Knowledge-Based Systems 84 (2015) 18–33

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# SG-WSTD: A framework for scalable geographic web search topic discovery

Di Jiang [a], Jan Vosecky [b], Kenneth Wai-Ting Leung [b], Lingxiao Yang [c,*], Wilfred Ng [b]

[a] Baidu Inc, China
[b] Hong Kong University of Science and Technology, Hong Kong
[c] London School of Economics and Political Science, UK

ARTICLE INFO

*Article history:*
Received 25 October 2014
Received in revised form 14 March 2015
Accepted 21 March 2015
Available online 30 March 2015

*Keywords:*
Topic model
Search engine
Information retrieval

ABSTRACT

Search engine query logs are recognized as an important information source that contains millions of users' web search needs. Discovering *Geographic Web Search Topics* (G-WSTs) from a query log can support a variety of downstream web applications such as finding commonality between locations and profiling search engine users. However, the task of discovering G-WSTs is nontrivial, not only because of the diversity of the information in web search but also due to the sheer size of query log. In this paper, we propose a new framework, *Scalable Geographic Web Search Topic Discovery* (SG-WSTD), which contains highly scalable functionalities such as search session derivation, geographic information extraction and geographic web search topic discovery to discover G-WSTs from query log. Within SG-WSTD, two probabilistic topic models are proposed to discover G-WSTs from two complementary perspectives. The first one is the *Discrete Search Topic Model* (DSTM), which discovers G-WSTs that capture the commonalities between discrete locations. The second is the *Regional Search Topic Model* (RSTM), which focuses on a specific geographic region on the map and discovers G-WSTs that demonstrate geographic locality. Since query log is typically voluminous, we implement the functionalities in SG-WSTD based on the MapReduce paradigm to solve the efficiency bottleneck. We evaluate SG-WSTD against several strong baselines on a real-life query log from AOL. The proposed framework demonstrates significantly improved data interpretability, better prediction performance, higher topic distinctiveness and superior scalability in the experimentation.



## 1. Introduction

Web search reflects the user's information needs in their daily life. A search engine query log, which records millions of users' web search history, becomes a valuable information source and serves as the basis of many applications of search engines [1–4]. Recently, the technique of topic modeling has been successfully utilized to empower many geographic applications in platforms such as Flickr and Twitter [5–7]. Since query log has its own advantages over the other Web 2.0 media in capturing the users' interests and preferences [8,9], it is valuable to investigate how to discover geographic web search topics (G-WSTs) from query log and explore the potential downstream applications of the discovered G-WSTs.

Query log brings several unique challenges that render the traditional topic models inapplicable, or they can only work suboptimally. First, the geographic information is not always explicitly available in a query log. For example, the AOL query log (see Table 1) does not record any geographic information explicitly. This feature of query log is quite different from web services with geo-tagging functionality, where the text snippets are explicitly associated with GPS locations. Therefore, the geographic information needs to be discovered somehow from raw log entries. This challenge also illustrates the fundamental difference between our work and those that study Location-based Services (LBS) [10], which utilize the users' real-time locations to carry out geographic-aware services. We aim to capture the search topics that are related to some locations in which the users are interested, and these locations are not necessarily the same as the users' real-time locations. Second, each search query is very short and on average contains only 2.4–2.7 terms [11]. Therefore, simply considering each query as a document and applying traditional topic models results in poor topic estimation due to the shortness of each query [12]. Third, different from the typical scenario of topic modeling which only faces homogeneous items (e.g., the *words* in

* Corresponding author.
*E-mail address:* audreyyoung@126.com (L. Yang).

http://dx.doi.org/10.1016/j.knosys.2015.03.020
0950-7051/

**Table 1**
Examples of AOL query log.

| Query ID | User ID | Query | Clicked URL | Rank of URL |
|---|---|---|---|---|
| $q_1$ | $u_1$ | Disneyland orlando | disneyworld.disney.go.com | 1 |
| $q_2$ | $u_1$ | Disney resort tokyo | www.swandolphin.com | 3 |
| $q_3$ | $u_1$ | Paris disney | www.disneylandparis.com | 1 |
| $q_4$ | $u_1$ | Glasgow scottish culture | www.bbc.com | 2 |
| $q_5$ | $u_1$ | Edinburgh bagpipe | | |
| $q_6$ | $u_1$ | Glasgow haggis | www.yelp.co.uk | 1 |

documents), query log is basically composed of two heterogeneous items, the *query terms* and the *URLs*, which have different natures and are not independent of each other. Thus, topic modeling on query log needs to handle the heterogeneous items and capture the complicated co-occurrence between them. Fourth, query log is typically voluminous and how to efficiently apply topic modeling to massive query log is still an open problem.

In this paper, we propose a new framework named *Scalable Geographic Web Search Topic Discovery* (SG-WSTD), which addresses the aforementioned challenges and conducts geographic web search topic discovery from two complimentary perspectives: the *discrete perspective* and the *regional perspective*. The discrete perspective assumes that each G-WST is associated with some discrete locations. The geographic relations between the locations, such as the distances between them, are not considered in the discrete perspective. In contrast, the regional perspective assumes that G-WSTs have geographic locality, and each G-WST is adherent to a region on the map. We now illustrate the two perspectives via the example shown in Table 1.

Consider the search queries in Table 1, the queries $q_1, q_2$ and $q_3$ are about the G-WST of Disneyland. Just like the reality that Disneyland resorts are located in different countries, the locations related to this G-WST, such as Orlando, Tokyo and Paris, are remote from each other. In contrast, the queries $q_4, q_5$ and $q_6$ are about the G-WST of Scottish culture, the locations related to this G-WST, Glasgow and Edinburgh, are in geographic proximity.

If we view the queries from the discrete perspective, the G-WST about Disneyland is related to Orlando, Tokyo and Paris while the topic about Scottish culture is associated with Glasgow and Edinburgh. The latent semantics that the G-WST about Scotland culture characterizes the features of a region (i.e., the Central Belt[1]) is lost. Even worse, since we do not impose proximity on the geographic information of a G-WST, the probability that q4, q5 and q6 are assigned as the same G-WST can be very low, and the G-WST about Scottish culture may not be discovered at all.

If we view the queries from the regional perspective, the G-WST about Scottish culture covers a region such as the Central Belt. However, the G-WST about Disneyland covers a wide region, which covers America, Europe and Asia. It is not very informative because the region is overly broad. Even worse, since the regional perspective assumes that each G-WST has the geographic locality, a G-WST that is associated with widely dispersed locations can have very low probability, and the G-WST about Disneyland may not be discovered at all.

The aforementioned example shows that the G-WSTs generated from the two geographic perspectives are complimentary and a framework which supports discovering G-WSTs from both perspectives is able to provide comprehensive coverage for the latent semantics in query log. Thus, in SG-WSTD, we utilize two probabilistic topic models to discover G-WSTs from these two perspectives. The first model is the *Discrete Search Topic Model* (DSTM), which differentiates the locations from the general query terms and assumes that the locations follow a different multinomial distribution given the G-WST. In DSTM, each geographic location is considered as a *toponym* and we aim to find the commonality across different locations. We further propose the *Regional Search Topic Model* (RSTM), which discovers G-WSTs that are adherent to a specific region on the map. The region's spectrums are modeled by two Gaussian distributions over the latitude and the longitude. In RSTM, each geographic location is considered as a *coordinate pair* that contains latitude and longitude. As long as G-WSTs are successfully generated from the two perspectives, they can be utilized together and support a series of new web applications. Besides supporting two geographic perspectives, another salient advantage of SG-WSTD lies in its scalability in processing massive query log. In commercial search engines, the MapReduce computing paradigm has been widely adapted to process the massive search engine query log [13]. In SG-WSTD, all the functionalities such as search session derivation, geographic information extraction and geographic web search topic discovery are developed based on MapReduce and can be seamlessly integrated in parallel computing platforms such as Hadoop [14]. The experimental results show that DSTM and RSTM are effective in discovering semantically coherent G-WSTs. We also evaluate SG-WSTD against several strong baselines with respect to quantitative metrics. The SG-WSTD framework demonstrates significantly improved data interpretability, better prediction performance, higher topic distinctiveness and superior scalability in the experimentation.

The contributions of this paper are summarized as follows:

- We study a new problem in geographic web search topic discovery. Techniques such as search session derivation, geographic information extraction and topic discovery are seamlessly integrated within a framework.
- We address the challenging issues of topic modeling on a search engine query log and propose two complimentary topic models to discover G-WSTs from both the discrete and regional perspectives.
- We propose parallel algorithms for the functionalities in SG-WSTD based on the MapReduce paradigm. These algorithms can be seamlessly integrated into parallel processing platforms such as Hadoop and demonstrate good efficiency in query log analysis.
- We conduct extensive experiments on a real-life query log from AOL. The effectiveness of the proposed framework is unequivocally verified by its superiority over several strong baselines.

The rest of the paper is organized as follows: we review the related work in Section 2. In Section 3, we outline the architecture of SG-WSTD. In Section 4, we discuss the generative processes and the parameter inference strategies for DSTM and RSTM. In Section 5, we present the experimental results. Finally, we conclude the paper in Section 6.

## 2. Related work

In this section, we discuss the related work, including topic modeling, geographic-related knowledge discovery, geographical information retrieval and parallel text processing.

*Topic modeling:* Probabilistic topic modeling is gaining significant momentum in recent years. Blei et al. [15] proposed the pioneering Latent Dirichlet Allocation (LDA) to analyze electronic archives by deriving latent topics. Following LDA, many topic models that specialize in different tasks are proposed. Wang et al. [16]

[1] http://en.wikipedia.org/wiki/Central_Belt.