind our



(a) Perplexity Against Number of Iteration



(b) Perplexity Against Relative Elapsed Time

Fig. 7. Efficiency evaluation.

approach of cross-lingual query recommendation is detailed as follows: the cross-lingual topics are utilized to evaluate the topic-based similarity between the query in language S and that in another language T. Queries with high topic-based similarities are identified as the query recommendation candidates. These query recommendation candidates are further ranked according to the topic-based similarity with respect to the original query in language S, and the ranking list of queries in language T is presented to the end user. Formally, the formula of calculating the topic-based similarity is defined as follows:

$$Score(q_S, q_T) = \sum_{k=1}^{K} \sum_{w \in q_S, w' \in q_T} p(w, w'|\theta_k), \tag{26}$$

where q_S is the search query in language S and q_T is the search query in language T. Essentially, we need to suggest the top-N queries with the highest probabilities, where N is a user-specified parameter.

In order to evaluate the effectiveness of the cross-lingual query recommendation, we prepare six baselines as follows. In the first baseline (i.e., Translation Baseline), the

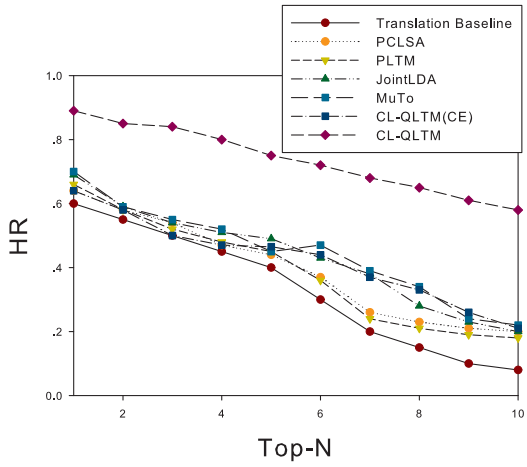
English query words are first translated into Chinese words, then we select the Chinese queries which have the highest cosine similarity with the Chinese words as the query recommendation candidates. The second baseline (i.e., PCLSA) is straightforwardly implemented by the topics generated by PCLSA and Equation (26). The third and fourth baselines are the PLTM [Mimno et al. 2009] and JointLDA [Jagarlamudi and Daumé III 2010], which are trained based on parallel English/Chinese Wikipedia corpus. The fifth baseline is MuTo [Boyd-Graber and Blei 2009], which uses dictionary as the measurement for word distance. Finally, the baseline CL-QLTM(CE) is a variant of CL-QLTM by simply considering English and Chinese words as being from the same language.

For the purpose of performance evaluation, we recruit five human experts who give explicit relevance evaluation of the query suggestion lists. We design a metric named *Human Relevance* (HR) to evaluate the effectiveness of the ranking from the users' explicit feedbacks. Similar to Leung et al. [2010], a web search middleware is implemented to record the experts' feedback. The human experts are required to submit search queries to middleware and rate the suggested queries on a 6-point scale (0, 0.2, 0.4, 0.6, 0.8, and 1), where 0 means "totally irrelevant" and 1 indicates "entirely relevant." We report the average HR in Figure 8(a). We can see that the CL-QLTM generates query recommendations that better align with the users' latent information needs and it significantly outperforms the baselines with respect to the HR. Particularly, the HR of the CL-QLTM is at least 1.5 times better than that of the other five baselines when $N \geq 2$. The results support our idea that a better cross-lingual query suggestion paradigm should be able to utilize the underlying relations between query words in different languages. The CL-QLTM captures the underlying semantic similarity between query words in different languages and it is more suitable for the web search scenario. Hence, CL-QLTM outperforms the five baselines in the task of cross-lingual query recommendation. Furthermore, JointLDA and CL-QLTM(CE) are the runner-ups when $N \leq 5$ and $N > 5$, respectively, and Translation Baseline is the worst. Also, the HR of all the algorithms decrease as N increases overall, which is because the difference between the result of cross-lingual query recommendation and the result from human experts becomes larger as N becomes larger. Meanwhile, the gap of HR between the algorithms becomes wider as N increases, which also verifies that the effectiveness of CL-QLTM is more stable.

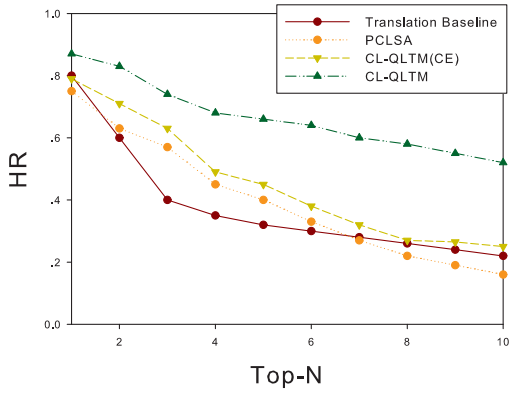
6.5.2. Cross-Lingual URL Recommendation. In the task of cross-lingual URL recommendation, for each search query q in language S, we recommend the URLs whose corresponding web pages are written in the language T. The logic of cross-lingual URL recommendation is similar to the cross-lingual query recommendation. We utilize the cross-lingual topics to calculate the similarity between the query words and the URLs. The URLs are ranked according to their topic-based similarity, with respect to the query. The formula of calculating the similarity between a query q and a URL u is defined as follows:

$$Score(q, u) = \sum_{k=1}^{K} \sum_{w \in q_S, u \in U_T} p(w, u|\theta_k). \tag{27}$$

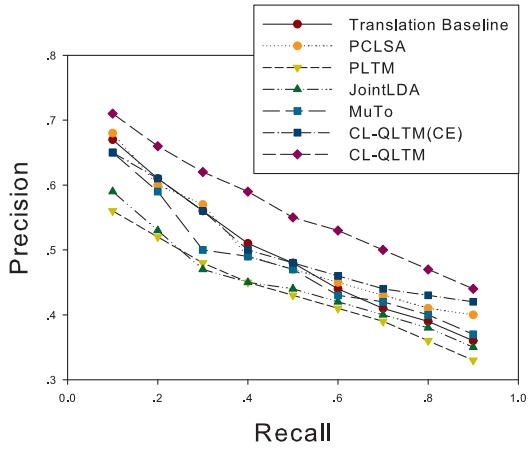
Essentially, we need to suggest the top-N URLs with the highest probabilities, where N is a user-specified parameter. To evaluate the effectiveness of the proposed method in the cross-lingual URL recommendation, we design three baselines. The first baseline is Translation Baseline, which first translates the English query words into Chinese words, then it selects the URLs which have the highest relevance with the Chinese words as the URL recommendation candidates. The second baseline (i.e., PCLSA) is straightforwardly implemented by the topics generated by PCLSA and Equation (27). The third baseline CL-QLTM(CE) is a variant of CL-QLTM by simply considering



(a) Cross-lingual Query Recommendation



(b) Cross-lingual URL Recommendation



(c) Cross-lingual Information Retrieval

Fig. 8. Application evaluation.

English and Chinese words as being from the same language. The PLTM and JointLDA are not chosen for this task since they are trained on parallel/comparable corpus, the URLs in which are significantly limited compared to the query log. Similar to the evaluation of the cross-lingual query recommendation, we evaluate the performance of URL recommendation by using the 6-point scale (0, 0.2, 0.4, 0.6, 0.8, and 1) and the ground truth generated from five human experts. The average HR is presented in Figure 8(b). We observe that the HR of all the algorithms decreases as N increases, and the CL-QLTM always significantly outperforms more than the three competitors. Moreover, CL-QLTM(CE) is better than Translation Baseline and PCLSA for most of the time, and Translation Baseline is instable since its HR sharply decreases when $N \leq 3$. The experimental results demonstrate that CL-QLTM is effective in capturing the underlying semantic relations between query words and URLs in different languages. Hence, the CL-QLTM is superior in the task of cross-lingual URL recommendation.

6.5.3. Cross-Lingual Information Retrieval. Another application we conduct is the cross-lingual information retrieval, which deals with retrieval of documents that are written in a language different from the language of the query [Vulić et al. 2015]. We conduct this experiment based on a corpus containing 100,000 English and Chinese web pages. We randomly select 1,000 English search queries and 1,000 Chinese search queries. The cross-lingual information retrieval methods follow the setting in Vulić et al. [2015]. We compare the CL-QLTM with the following six baselines: Translation Baseline, which simply translates the source query to the target query; PCLSA, which is implemented by the topics generated by PCLSA and Equation (26); and the PLTM, JointLDA, MuTo, and CL-QLTM(CE), which are variants of the CL-QLTM by simply considering English and Chinese words as being from the same language. The results are evaluated in terms of precision and recall, which are well-adopted metrics for information retrieval. The experimental result is shown in Figure 8(c). We observe that CL-QLTM performs the best among all methods. Furthermore, the performance of CL-QLTM(CE) and PCLSA are quite similar and better than the other three baselines. JointLDA and the PLTM are among the worst ones. Particularly, when the recall becomes larger, the precision of Translation Baseline sharply decreases. Hence, it is harder for Translation Baseline to keep the balance between its precision and recall. The major disadvantage of the PLTM, JointLDA, and MuTo lies in their inapplicability on a query log, and the cross-lingual topics derived from parallel/comparable corpus are not as effective as those learned from a query log. Hence, although the PLTM, JointLDA, and MuTo are able to derive plausible topics from parallel/comparable corpus, these topics are not necessarily effective for web search scenarios. In contrast, the CL-QLTM, which is highly calibered for web search and, thus, its topics are more suitable for web information retrieval tasks. Another interesting observation is gained from comparing the CL-QLTM and CL-QLTM(CE). The performance superiority of the CL-QLTM over CL-QLTM(CE) is caused by differentiating languages and delicately modeling their subtle relations. The above result demonstrates that the CL-QLTM is promising for cross-lingual information retrieval.

7. CONCLUSION

In this article, we study the problem of cross-lingual topic extraction from multilingual search engine query log. We propose a novel probabilistic topic model (i.e., the CL-QLTM) that can incorporate translation knowledge in cross-lingual dictionaries as a regularizer to constrain the parameter estimation so that the topics would be synchronized in multiple languages. We evaluated the model using a real-life query log from a major commercial search engine. The experimental results show that

CL-QLTM is effective to cross the language barrier and is superior in extracting latent topics from a multilingual query log. CL-QLTM outperforms several strong baselines with regards to both quantitative metrics and downstream applications. This work opens up some interesting research directions to explore in the future. For example, the topics of CL-QLTM can be considered as clusters of semantically related words in different languages. Hence, the semantical similarity between words in different languages can be straightforwardly calculated through these topics. Such semantical similarity may be promising for translating the new slangs (which is out of the vocabulary of translative dictionary) on the Web. Hence, we plan to investigate how to apply the derived cross-lingual topics in SMT in future work.

REFERENCES

Vamshi Ambati and U. Rohini. 2006. Using monolingual clickthrough data to build cross-lingual search systems. *New Directions in Multilingual Information Access* (2006), 28.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 113–120, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003), 993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 75–82, 2009.

M. J. Carman, F. Crestani, M. Harvey, and M. Baillie. 2010. Towards query log based personalization using topic models. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*. ACM, 1849–1852, 2010.

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 536–544, 2012.

Kosuke Fukumasu, Koji Eguchi, and Eric P. Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In *Advances in Neural Information Processing Systems*. 1286–1294, 2012.

Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 463–470, 2007.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101, suppl 1 (2004), 5228–5235.

Carrie Grimes, Diane Tang, and Daniel M. Russell. 2007. Query logs alone are not enough. In *Proceedings of the Workshop on Query Log Analysis at the 16th International Conference on World Wide Web*. ACM, 2007.

Tom Hebert and Richard Leahy. 1989. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Transactions on Medical Imaging* 8, 2 (1989), 194–202.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42, 1–2 (2001), 177–196.

J. Huang and E. N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 77–86, 2009.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*. Springer, 444–456, 2010.

Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. 2014. Fast topic discovery from web search streams. In *Proceedings of the 23rd International World Wide Web Conference*. ACM, 949–960, 2014.

Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. 2016. Query intent mining with multiple dimensions of web search data. *World Wide Web* 19, 3 (2016), 475–497.

Di Jiang, Kenneth Wai-Ting Leung, Lingxiao Yang, and Wilfred Ng. 2015a. TEII: Topic enhanced inverted index for top-k document retrieval. *Knowledge-Based Systems* 89 (2015), 346–358.

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. 2012. G-WSTD: A framework for geographic web search topic discovery. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management*. ACM, 1143–1152, 2012.

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. 2013. Panorama: A semantic-aware application search framework. In *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 371–382, 2013.

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. 2014. Personalized query suggestion with diversity awareness. In *Proceedings of the 30th International Conference on Data Engineering*. IEEE, 400–411, 2014.

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, Lingxiao Yang, and Wilfred Ng. 2015. SG-WSTD: A framework for scalable geographic web search topic discovery. *Knowledge-Based Systems* 84 (2015), 18–33.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 175–182, 2002.

K. W.-T. Leung, Dik Lun Lee, and Wang-Chien Lee. 2010. Personalized web search with location preferences. In *Proceedings of the 26th International Conference on Data Engineering*. IEEE, 701–712, 2010.

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: Volume 2*. Association for Computational Linguistics, 880–889, 2009.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. 2009. Distributed algorithms for topic models. *The Journal of Machine Learning Research* 10 (2009), 1801–1828.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 1155–1156, 2009.

Zhaochun Ren and Maarten de Rijke. 2015. Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 93–102, 2015.

Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. 2013. Personalized time-aware tweets summarization. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 513–522, 2013.

Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn Van Dolen, and Maarten De Rijke. 2014. Hierarchical multi-label classification of social text streams. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 213–222, 2014.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 487–494, 2004.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. ACL, 479–484, 2011.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management* 51, 1 (2015), 111–147.

Ivan Vulić and Marie-Francine Moens. 2013. A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models. In *Advances in Information Retrieval*. Springer, 98–109, 2013.

Ivan Vulic and Marie-Francine Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. ACL, 349–362, 2014.

Xuerui Wang, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Bo Pang. 2008. Cross-lingual query classification: A preliminary study. In *Proceedings of the 2nd ACM workshop on Improving Non English Web Searching*. ACM, 101–104, 2008.

X. Wang and A. McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 424–433, 2006.

Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 247–256, 2011.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 879–888, 2012.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1128–1137, 2010.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17*. ACL.

Received October 2015; revised June 2016; accepted June 2016