Mining Web Search Topics With Diverse Spatiotemporal Patterns

Di Jiang, Wilfred Ng
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong, China
{dijiang, wilfred}@cse.ust.hk

ABSTRACT

Mining the latent topics from web search data and capturing their spatiotemporal patterns have many applications in information retrieval. As web search is heavily influenced by the spatial and temporal factors, the latent topics usually demonstrate a variety of spatiotemporal patterns. In the face of the diversity of these patterns, existing models are increasingly ineffective, since they capture only one dimension of the spatiotemporal patterns (either the spatial or temporal dimension) or simply assume that there exists only one kind of spatiotemporal patterns. Such oversimplification risks distorting the latent data structure and hindering the downstream usage of the discovered topics. In this paper, we introduce the *Spatiotemporal Search Topic Model* (SSTM) to discover the latent topics from web search data with capturing their diverse spatiotemporal patterns simultaneously. The SSTM can flexibly support diverse spatiotemporal patterns and seamlessly integrate the unique features in web search such as query words, URLs, timestamps and search sessions. The SSTM is demonstrated as an effective exploratory tool for large-scale web search data and it performs superiorly in quantitative comparisons to several state-of-the-art topic models.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Experimentation

Keywords

Spatiotemporal, Search Engine, Query Log

1. INTRODUCTION

With the rapid growth of web search data, there is a great demand for developing effective text mining models to analyze search engine query log, since discovering the web search topics and capturing their spatiotemporal patterns are vital for applications such as Google Trends[1] and Foursquare [11]. However, little work has been done on analyzing query log from the spatiotemporal perspective. Although some admixture topic models [2] have been proposed to accommodate the demands of analyzing timestamped or GPS-labeled data, to the best of our knowledge, none of them supports discovering topics by considering different spatiotemporal patterns simultaneously. However, in the field of query log analysis, it is critical to model the existence of diverse spatiotemporal patterns. For instance, a topic about *Terrorism* may widely exist from the spatial perspective and lasts for a long time period from the temporal perspective. In contrast, a topic about *Storm Sandy* may be primarily related to the coastal regions of the United States and has a relatively short longevity. Besides the existence of diverse spatiotemporal patterns, query log data have some unique features such as query words, URLs, timestamps, search sessions, etc. These information should be seamlessly integrated for discovering the web search topics from query log.

To handle the aforementioned challenges, we propose the *Spatiotemporal Search Topic Model* (SSTM) to discover the latent topics from query log and capture their diverse spatiotemporal patterns simultaneously. To the best of our knowledge, the SSTM is the first model that accommodates a variety of spatiotemporal patterns in a unified fashion. We carry out large-scale experiments and the experimental results show that the SSTM outperforms several state-of-the-arts in terms of perplexity as well as in the tasks such as location prediction and time predication. The remainder of the paper is organized as follows. In Section 2, the related work is reviewed. Section 3 discusses the spatiotemporal patterns. Section 4 presents the SSTM and details its parameter inference method. Section 5 presents experimental results, and conclusions are provided in Section 6.

2. RELATED WORK

We review some recent related work in this section. In [1], a framework for quantifying the spatial variation of search queries is developed on complete Yahoo query log. The Topic-Over-Time model [14] is proposed to associate each topic with a continuous distribution over timestamps. In [4], the authors proposed a model for the monitoring topic and vocabulary evolution over documents. In [16], LPTA is proposed to exploit the periodicity of the terms as well as term co-occurrences. Wang *et al.* [13] proposed a topic model to capture the relationships between locations and words in news and blogs. Yin *et al.* [15] studied the problem of discovering geographic topics from GPS-associated tweets. Sizov *et al.* [12] proposed models for characterization of social media by combining text features with spatial knowledge. Eisenstein *et*

[1]http://www.google.com/trends/

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

al. [3] proposed a multi-level generative model that reasons about lexical variation across different geographic regions. Hao *et al.* [5] proposed the location-topic model to mine location-representative knowledge from a collection of travelogues. Hong *et al.* [6] presented an algorithm by modeling diversity in tweets based on topical diversity, geographical diversity, and an interest distribution of the user. In [10], a probabilistic approach is proposed to model the subtopic themes and spatiotemporal theme patterns simultaneously. With the popularity of applying topic modeling to spatial, temporal or even spatiotemporal information, few prior work has been done on analyzing web search data by jointly modeling diverse spatiotemporal patterns. Hence, the major distinction of the proposed SSTM is its capability of discovering web search topics while capturing their diverse spatiotemporal patterns simultaneously.

3. SPATIOTEMPORAL PATTERNS

In this section, we review the spatial temporal patterns that are used in existing work. The existing spatial patterns can be broadly divided into two categories: the *local* pattern (S_l) and the *global* pattern (S_g) [12, 13]. The logic behind this categorization is that some topics demonstrate geographic locality while the others do not. Specifically, S_g assumes that each topic is related to some locations and the geographic distance between the locations is not considered while S_l assumes that topics have geographic locality and each topic is related to a specific area on the map. The existing temporal patterns can be broadly classified into three types: the *periodic* pattern (T_p), the *background* pattern (T_g) and the *bursty* pattern (T_b) [8, 16]. A periodic topic is one that repeats in regular intervals; a background topic is one covered uniformly over the entire period; a bursty topic is a transient topic that is intensively covered only in a certain time period. We now present the spatiotemporal patterns that will be used in this paper. Assume that we have the spatial pattern set $\mathcal{S}$ and the temporal pattern set $\mathcal{T}$ from the existing work, we create a set of spatiotemporal patterns $\mathcal{P}$ by applying the Cartesian product on $\mathcal{S}$ and $\mathcal{T}$. In this way, we get very diverse spatiotemporal patterns, which are presented in Table 1. Note that the spatiotemporal pattern proposed in [10] is captured by p_2 in Table 1.

Table 1: Spatiotemporal Patterns

ID	Pattern	Description
p_1	(S_g, T_g)	global-background pattern
p_2	(S_g, T_b)	global-bursty pattern
p_3	(S_g, T_p)	global-periodic pattern
p_4	(S_l, T_g)	local-background pattern
p_5	(S_l, T_b)	local-bursty pattern
p_6	(S_l, T_p)	local-periodic pattern

4. SPATIOTEMPORAL SEARCH TOPIC MODEL

4.1 Description of SSTM

We represent each query log entry in a format as follows:

$$\{uid, \mathbf{w}, t, (\mathbf{l}, \mathbf{l_{lat}}, \mathbf{l_{lon}})^?, \mathbf{u}^?\}$$

Here uid is the user identifier, $\mathbf{w}$ is the word vector for the query, t is the timestamp of the query, $(\mathbf{l}, \mathbf{l_{lat}}, \mathbf{l_{lon}})^?$ is a vector of triplets where $\mathbf{l}$ is the names of the locations, $\mathbf{l_{lat}}$ and $\mathbf{l_{lon}}$ represent the latitudes and longitudes of the corresponding locations, $\mathbf{u}^?$ is the clicked URLs of this query. The location information is not readily available in query log, we employ the method in [9] to extract the locations $\mathbf{l}$ and obtain the latitudes $\mathbf{l_{lat}}$ and longitudes $\mathbf{l_{lon}}$ of the locations via a geographic dictionary. The question mark indicates that the corresponding element may not exist for some log entries. In web search, the queries are not independent from each other. The users usually submit search queries consecutively as a session to satisfy a single information need. We utilize the method in [7] to segment query log into search sessions. Finally, we group each user's log entries together as a document and then organize each document via sessions.

The generative story of SSTM is presented in Algorithm 1. We assume that each user has a topic distribution and each topic is related to a specific spatiotemporal pattern. When conducing web search to satisfy an information need, the user first decides the topic and then selects some query words according to the chosen topic. For each search session, the user needs to decide whether to click some URLs. If so, the clicked URLs are generated according to the chosen topic as well. Since the information within a session is coherent and is used to satisfy the same information need, we constrain that the information in the same session shares the same topic. Finally, the spatiotemporal information such as the timestamps and the locations (or the latitudes and longitudes) is generated based on the spatiotemporal pattern of the chosen topic.

Algorithm 1 Generative Procedure of SSTM

for topic $k \in 1, ..., K$ **do**
  draw a query word distribution $\phi_k \sim$ Dirichlet(β);
  draw a URL distribution $\Omega_k \sim$ Dirichlet(δ);
end for
for each document $d \in 1, ..., D$ **do**
  draw d's topic distribution $\theta_d \sim$ Dirichlet(α);
  for each search session s in d **do**
    choose a topic $z \sim$ Multinomial(θ_d);
    generate query words $w \sim$ Multinomial(ϕ_z);
    if $X_s = 1$ **then**
      generate URLs $u \sim$ Multinomial(Ω_z);
    end if
    generate the temporal information $t \sim p(t|z)$;
    if $Y_s = 1$ **then**
      generate the spatial information $l \sim p(l|z)$;
    end if
  end for
end for

4.2 Parameter Inference

We proceed to discuss a sampling method for the parameter inference of SSTM. The joint likelihood of the observed query words, the URLs and the spatiotemporal information with the hyperparameters is listed as follows:

$$P(\mathbf{w}, \mathbf{u}, \mathbf{t}, \mathbf{l}, \mathbf{z}|\alpha, \beta, \delta, \mathbf{X}, \mathbf{Y}) = P(\mathbf{u}|\mathbf{z}, \delta, \mathbf{X})P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{l}|\mathbf{z}, \mathbf{Y})P(\mathbf{t}|\mathbf{z})P(\mathbf{z}|\alpha). \quad (1)$$

The probability of generating the query words $\mathbf{w}$ in the corpus is given as follows:

$$P(\mathbf{w}|\mathbf{z}, \beta) = \int \prod_{d=1}^{D} \prod_{s=1}^{S_d} \prod_{i=1}^{W_s} P(w_{dsi}|\phi_{z_{ds}})^{N_{dst_{dsi}}} \prod_{z=1}^{K} P(\phi_z|\beta) d\Phi. \quad (2)$$

The probability of generating the URLs $\mathbf{u}$ in the corpus is given as

follows:

$$P(\mathbf{u}|\mathbf{z},\delta,\mathbf{X}) = \int \prod_{d=1}^{D}\prod_{s=1}^{S_d}\{\prod_{i=1}^{U_{ds}} P(u_{dsi}|\Omega_{z_{ds}})^{N_{dsu_{dsi}}}\}^{I(X_{ds}=1)} \prod_{z=1}^{K} P(\Omega_z|\delta)d\Omega. \tag{3}$$

The conditional probability of generating the temporal information t given the topic z can be written as:

$$\begin{aligned} p(t|z) = p(I_z^T = 0)p'(t|z) + \\ p(I_z^T = 1)p''(t|z) + p(I_z^T = 2)p'''(t|z), \end{aligned} \tag{4}$$

where

$$\begin{cases} p'(t|z) = \dfrac{1}{t_e - t_s}, & (5) \\ p''(t|z) = \dfrac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z)^2}{\sigma_z^2}}, & (6) \\ p'''(t|z) = \sum\limits_{n} p(t|z,n)p(n). & (7) \end{cases}$$

The background pattern $p'(t|z)$ is modeled by a uniform distribution. In $p'(t|z)$, t_e and t_s are the newest and the oldest timestamps in the query log. The bursty pattern $p''(t|z)$ is modeled by a Gaussian distribution. The periodic pattern $p'''(t|z)$ is modeled as a mixture of Gaussian distributions. In $p'''(t|z)$, 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.

The conditional probability of generating the spatial information l given the topic z can be written as:

$$p(l|z) = p(I_z^L = 0)p'(l|z) + p(I_z^L = 1)p''(l|z). \tag{8}$$

If topic z is a global pattern, we model that spatial information by a Multinomial distribution over the locations. If topic z is a local pattern, we model the spatial information as a 2-dimensional Gaussian distribution over the latitude and longitude:

$$\begin{aligned} p''(l|z) = \frac{1}{2\pi\sigma_z^{lat}\sigma_z^{lon}\sqrt{1-r^2}} e^{\frac{1}{2(1-r^2)}[\frac{(l^{lat}-\mu_z^{lat})^2}{\sigma_z^{lat\,2}}} \\ -2r\frac{(l^{lat}-\mu_z^{lat})(l^{lon}-\mu_z^{lon})}{\sigma_z^{lat}\sigma_z^{lon}} + \frac{(l^{lon}-\mu_z^{lon})^2}{\sigma_z^{lon\,2}}]. \end{aligned} \tag{9}$$

After combining the aforementioned formula terms, applying Bayes rule and folding terms into the proportionality constant, the conditional probability of assigning the kth topic for the ith session can be determined by a set of formulas. For instance, if the topic k is related to the spatiotemporal pattern p_1 in Table 1, the conditional probability is defined as follows:

$$\begin{aligned} P(z_i = k, I_k^T = 0, I_k^L = 0|\mathbf{z_{-i}}, \mathbf{w}, \mathbf{u}, \mathbf{t}, \mathbf{l}, \mathbf{X}, \mathbf{Y}) \propto \\ \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^{K} C_{dk'}^{DK} + \alpha_{k'}} \prod_{j=1}^{T} \frac{1}{t_{end} - t_{start}} \\ \frac{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w + N_{iw}))} \prod_{w=1}^{W_i} \frac{\Gamma(C_{kw}^{KW} + \beta_w + N_{iw})}{\Gamma(C_{kw}^{KW} + \beta_w)} \\ \{\frac{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u))}{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u + N_{iu}))} \prod_{u=1}^{U_i} \frac{\Gamma(C_{ku}^{KU} + \delta_u + N_{iu})}{\Gamma(C_{ku}^{KU} + \delta_u)}\}^{I(X_i=1)} \\ \{\frac{\Gamma(\sum_{l=1}^{L}(C_{kl}^{KL} + \lambda_l))}{\Gamma(\sum_{l=1}^{L}(C_{kl}^{KL} + \lambda_l + N_{il}))} \prod_{l=1}^{L_i} \frac{\Gamma(C_{kl}^{KL} + \lambda_l + N_{il})}{\Gamma(C_{kl}^{KL} + \lambda_l)}\}^{I(Y_i=1)}, \end{aligned} \tag{10}$$

where C_{dk}^{DK} is the number of sessions that have been assigned to topic k in document d, C_{kw}^{KW} is the number of query words that have been assigned to topic k, C_{ku}^{KU} is the number of URLs that have been assigned to topic k, N_{iw} is the number of w in the ith session and N_{iu} is the number of u in the ith session. Similarly, we can derive the remaining conditional probabilities when the corresponding topic is related to other spatiotemporal patterns such as p_2, p_3, ..., p_6. For the sake of simplicity and efficiency, we update the distribution parameters of the bursty pattern, the periodic pattern and the local pattern after each iteration of the sampling procedure. For example, we update the Gaussian distribution of the bursty pattern by the sample mean and sample variance as follows:

$$\mu_z = \frac{1}{n}\sum_{i=1}^{n} t_i, \tag{11}$$

$$\sigma_z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \mu_z)^2}, \tag{12}$$

where t_i is the ith timestamp which exists in sessions that are assigned search topic z. The update formulas for other spatiotemporal patterns can be straightforwardly obtained in a similar way.

5. EXPERIMENTS

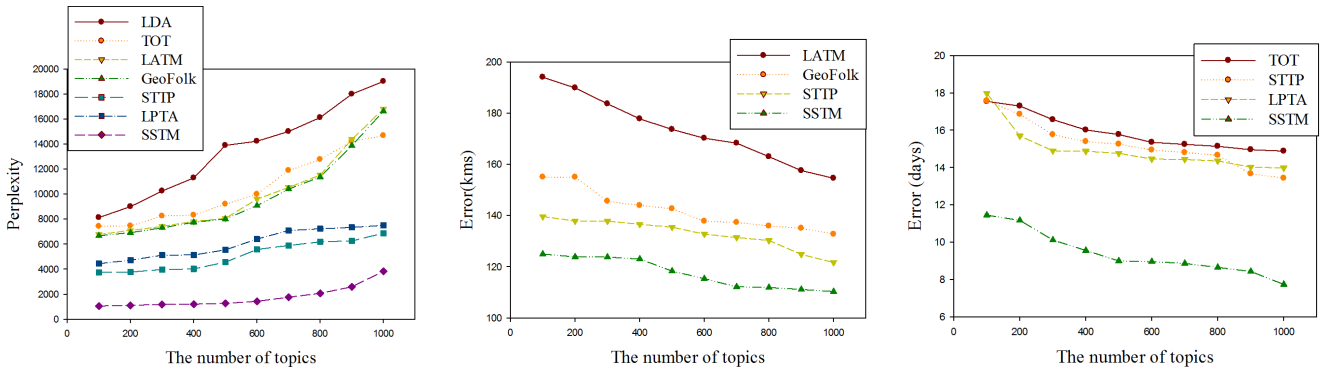
In this section, we evaluate the SSTM on the query log of a major commercial search engine in the United States. The dataset contains 1,200,945 search queries that were submitted by 10,213 users. After carrying out the session derivation process, we obtain 520,131 search sessions.

5.1 Perplexity Comparison

We first evaluate SSTM by the quantitative metric of perplexity. The baselines we choose are summarized as follows: Latent Dirichlet Allocation (LDA) [2], Topics-Over-Time model (TOT)[14], Location Aware Topic Model (LATM)[13], Geodata in Folksonomies (GeoFolk)[12], Spatiotemporal Theme Pattern model (STTP) [10] and Latent Periodic Topic Analysis (LPTA)[16]. Perplexity measures the ability of a model to generalize to unseen data and better generalization performance is indicated by a lower perplexity. We compare the models by a ten-fold cross validation. Figure 1(a) illustrates the average perplexity for each model when the number of topics is set to different values. We can see that the STTM provides significantly better fit for the data than the baselines. For instance, when the number of topics is 600, the perplexity of LDA, TOT, LATM, GeoFolk, STTP and LPTA are 14220, 9986, 9601, 9072, 5569 and 6401 while the perplexity of SSTM is 1423.

5.2 Location Prediction

Geographical locations can be used to predict users' behaviors and uncover users' interests and therefore it is potentially invaluable for many perspectives, such as behavior targeting and online advertisements. In this subsection, we focus on the task of location prediction. Our goal is to predict the location for a new search session based on the words and URLs in the session and the user's information. In this experiment, we filter out the original geographic information from each session and the original geographic information is utilized as the ground truth. For each new session, we predict its location as $\hat{l}_d$. We calculate the Euclidean distance between the predicted value and the ground truth locations and average them over the whole test set. For all the models, we adopt a ten-fold cross validation setting and the numbers reported here are averaged across different folds. The experimental result is shown in Figure 1(b). We can see that LATM and GeoFolk demonstrate



(a) Perplexity Comparison (b) Location Prediction (c) Time Prediction

Figure 1: Experimental Results

relatively high error because they only captures the spatial information. By considering both the spatial and the temporal information, STTP achieves better performance in location prediction, although it only captures one kind of spatiotemporal pattern. By capturing the diverse spatiotemporal patterns, SSTM further reduces the error of location prediction. For instance, when the number of topics is 1000, the error of LATM, GeoFolk and STTP are 154km, 132km and 121km while the error of SSTM is 110km.

5.3 Time Prediction

Another interesting feature of SSTM is the capability of predicting the timestamps given the textual information in a search session. This task also provides another opportunity to quantitatively compare SSTM against the models that capture the temporal information such as TOT, STTP and LPTA. We measure their ability of predicting the date given the query terms in a session. We use 5,000 held-out search sessions as the evaluation data and then evaluate each model's ability to predict the date of a search session. The result is presented in Figure 1(c). SSTM demonstrates the highest date prediction accuracy. For example, when the number of topics is set to 1000. The error of SSTM is 7.73 days while those of TOT, STTP and LPTA are 14.87, 13.26 and 13.97 days.

6. CONCLUSION

Search engine query log usually has a mixture of latent topics, which exhibit diverse spatiotemporal patterns. Discovering the latent topics from query log and capturing their spatiotemporal patterns are critical for many applications in information retrieval. In this paper, we develop the *Spatiotemporal Search Topic Model* (SSTM), which explains the generation of web search topics and a variety of spatiotemporal patterns simultaneously. We evaluate the SSTM against several strong baselines on a commercial search engine query log. Experimental results show that the SSTM provides better fit for the latent structure of query log data. We also demonstrate that SSTM can serve as the basis for higher level tasks, such as predicting the location and time of web search behaviors. In future work, we plan to apply the SSTM for applications such as user profiling and personalized spatiotemporal web search.

7. ACKNOWLEDGEMENTS

This work is partially supported by RGC GRF under grant number HKUST 617610. We also wish to thank the anonymous reviewers for their comments.

8. REFERENCES

[1] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak, *Spatial variation in search engine queries*, WWW, 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, JMLR (2003).

[3] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, *A latent variable model for geographic lexical variation*, EMNLP, 2010.

[4] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou, *Topic evolution in a stream of documents*, 2009.

[5] Q. Hao, R. Cai, C. Wang, R. Xiao, J. M. Yang, Y. Pang, and L. Zhang, *Equip tourists with knowledge mined from travelogues*, WWW, 2010.

[6] L. Hong, A. Ahmed, S. Gurumurthy, A.J. Smola, and K. Tsioutsiouliklis, *Discovering geographical topics in the twitter stream*, WWW, 2012.

[7] Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng, *Context-aware search personalization with concept preference*, CIKM, 2011.

[8] Di Jiang, Kenneth Wai-Ting Leung, Wilfred Ng, and Hao Li, *Beyond click graph: Topic modeling for search engine query log analysis*, DASFAA, 2013.

[9] Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng, *G-wstd: a framework for geographic web search topic discovery*, CIKM, 2012.

[10] Q. Mei, C. Liu, H. Su, and C.X. Zhai, *A probabilistic approach to spatiotemporal theme pattern mining on weblogs*, WWW, 2006.

[11] Blake Shaw, Jon Shea, Siddhartha Sinha, and Andrew Hogue, *Learning to rank for spatiotemporal search*, WSDM, 2013.

[12] S. Sizov, *Geofolk: latent spatial semantics in web 2.0 social media*, WSDM, 2010.

[13] C. Wang, J. Wang, X. Xie, and W. Y. Ma, *Mining geographic knowledge using location aware topic model*, GIR, 2007.

[14] X. Wang and A. McCallum, *Topics over time: a non-markov continuous-time model of topical trends*, SIGKDD, 2006.

[15] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, *Geographical topic discovery and comparison*, WWW, 2011.

[16] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, *Lpta: A probabilistic model for latent periodic topic analysis*, ICDM, 2011.