Mining Web Search Topics With Diverse Spatiotemporal Patterns

ABSTRACT

Mining the topics (or themes) from web search and analyzing their spatiotemporal patterns have many applications in various domains such as searching information by mobile devices. As web search is heavily influenced by many spatial and temporal factors, the 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 the spatial or temporal information or assume there exists only one kind of spatiotemporal pattern. These simplistic assumptions risk heavily distorting the latent structure of web search data and hinder the downstream usage of the discovered topics. In this paper, we introduce a new model, the *Spatiotemporal Search Topic Model* (SSTM) to discover the topics from web search data, taking into consideration their diverse spatiotemporal patterns simultaneously. The SSTM flexibly supports different spatiotemporal assumptions in the process of topic discovery. Besides following spatiotemporal patterns, web search data also has some unique features, such as query words, URLs, timestamps and search sessions. We show that the proposed SSTM can integrate all the aforementioned information together. The SSTM is demonstrated as an effective exploratory tool of a large-scale query log and it performs superiorly in quantitative comparisons to several state-of-the-art models.

1. INTRODUCTION

With the huge size and the rapid growth of the web search data, there is a great demand for developing more effective text mining models to analyze search engine query log, since discovering topics from the log and capturing their spatiotemporal patterns are vital for many applications such as Google Trends[1] and Foursquare [10]. However, very little work has been done on analyzing web search data from the spatiotemporal perspective. On the other hand, more web search is done in a mobile environment as smartphones become a common web search platform. Although some admixture topic models [1] 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 their diverse spatiotemporal patterns simultaneously, which, however, is critical for analysing web search data, since web search topics are usually subjected to many different spatiotemporal patterns. For instance, a topic about *Terrorism* widely exists 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 region of the United States and has a relatively short longevity.

Search engine query log has multiple types of information such as query words, URLs, timestamps, etc. Thus, it is more challenging than mining topics from traditional documents, articles or microblogs. To handle the challenge, we propose the *Spatiotemporal Search Topic Model* (SSTM) to discover the latent topics from web search data and capture their diverse spatiotemporal patterns. To the best of our knowledge, the proposed SSTM is the first model that accommodates a wide range of spatiotemporal assumptions in a unified framework. Besides the variety of spatiotemporal patterns, SSTM also seamlessly integrates many unique features of web search data, such as query words, URLs and search sessions. We carry out large-scale experiments and the results show that SSTM outperforms several state-of-the-art algorithms in terms of perplexity as well as in other interesting tasks in a mobile environment such as location prediction and time predication. The remainder of the paper is organized as follows. In Section 2, related work is reviewed. Section 3 reviews the existing spatial and temporal assumptions. 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

In this section, we review some related work. The *Topic-Over-Time model* [14] was proposed to associate each topic with a continuous distribution over timestamps. [2] proposed "topic monitors" for monitoring topic and vocabulary evolution over documents. In [17], the authors proposed LPTA that exploits the periodicity of the terms as well as term co-occurrences. Wang *et al.* [12] proposed a topic model that explicitly captures the relationships between locations and words in the news and blogs. Yin *et al.* [16] studied the problem of discovering geographic topics from GPS-associated tweets. Sizov *et al.* [11] proposed Bayesian models for characterization of social media by combining text features with spatial knowledge. Eisenstein *et 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* that mines location-representative knowledge from a large collection of travelogues. Hong *et al.* [6] presented an algorithm by modeling diversity in tweets based on topical diversity, geographi-

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

cal diversity, and an interest distribution of the users. [9] explores spacetime structures of the topical content of short textual messages in a stream available from Twitter in Ireland. In [8], a probabilistic approach is proposed to model the subtopic themes and spatiotemporal theme patterns simultaneously. With the popularity of topic modeling on spatial or temporal information, few existing work has been done in the context of web search data. There is even no work studying how to jointly model diverse spatiotemporal patterns. The major distinction of our work is to discover diverse spatiotemporal patterns based on a single model.

3. SPATIOTEMPORAL PATTERNS

We first review what spatial patterns and temporal patterns are typically used in existing work. Spatial patterns can be broadly divided into two categories: the *local* pattern (S_l) [11] and the *global* pattern (S_g) [12]. The rationale behind this categorization is that some topics demonstrate geographic locality but others do not. Informally, S_g assumes that each topic is related to some locations and the geographic distance between these locations is not considered while S_l assumes that topics have geographic locality and each topic is related to an area on the map. 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) [17]. A periodic topic is one repeating 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 significantly extend the horizon of the spatiotemporal patterns as follows. Assume that we have the spatial pattern set S and the temporal pattern set T from the existing work, we create a set of spatiotemporal patterns P by applying the Cartesian product on S and T. In this way, we get very diverse spatiotemporal patterns such as local-background, global-bursty, etc. The spatiotemporal patterns we utilize in this paper is summarized in Table 1. Note that the spatiotemporal pattern proposed in [8] can be captured by p_2.

Table 1: Spatiotemporal Patterns

IDs	Patterns	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 Model Description

We now represent each query log entry 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 name of the location, $\mathbf{l_{lat}}$ and $\mathbf{l_{lon}}$ represent the latitude and longitude of the corresponding location, $\mathbf{u}^?$ is the clicked URL of this query. The location information is not readily available in query log, we employ the method in [13] to extract the locations $\mathbf{l}$ and then utilize Google GeoCoding API[2] to obtain the latitudes $\mathbf{l_{lat}}$ and longitudes $\mathbf{l_{lon}}$ of the locations. Be advertised that the question mark indicates that the element may not exist for some query log entry. 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. We proceed to discuss the generative story of SSTM. We assume that each user has a topic distribution and each topic is related to a 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 an 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 is generated based on the spatiotemporal pattern of the chosen topic. With the above specifications, the generative process of SSTM is presented in Algorithm 1.

Algorithm 1 Generative Process 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 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
    draw timestamps $t \sim p(t|z)$
    if $Y_s = 1$ **then**
      draw locations $l \sim p(l|z)$
    end if
  end for
end for

4.2 Inference Algorithm

Now, we 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:

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

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

[2]https://developers.google.com/maps/documentation/geocoding/

follows:

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

The conditional probability of generating a timestamp t in document d given the topic z can be written as:

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

The background pattern is modeled by a uniform distribution:

$$p'(t|z) = \frac{1}{t_{end} - t_{start}},$$

where t_{end} and t_{start} are the newest and the oldest timestamps in the query log. The bursty pattern is modeled by a Gaussian distribution:

$$p''(t|z) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z)^2}{\sigma_z^2}}.$$

The periodic pattern is modeled as a mixture of Gaussian distributions,

$$p'''(t|z) = \sum_n p(t|z,n)p(n),$$

where 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 location l in document d given the topic z can be written as:

$$p(l|z) = p(I_z^L = 0|d)p'(l|z) + p(I_z^L = 1|d)p''(l|z). \tag{5}$$

If topic z is a global pattern, we model that spatial information by a Multinomial distribution over the locations:

$$p'(l|z) = \frac{C_{zl}^{KL} + \lambda_l}{\sum_{l'}(C_{zl'}^{KL} + \lambda_{l'})}, \tag{6}$$

where C_{zl}^{KL} is the number of locations that have been assigned to topic z and λ_l is the Dirichlet prior of location l. If topic z is a local pattern, we model the spatial information as a 2-dimensional Gaussian distribution over the latitude and longitude information as follows:

$$p''(l|z) = \frac{1}{2\pi\sigma_z^{lat}\sigma_z^{lon}\sqrt{1-r^2}} e^{\frac{1}{2(1-r^2)}} [\frac{(l_{lat}-\mu_z^{lat})^2}{\sigma_z^{lat\,2}} - 2r\frac{(l_{lat}-\mu_z^{lat})(l_{lon}-\mu_z^{lon})}{\sigma_z^{lat}\sigma_z^{lon}} + \frac{(l_{lon}-\mu_z^{lon})^2}{\sigma_z^{lon\,2}}]. \tag{7}$$

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:

$$P(z_i = k, I_k^T = 0, I_k^L = 0|\mathbf{z_{-i}}, \mathbf{w}, \mathbf{u}, \mathbf{t}, \mathbf{l}, \mathbf{X}, \mathbf{Y}) \propto \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^{K} C_{dk'}^{DK} + \alpha_{k'}} \prod_{j=1}^{T} \frac{1}{t_{end} - t_{start}} \frac{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w + N_{iw}))} \prod_{w=1}^{W_i} \frac{\Gamma(C_{kw}^{KW} + \beta_w + N_{iw})}{\Gamma(C_{kw}^{KW} + \beta_w)} \{\frac{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u))}{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u + N_{iu}))} \prod_{u=1}^{U_i} \frac{\Gamma(C_{ku}^{KU} + \delta_u + N_{iu})}{\Gamma(C_{ku}^{KU} + \delta_u)}\}^{I(X_i=1)} \{\frac{\Gamma(\sum_{l=1}^{L}(C_{kl}^{KL} + \lambda_l))}{\Gamma(\sum_{l=1}^{L}(C_{kl}^{KL} + \lambda_l + N_{il}))} \prod_{l=1}^{L_i} \frac{\Gamma(C_{kl}^{KL} + \lambda_l + N_{il})}{\Gamma(C_{kl}^{KL} + \lambda_l)}\}^{I(Y_i=1)}, \tag{8}$$

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 the conditional probability of assigning the kth topic for the ith session when the kth topic is related to spatiotemporal patterns such as p_2,..., p_6. We do not list all the formulas here due to space limitations. We simply 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:

$$\mu_z = \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i, \tag{9}$$

$$\sigma_z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2}, \tag{10}$$

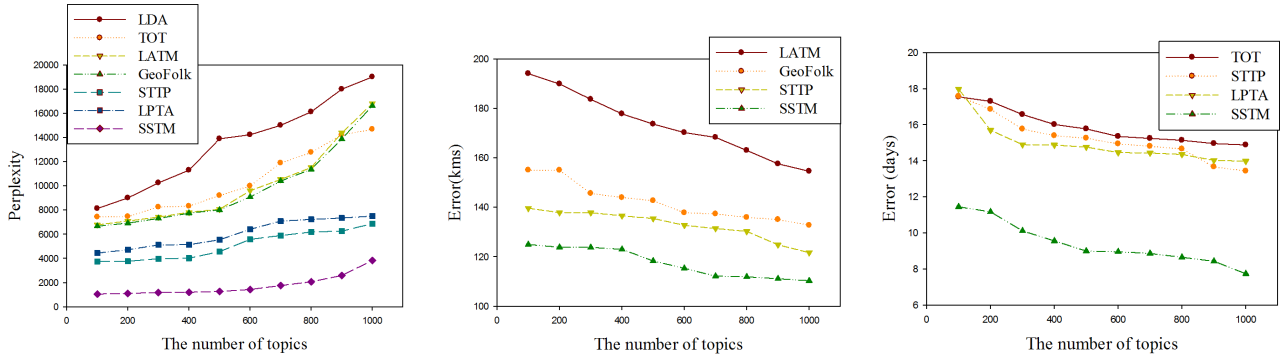
where t_i is the ith timestamp which exists in sessions that are assigned search topic z. The parameter update formulas for other patterns can be straightforwardly obtained by using a similar approach.

5. EXPERIMENTS

In this section, we evaluate SSTM by using the query log of a commercial search engine. 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 from the data set. We adopt symmetric hyperparameters in SSTM like [4] and this setting demonstrates good performance in our experiments.

5.1 Perplexity Comparison

We first evaluate SSTM in the quantitative metric of perplexity. The chosen baselines are summarized as follows: Latent Dirichlet Allocation (LDA) [1], Location Aware Topic Model (LATM)[12], Geodata in Folksonomies (GeoFolk)[11], Latent Periodic Topic Analysis (LPTA)[17], Topics-Over-Time model (TOT)[15], Spatiotemporal Theme Pattern model (STTP) [8] and the LPTA model [17]. Perplexity measures the ability of a model to generalize to unseen data. 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 topic amount is set to different values. We can see that the STTM provides significantly better fit than the baselines. For example, 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.



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

Figure 1: Experimental Results

5.2 Location Prediction

Geographical locations can be used to predict users' behaviors and uncover users' interests and, thus, it is invaluable for many applications, such as behavior targeting and just-in-time advertisement. We now 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 used in the session and the user's information. In this experiment, we filter out the geographic information from each session for location prediction and the original geographic information is utilized as the ground truth. For each new session, we predict its location as $\hat{l}_d$. We compute 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. The values reported here are averaged across different folds. The experimental result is shown in Figure 1(b). We can see that LATM and GeoFolk demonstrate the highest and the second highest error, since they only capture the spatial information. By considering both the spatial information and the temporal information, STTP achieves better performance in location prediction, although it only captures only kind of spatiotemporal pattern. By capturing more 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 words in a session. This task also provides another opportunity to quantitatively compare SSTM against the baselines. We measure the ability to predict 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 with an average error. For example, when the number of topics is set 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 has a mixture of topics and exhibit spatiotemporal patterns. Discovering the latent topics and modeling their spatiotemporal patterns are useful for many applications running in a mobile environment. In this paper, we propose a novel topic model SSTM used in web search. SSTM discovers topics from web log data and captures a variety of spatiotemporal topic patterns simultaneously. We evaluate SSTM against several strong baselines on a commercial search engine query log. Experiment results show that SSTM provides better fit for the latent structure of web search data. We also demonstrate that SSTM serves as a fundamental utility for higher level tasks based on spatiotemporal patterns, such as the prediction of location and time of web search behaviors. Our future work applies SSTM for enhancing applications running on mobile devices, such as smartphone web user profiling and personalized spatiotemporal web search.

7. REFERENCES

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

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

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

[4] T. L. Griffiths and M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences of the United States of America (2004).

[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] J. Huang and E. N. Efthimiadis, *Analyzing and evaluating query reformulation strategies in web search logs*, CIKM, 2009.

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

[9] A. Pozdnoukhov and C. Kaiser, *Space-time dynamics of topics in streaming text*, SIGSPATIAL, 2011.

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

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

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

[13] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.Y. Ma, and Y. Li, *Detecting dominant locations from search queries*, SIGIR, 2005.

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

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

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

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