

Burstiness in Query Log: Web Search Analysis by Combining Global and Local Evidences

Chen Zhang^{†‡}, Sen Zhang[†], Lei Chen[‡], Peiguang Lin[†]

[†]School of Computer Science & Technology, Shandong University of Finance & Economics, Jinan, China

[‡]HongKong University of Science & Technology, HongKong

czhangad@cse.ust.hk, zhangsen.sdufe@gmail.com, leichen@cse.ust.hk, linpg@sdufe.edu.cn

Abstract—Web search analysis plays a critical role in improving the performance of cutting-edge search engines. Most of the existing models, such as the click graph and its variants, focus on utilizing the wisdom of the crowd. However, how to design a model supporting both the collective wisdom as well as the unique characteristic of individuals is rarely studied. In this paper, our goal is to solve the new problem of *user-specific web search analysis*. We go beyond click graph and propose two probabilistic topic models, Topic Independence Model (TIM) and Topic Dependence Model (TDM). TIM adopts an assumption that the generation of query terms and URLs are topically independent; TDM captures the coupling between search queries and URLs. We also capture the temporal burstiness of topics by utilizing the continuous Beta distribution. Through a large-scale analysis of a real-life search query log, we observe that each user's web search trail enjoys multiple kinds of user-based unique characteristics. On a massive search query log, the new models achieve a better held-out likelihood than standard LDA, DCMLDA and TOT, and they can also effectively reveal the latent evolutions of topics on the corpus level and user-based level.

I. INTRODUCTION

The term “burstiness” describes the behavior of a rare word appearing many times in a single document. It has since been discovered that there are many natural and man made quantities that demonstrate such burstiness phenomenon, such as in financial realm, gene expression and computer vision data. Now web search has become an indispensable part of people's daily life, and the search queries that the user submits have become a huge pool of human knowledge, which demonstrate its unique characteristics against natural language used in other digital formats such as articles, microblog, etc. However, with the importance of web search analysis and its clear uniqueness among other natural-language-based text, very few work has been done to analyze the burstiness phenomenon in web search behaviors. In this paper, we systematically analyze three kinds of burstiness phenomenon in web search and proposed different probabilistic topic models to utilize the burstiness phenomenon.

The topic modeling approach plays a significant role in latent knowledge exploration and becomes more and more popular in data mining. In order to find out the latent topics of the document, Blei *et al.*[1] proposed Latent Dirichlet Allocation (LDA). LDA have been widely used for both academia and industry. And many variants of LDA also performed quite so good on its area, such as twitter analysis, digital articles. Jiang *et al.* discovered latent search topics via mining web

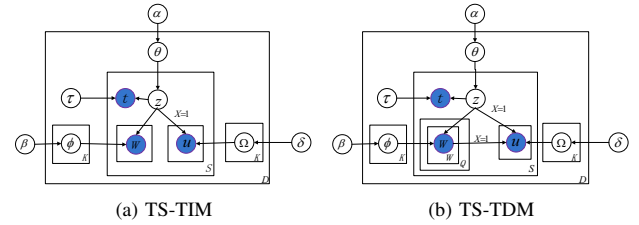


Fig. 1: Two Interpretations of Web Search

search query log[2–4]. But LDA and its variants can't capture the burstiness in an article. Although model likes DCMLDA[5] has been proposed to model the word burstiness in general documents, they are not good candidates to model the web search behaviors. The reason is that the individual web search trail has two different component, one is query term and the other is URLs - this is totally different with natural language.

In this paper, we study the burstiness phenomenon in web search via analyzing search query log. Explicitly, we present probabilistic models that are far better suited for representing search query log and capture multiple domain characteristics. The models we proposed are quite different in natural: the TIM is a one-stage model and the burstiness information is essentially stored in the skewed Dirichlet priors. The second model TDM captures the coupling between queries and URLs. Finally, we propose two variants TIM-T and TDM-T to enable the proposed models to capture the temporal burstiness. After comparing the proposed models with traditional topic models, we draw a conclusion that the proposed models perform more effectively in a huge of real query log.

II. USER-CENTRIC PROBABILISTIC TOPIC MODELS

In this section, we propose a series of topic models to capture the web search burstiness. The models have the following desiderata:

- The burstiness of search query terms and URLs are modeled separately.
- Web search characteristics, query terms, URLs and sessions are all taken into consideration. A search session refers to users who submitted some queries for satisfying that same query need in a time period.

A. Topical Independence Model (TIM)

A search engine user has many different search topics. Therefore, our model must allow a single document have

multiple topics, and account for search burstiness by making the topics document specific. For simplicity, TIM adopts an assumption that the generation of query terms and URLs are topically independent, which is shown in the graphical model in Figure 1(a).

Figure 1(a) presents the generative process of TIM. At first, for each document, we draw a document-specific mix θ_d over topics that is drawn from a symmetric Dirichlet prior α . Then in the document d , the distribution of word ϕ_{kd} and the URL distribution Ω_{kd} are drawn from symmetric Dirichlet prior β_k and δ_k . Because a session is related to the same search topic, a topic z is session-specific and drawn from θ . Next, in each session, ϕ_{zd} is a multinomial distribution based on the search topic z and the document d . The binomial distribution X is the indicator to check whether users click the URL in a search session. $X = 1$ means that there exists clickthrough and the URL are drawn from Ω_{zd} based on the search topic z and the document d . Finally, each word w and URL u (if any) are selected by the preference of the document-topic z , and the topic-word w and topic-URL u (if any). The parameters β and δ are document specific so that the TIM captures the burstiness of query terms as well as URLs for each user.

The method of Gibbs sampling [6] for TIM is similar to LDA. The complete likelihood is calculated as follow:

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

If users didn't click any URLs in the session, the conditional probability of the i th session's k th topic is:

$$P(z_i = k | X_i = 0, \mathbf{z}_{-i}, \mathbf{w}, \mathbf{u}, \alpha, \beta, \delta) \propto \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^K (C_{dk'}^{DK} + \alpha_{k'})} \frac{\Gamma(\sum_{t=1}^W (C_{kwd}^{KWD} + \beta_{wk}))}{\Gamma(\sum_{t=1}^W (C_{kwd}^{KWD} + \beta_{wk} + N_{iw}))} \prod_{w=1}^W \frac{\Gamma(C_{kwd}^{KWD} + \beta_{wk} + N_{iw})}{\Gamma(C_{kwd}^{KWD} + \beta_w)} \quad (2)$$

When there is clickthrough in the session, the conditional probability of the i th session's k th topic is:

$$P(z_i = k | X_i = 1, \mathbf{z}_{-i}, \mathbf{w}, \mathbf{u}, \alpha, \beta, \delta) \propto \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^K (C_{dk'}^{DK} + \alpha_{k'})} \frac{\Gamma(\sum_{w=1}^W (C_{kwd}^{KWD} + \beta_{wk}))}{\Gamma(\sum_{w=1}^W (C_{kwd}^{KWD} + \beta_w + N_{iw}))} \prod_{w=1}^W \frac{\Gamma(C_{kwd}^{KWD} + \beta_{wk} + N_{iw})}{\Gamma(C_{kwd}^{KWD} + \beta_{wk})} \frac{\Gamma(\sum_{u=1}^U (C_{kud}^{KUD} + \delta_{uk}))}{\Gamma(\sum_{u=1}^U (C_{kud}^{KUD} + \delta_{uk} + N_{iu}))} \prod_{u=1}^U \frac{\Gamma(C_{kud}^{KUD} + \delta_{uk} + N_{iu})}{\Gamma(C_{kud}^{KUD} + \delta_{uk})} \quad (3)$$

B. Topical Dependence Model (TDM)

What makes the problem more complicated is the fact that query terms and URLs are closely coupled via the search engine, clicked URLs is raised by the corresponding query terms. In the case of web search, the URLs are the results of submitting queries to the search engine. Thus, URLs and search queries are coupled. Consequently, we introduce the variable Δ_{qku} to represent the query-URL multinomial whose prior is denoted by δ , and depict the relation between query and URL. Since the query-URL multinomial distribution can be easily obtained via the widely used click graph. We denote TDM utilize the global bipartite as TDM-G. Since we want to investigate whether focus on the user-centric information can boost the performance of topic modeling. We also build

user-based query-URL bipartite and denote TDM utilizes the user-based query-URL multinomial as TDM-U.

Figure 1(b) presents the generative process of TDM, which is similar to the TIM. The key difference between them is that when $(X = 1)$, the URL is recognized by its search topic z and related query q rather than document d in TIM.

The joint likelihood of generating the query items and URLs is as follows:

$$P(\mathbf{w}, \mathbf{u}, \mathbf{z} | \alpha, \beta, \delta) = P(\mathbf{u} | \mathbf{z}, \delta) P(\mathbf{w} | \mathbf{z}, \beta) P(\mathbf{z} | \alpha). \quad (4)$$

In TDM, $P(\mathbf{z} | \alpha)$ and $P(\mathbf{w} | \mathbf{z}, \beta)$ are the same as TIM. The major distinction between them is that the generation of URL u is decided by the search topic z and the matching query q . Because the query item w and URL u are related in the given search topic. When $X = 1$, the conditional probability of the i th session's k th topic is the same with TIM. But if there exists clickthrough the conditional probability is defined as follows:

$$P(z_i = k | X_i = 1, \mathbf{z}_{-i}, \mathbf{w}, \mathbf{t}, \mathbf{u}, \alpha, \beta, \delta) \propto \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^K (C_{dk'}^{DK} + \alpha_{k'})} \frac{\Gamma(\sum_{t=1}^W (C_{kwd}^{KWD} + \beta_{wk}))}{\Gamma(\sum_{t=1}^W (C_{kwd}^{KWD} + \beta_{wk} + N_{iw}))} \prod_{w=1}^W \frac{\Gamma(C_{kwd}^{KWD} + \beta_{wk} + N_{iw})}{\Gamma(C_{kwd}^{KWD} + \beta_w)} \prod_{q \in s_i} \frac{\Gamma(\sum_{u=1}^U (C_{qzu}^{QZU} + \delta_{qu}))}{\Gamma(\sum_{u=1}^U (C_{qzu}^{QZU} + \delta_{qu} + N_{iu}))} \prod_{u \leftarrow q} \frac{\Gamma(C_{qzu}^{QZU} + \delta_{qu} + N_{iu})}{\Gamma(C_{qzu}^{QZU} + \delta_u)} \quad (5)$$

C. Including Temporal Information

Another phenomenon in web search is the temporal burstiness. A user tend to intensively search some content within a short time period. Therefore, we assume that each user's search trail has a temporal burstiness, which is embodied by the timestamps associated with each query. Since it is tricky to determine the temporal granularity, the temporal burstiness of topics can be captured by a continuous Beta distribution. By introducing the Beta distribution, we enable a topic to be more likely to appear within a short time period. Since the topic-term multinomial distribution is fixed (as for each user), the terms that exist in the topic will demonstrate burstiness. We observe that the burstiness phenomenon can be observed on the day level as well as the hour level, which suggests that a model that do not needs the discretization is preferable.

Based on TIM and TDM model, Figure 1 also gives the main generative process of temporal information. Within a session, if the temporal prominence is on the corpus level, the timestamps are drawn from a Beta distribution $\tau_z(X-TG)$ based on the search topic z , and if the temporal prominence on the user-based level, the timestamps are drawn from a Beta distribution $\tau_{dz}(X-TU)$ based on the session topic z and the document d .

In order to implement Gibbs sampling for TIM-T and TDM-T, we proposed a condensed inference method for their sampling, which is similar to the Gibbs sampling in DCMLDA. We also calculate the complete likelihood of the model:

$$P(\mathbf{w}, \mathbf{u}, \mathbf{t}, \mathbf{z} | \alpha, \beta, \delta, \tau) = P(\mathbf{t} | \mathbf{z}, \tau) P(\mathbf{u} | \mathbf{z}, \delta) P(\mathbf{w} | \mathbf{z}, \beta) P(\mathbf{z} | \alpha). \quad (6)$$

When the temporal prominence on the corpus level, the generative process of timestamp and the temporal parameters

update rules are the same as Topic over Times model(TOT)[7]. But if the temporal prominence is on the user-based level, the generative process of timestamp will be different. We utilize follow equations to update temporal parameters:

$$\tau_{dk1} = t_{dk} \left(\frac{t_{dk}(1 - t_{dk})}{s_{dk}^2} - 1 \right), \quad (7)$$

$$\tau_{dk2} = (1 - t_{dk}) \left(\frac{t_{dk}(1 - t_{dk})}{s_{dk}^2} - 1 \right), \quad (8)$$

where t_{dk} is the mean of sampling, and s_{dk}^2 denote the biased sample variance of topic z 's timestamps in document d .

D. Parameter Estimation

Both TIM model and TDM model have seven unobserved variables, $\alpha, \beta, \delta, \phi, \Omega, \theta$ and z . α, β and δ are hyper-parameters, while others are topic parameters. Given a training set of documents, we learn appropriate values for the variables by alternating between optimizing the topic parameters given the hyper-parameters, and optimizing the hyper-parameters given the topic parameters. We can compute the complete likelihood of TIM and TDM. Based on this, we use single-sample Monte Carlo EM to learn α, β and δ . Algorithm1 summarizes the method.

Algorithm 1 Single-Sample Monte Carlo EM

- 1: start with initial α, β and δ ;
- 2: **repeat**
- 3: Run Gibbs sampling to steady-state;
- 4: Choose a specific topic assignment for each word using Gibbs sampling
- 5: Choose α, β and δ to maximize complete likelihood $p(w, u, t, z | \alpha, \beta, \delta, \tau)$
- 6: **until** Convergence of α, β and δ .

III. EXPERIMENTS

In this experiment, we select a real world major commercial search query data set as the train set. We should divide the query log into many documents, based on its user id, and then we can utilize our proposed probabilistic topic model to analyze the query log. In each document, we segment these query log into sessions by adopting methods proposed in [8]. After processing, we get about 6,500,000 sessions. In order to filter out meaningless queries and URLs, we follow the stopwords lists to filter out queries. As for URLs, the websets such as 'www.google.com' and other popular portals will be removed. Each session's timestamp is decided by the data and time on which the query log was given.

A. Discovered Search Topic and Capture Burstiness

We present the discovered search topics and illustrate the proposed model can accurately predict the timestamps of search query log. For simplicity, we set 50 search topics($K = 50$), and run Gibbs sampler for 1000th iteration to extract the topics.

We present four search topic examples discovered by TIM-TG and TDM-TG on the corpus level in Table I. As a comparison, We also show the TIM-TU and TDM-TU search topics based on the user-based level in Table II. The topic titles in the table are manually added by our own judgement. The

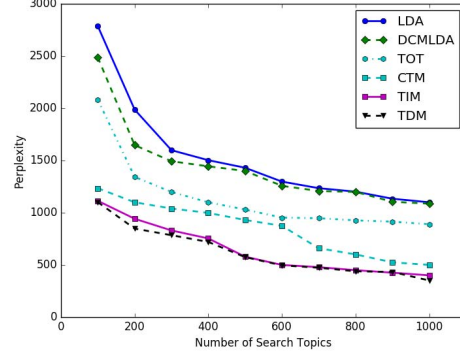


Fig. 2: Perplexity of Models

histograms in the table is the topics' distribution over time, and the curve is the fitted beta PDF. Below the histograms, we present the top five queries and sort by the probability in their topic.

In the leftmost topic, *War*, is an example of how TIM and TDM successfully capture the temporal burstiness. The beta distribution of times show an extra raised in the last few days on May. It is strong associated with Memorial Day on the last Monday of May. "war", "world", "civil" and "cold" are the most frequent words on global corpus level. On the user-based level, we choose a user that with a large number of search query log to analyze the search topic via TIM-TU. In this topic we also observed that words like "iraq", "war" carry out significant information on the user-based level, which is about Iraq war. The result suggests that people pay more attention to the topic of war on Memorial Day, and for the user in Table II, he just concerns about Iraq war.

The above results show the discovered search topics of the proposed models and reveal their temporal burstiness of the search topics.

B. Quantitative Measure

In this section, we compare TIM and TDM with four baseline models by utilizing the perplexity of held-out data. In fact, it is difficult to conduct direct comparison for proposed models since few works focus on using topic models to capture web search burstiness and temporal prominence of topic. As a result, we select three general topic models and a Click-through Topic Model(CTM)[9] as the baselines, namely LDA, DCMLDA, TOT and CTM. Then we perform experimental studies under a general evaluation metric - held-out method.

The Perplexity of Held-out Data: The original training search query log data is separated into two parts, one as train set (training corpus), for the initial frequency estimation; the other is called held-out data. The perplexity of held-out data is a standard measure to evaluate the capability of the generalization and forecasting unknown data. The perplexity is formally defined as follows:

$$Perplexity_{held-out}(M) = \left(\prod_{d=1}^D \prod_{i=1}^{N_d} p(w_i | M) \right)^{\frac{-1}{\sum_{d=1}^D N_d}}, \quad (9)$$

TABLE I: Four topics discovered by TIM-TG and TDM-TG for the data set on the corpus level.

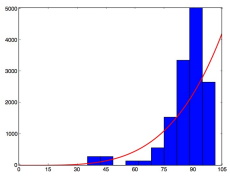
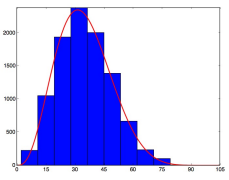
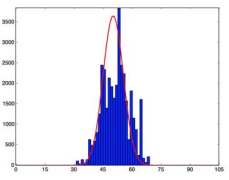
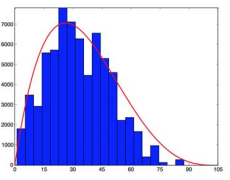
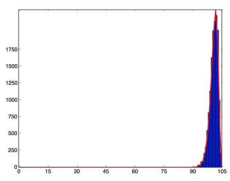
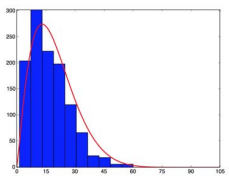
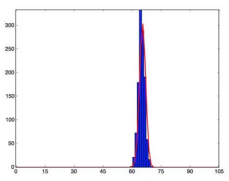
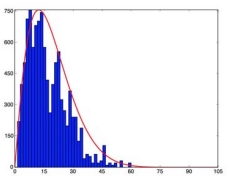
TIM-TG		TDM-TG	
War	Computer	Immigration	Health Care
			
war 0.02864	online 0.02495	immigration 0.06181	care 0.03455
world 0.02766	ebay 0.02451	illegal 0.06067	insurance 0.03072
civil 0.02128	chat 0.02416	law 0.05879	welfare 0.02918
cold 0.01975	myspace 0.02336	us 0.05313	nutrition 0.02883
kill 0.01849	games 0.02092	senate 0.05086	medical 0.02649

TABLE II: Four topics discovered by TIM-TU and TDM-TU for the data set on the user-based level.

TIM-TU		TDM-TU	
War	Computer	Immigration	Health Care
			
war 0.04762	game 0.04225	immigration 0.05382	care 0.03894
iraq 0.04358	pogo 0.03912	Mexican 0.04971	mental 0.03803
navy 0.04302	online 0.03641	illegal 0.04406	nutrition 0.03389
American 0.03275	yahoo 0.03379	reform 0.04397	vitamin 0.03064
kill 0.03028	myspace 0.02311	spanish 0.04081	calcium 0.02802

ACKNOWLEDGMENT

This work is supported in part by the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions, the National Natural Science Foundation of China under Grant 6170128, Shandong Provincial Natural Science Foundation under Grant ZR2017QF00, Human & Social Science Project of MOE (15YJAZH042), the Hong Kong RGC GRF Project 16214716, the Hong Kong RGC Project 16202215, National Grand Fundamental Research 973 Program of China under Grant 2014CB340303, Science and Technology Planning Project of Guangdong Province, China, No. 2015B010110006, NSFC Grant No.61729201, 61232018, Microsoft Research Asia Collaborative Grant, WeBank Collaboration Research Project and NSFC Guang Dong Grant No. U1301253.

REFERENCES

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, 2003.
- [2] D. Jiang and W. Ng, "Mining web search topics with diverse spatiotemporal patterns," in *SIGIR*. ACM, 2013, pp. 881–884.
- [3] D. Jiang, J. Vosecky, K. W. Leung, L. Yang, and W. Ng, "SG-WSTD: A framework for scalable geographic web search topic discovery," *Knowl.-Based Syst.*, vol. 84, pp. 18–33, 2015.
- [4] D. Jiang, K. W. Leung, and W. Ng, "Query intent mining with multiple dimensions of web search data," *World Wide Web*, vol. 19, no. 3, pp. 475–497, 2016.
- [5] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 281–288.
- [6] C. M. Carlo, "Markov chain monte carlo and gibbs sampling," *Notes*, (April), 2004.
- [7] X. Wang and A. McCallum, "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*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 424–433.
- [8] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in *Proc. of the ACM CIKM conference*, 2009.
- [9] D. Jiang, K. W.-T. Leung, W. Ng, and H. Li, "Beyond click graph: Topic modeling for search engine query log analysis," in *International Conference on Database Systems for Advanced Applications*. Springer, 2013, pp. 209–223.

IV. CONCLUSION

In this paper, we propose a series of topic models to study the problem of the user-specific web search analysis from global level and user-based level. TIM focuses on capture the burstiness of web search query, while TDM analyzes the relationship between query terms and URLs. And in order to capture the temporal burstiness, we also propose two variant model TIM-T and TDM-T based on continuous Beta distribution. We also conduct a serious of experiments based on the real search query log, and get better experiment results than many baselines with respect to the perplexity of held-out data. For further plan, we intend to apply these models to targeted group-buying advertising via analyzing user's web search history.