Burstiness-Aware Web Search Analysis on Different Levels of Evidences

Chen Zhang, *Member, IEEE*, Haodi Zhang, Qifan Li, Kaishun Wu, *Member, IEEE*, Di Jiang, Yuanfeng Song, Peiguang Lin, and Lei Chen, *Fellow, IEEE*

Abstract—Personalizing the analysis for web search potentially improves the search experience. A good analytical model for web search should leverage not only collective wisdom but also individual characteristics. Most of the existing analytical models, however, such as the click graph and its variants, focus on how to utilize the collective wisdom, from a crowd, for instance. In this paper, we address the problem of user-specific web search analysis by considering the so-called burstiness in web search, which captures the behavior of rare words appearing many times in a single document. We go beyond click graph and propose two probabilistic topic models, namely, Topic Independence Model and Topic Dependence Model. The former adopts the assumption that the generation of query terms and URLs are topically independent, and the latter captures the coupling between search queries and URLs. We also capture the temporal burstiness of topics by utilizing the continuous Beta distribution. Based on the two proposed models, we propose a novel burstiness-aware search topic rank. 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 evolution of topics on the corpus level and user-based level.

Index Terms—Web search, burstiness, topic model, temporal topic modeling

1 Introduction

Web search analysis has been recognized as one of the centric issues in the past decades. It has been noticed for a long period that the word frequencies in natural languages are roughly inversely a proportion to their rank in the frequency table and therefore, follow a power-law distribution. The distribution was called word burstiness in the context of language models [1]. 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 [2]. 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 [3], [4], 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. Although models like DCM[5] and DCMLDA[2] have 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 components, one is query term and the other is URLs- this is totally different from natural language.

In this paper, we systematically study the burstiness phenomenon in web search via analyzing search query log. In particular, we categorized the burstiness phenomenon into three types, namely, '*meta-word burstiness*', '*query term burstiness*' and '*URL burstiness*', and 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 nature: 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 get a conclusion that the proposed models perform more effectively in a huge number of real query log. Based on the conference version of this paper [6], we propose

- *Chen Zhang is with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, SAR China. E-mail: c4zhang@comp.polyu.edu.hk.*
- *Haodi Zhang is with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai 200031, China, and also with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: zhanghd.ustc@gmail.com.*
- *Qifan Li and Kaishun Wu are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: {qfli, wu}@szu.edu.cn.*
- *Di Jiang and Yuanfeng Song are with the WeBank AI, Shenzhen 518040, China. E-mail: {dijiang, yfsong}@webank.com.*
- *Peiguang Lin is with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250100, China. E-mail: linpg@sdufe.edu.cn.*
- *Lei Chen is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, SAR China. E-mail: leichen@cse.ust.hk.*

Manuscript received 2 Nov. 2020; revised 18 May 2021; accepted 24 Aug. 2021. Date of publication 3 Sept. 2021; date of current version 3 Feb. 2023.
(Corresponding author: Haodi Zhang.)
Recommended for acceptance by C. Li.
Digital Object Identifier no. 10.1109/TKDE.2021.3109304

burstiness-aware search topic rank with TIM and TDM. We compare the topic distances with LDA, DCMLDA and TOT, and rank the significance of discovered search topics. Detailed result and discussion are available in Section 5.

We summarize our contributions in the following three aspects,

- We propose two models, TIM and TDM, to capture the multifaceted burstiness in web search the temporal burstiness via Beta distribution, which fills the gap of word burstiness and user behavioral analysis in web search.
- We propose a novel burstiness-aware method for search topic rank based on TIM and TDM, which achieves better performance in KL divergence compared with existing topic models.
- We experiment with and analyze competing models under different metrics.

We organize the rest of the paper with the following six sections. In Section 2 we present some related research work. In Section 3, we present the burstiness phenomenon of web search through empirical study. In Section 4, we formulate two user-based probability topic models and their temporal variants. In Section 5, we propose the burstiness-aware topic rank with TIM and TDM with hyper parameter estimation. The experiments are demonstrated in Section 6. Finally, Section 7 conclude this work and indicate the further direction.

2 Related Work

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.* [7] proposed Latent Dirichlet Allocation (LDA). LDA have been widely used in both academia and industry. Many variants of LDA also performed quite so good on its area, such as twitter analysis [8], [9], [10], digital articles [11], [12], [13], and web search query log analysis [14], [15], [16], [17], [18]. But LDA and its variants can't capture the burstiness in article. Madsen *et al.* [5] proposed the Dirichlet compound multinomial model (DCM). DCM can alternate multinomial distribution in capturing burstiness, but it can not find out the topic. Elkan [19] proposed the mixture of DCM distribution. This model can model a document that contains words with same topic, but when there exists words with different topics in a document, it can not work very well. Sunehag [1] separated the two-stage conditional presence/abundance issue by utilizing a framework. Doyle *et al.* [2] proposed DCMLDA model, which combines the extend DCM model and LDA, and it is used for capturing word burstiness. Lappas *et al.* [20] proposed that with the increment of time-stamped data collections, such as digitized periodicals, web news and blogs, it becomes more and more important to efficiently index and search these collections. Now, the research about term burstiness has been recognized as a standard mechanism to address event detection in the context of such collections. Sato *et al.* [21] proposed the PY topic model. This model adopts the Pitman-Yor(PY) process. The major distribution of PY process is that can get the power-law distribution of words and the various topics presence.

According to the description above, most of works focus on modeling homogenous items. In order to model web search query log, many works have been done in the last few years. Boldi *et al.* [22] proposed the query-flow graph, which presents the key information of latent search conduct. Fuxman *et al.* [23] analyzed the relation between search queries and URLs. Di *et al.* [24] analyze the limitations of click graph[25] and proposed three new perspectives of probabilistic topic models. These three models can successfully find the search topic models, as well as the relation of query terms and clicked URLs. However as the most extended LDA models, these models also can not capture the burstiness of web search query log.

Burstiness information plays a significant role in improving the performance of search personalization [26], [27]. In this paper, our goal is to improve the performance of search process by analyzing burstiness information. We accept the discrepancy theory concepts to model the query terms and URLs burstiness.

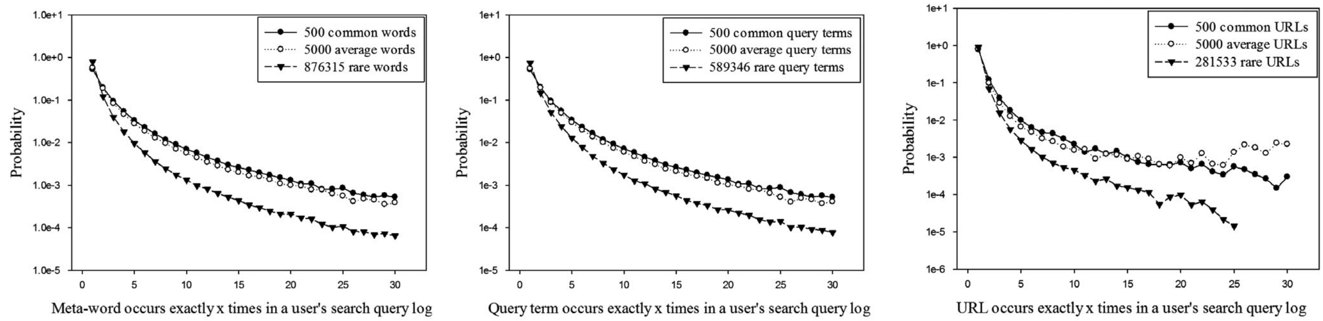
3 User-Based Empirical Analysis

In this section, we empirically analyze the burstiness of web search behaviors. In this experiments, we utilize the search query log data set from a major commercial search engine.

The user-based web search burstiness aims to model the *slang* of each search engine user's search behavior. For example, in order to express the same information, some user tend to use the term *football* while others may use the term *soccer*. Moreover, search query log contain abundance of URL information, which is not independent from the corresponding query terms, i.e., the burstiness of query term may result in the burstiness of clicked URLs.

In order to evaluate the user-based burstiness, we first organize the search query log entries on a user basis. We then regard both query term and URL as *meta-word*. Fig. 1a shows the probability that a word occurs in a document x times. According to the appearance frequency of meta words we split them into three categories-common, average and rare. The first 500 frequent words are named as common words, which represent 0.0453% of the words and 32.64% of the emission. The 501-5500 words represent 0.5705% of the vocabularies and 29.08% of the emission. The rest of vocabularies are rare words, they represent rest 99.3762% words and account for 38.28% of the emissions. In Figs. 1b and 1c, we present the results of similar experiments conducted on query terms and URLs, respectively. From Table 1, we observe that, comparing with the result on industry sector corpus reported in [5], web search enjoys much more rare *meta-words* that take up a much larger proportion of the emissions. Since the widely used multinomial distribution can only correctly model the common words, the phenomena mentioned above suggests the necessity of capturing burstiness in web search.

In Fig. 1, we should notice that the probability of common meta-words is higher than average words and rare words. The decay rate of rare meta-words is lower than common meta-words and average meta-words, although their curves are parallel. As for rare meta-words, once a meta-word has appeared, the probability that the meta-word occurs many times is higher than the common meta-words and average



(a) Counting probabilities of meta-words (b) Counting probabilities of query terms (c) Counting probabilities of URLs

Fig. 1. The phenomenon of user-based web search burstiness.

meta-words. If we unify query terms and URLs as words, it seems that URLs will be overwhelmed by query terms. Moreover, we observe that query terms and URLs have very clear difference. The phenomenon indicates that query terms and URLs should be modeled separately.

4 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 submitted some queries for satisfying that same query need in a time period.

In Section 4.1, we present the Topic Independence Model (TIM) to capture the burstiness of query terms and URLs by two sets of Dirichlet priors. In Section 4.2, we introduce the Topic Dependence Model (TDM), which captures the relation between query terms and URLs. In Section 4.3, we discuss the strategy of enabling TIM and TDM to capture the temporal burstiness.

4.1 Topical Independence Model (TIM)

A search engine user has many different search topics. Therefore, our model must allow a single document to have multiple topics, and account for search burstiness by making the topics and document-specific. For simplicity, TIM adopt an assumption that the generation of query terms and URLs are topically independent, which is shown in the graphical model in Fig. 2a.

TABLE 1
Distribution Comparison

	Common	Average	Rare
Articles(V)	1.5%	9.3%	89.2%
Log(V-W)	0.0453%	0.5705%	99.3762%
Log(V-Term)	0.0806%	0.7479%	99.1715%
Log(V-URL)	0.1094%	1.8152%	98.0754%
Articles(E)	73%	21%	6%
Log(E-W)	32.64%	29.08%	38.28%
Log(E-Term)	34.09%	32.61%	33.30%
Log(E-URL)	30.74%	22.03 %	47.23%

Algorithm 1 presents the generative process of TIM. At first, for each document, we draw a document-specific mix θ_d^z over topics that is drown from a symmetric Dirichlet prior α (Line 1 2). Then in the document d, the distribution of word θ_{kd}^w and the URL distribution θ_{kd}^u are drawn from symmetric dirichlet prior β_k and γ_k (Line 4 5). Because a session is related to the same search topic, a topic z is session-specific and drawn from θ^z. Next, in each session, θ_{zd}^w is a multinomial distribution based on the search topic z and the document d (Line 8 9). The binomial distribution X is the indicator to check whether users click the URL in a search session (Line 10). $X = 1$ means that there exists click-through and the URL are drawn from θ_{zd}^u 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) (Line 12). The parameters β and γ are document specific so that the TIM captures the burstiness of query terms as well as URLs for each user.

Algorithm 1. Generative Process of TIM

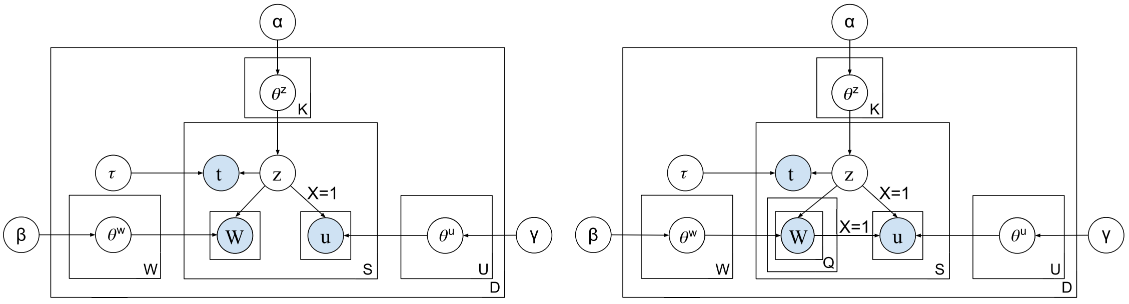
```
1: for document $d \in 1, \dots, D$ do
2:   draw $d$'s topic distribution $\theta_d^z \sim$ Dirichlet($\alpha$);
3:   for topic $k \in 1, \dots, K$ do
4:     draw a word distribution $\theta_{kd}^w \sim$ Dirichlet($\beta_k$);
5:     draw a URL distribution $\theta_{kd}^u \sim$ Dirichlet($\delta_k$);
6:   end for
7:   for each session $s$ in $d$ do
8:     choose a topic $z \sim$ Multinomial($\theta_d^z$);
9:     generate words $w \sim$ Multinomial($\theta_{zd}^w$);
10:    draw $X \sim$ Binomial($\rho$);
11:    if $X = 1$ then
12:      generate URLs $u \sim$ Multinomial($\theta_{zd}^u$);
13:    end if
14:  end for
15: end for
```

The method of Gibbs sampling [28] for TIM is similar to LDA. The complete likelihood is calculated as follow:

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

The probability $P(P(\mathbf{z}|\alpha))$ is the same as that in LDA

$$P(\mathbf{z}|\alpha) = \left(\frac{\Gamma(\sum_{z=1}^{K}\alpha_z)}{\prod_{z=1}^{K}\Gamma(\alpha_z)}\right)^D \prod_{d=1}^{D} \frac{\prod_{z=1}^{T}\Gamma(N_{d,z}+\alpha_z)}{\Gamma(\sum_{z=1}^{Z}(N_{d,z}+\alpha_z))}. \tag{2}$$



(a) TS-TIM

(b) TS-TDM

Fig. 2. Two interpretations of web search.

The probability $P(\mathbf{w}|\mathbf{z},\beta)$ is the same as that in DCMLDA

$$P(\mathbf{w}|\mathbf{z},\beta)=\prod_{d=1}^{D}\prod_{k=1}^{K}\left(\frac{\Gamma(\sum_{w=1}^{W}\beta_{k,w})\prod_{w=1}^{W}\Gamma(N_{d,k,w}+\beta_{k,w})}{\prod_{w=1}^{W}\Gamma(\beta_{k,w})\Gamma(\sum_{w=1}^{W}(N_{d,k,w}+\beta_{k,w}))}\right). \tag{3}$$

If users didn't click any URLs in the session, the conditional probability of the kth topic for the ith session is

$$P(z_i=k|X_i=0,\mathbf{z_{-i}},\mathbf{w},\mathbf{u},\alpha,\beta,\gamma)\propto\frac{C_{k,d}+\alpha_k}{\sum_{k'=1}^{K}(C_{k',d}+\alpha_{k'})}$$
$$\frac{\Gamma(\sum_{t=1}^{W}(C_{k,w,d}+\beta_{w,k}))}{\Gamma(\sum_{t=1}^{W}(C_{k,w,d}+\beta_{w,k}+N_{i,w}))}\prod_{w=1}^{W}\frac{\Gamma(C_{k,w,d}+\beta_{w,k}+N_{i,w})}{\Gamma(C_{k,w,d}+\beta_{w})}. \tag{4}$$

TABLE 2
Notations for Web Search Analysis

Notation	Meaning
D	the number of meta documents
K	the number of search topics
z	a topic
w	a query term
u	a URL
θ^z	multinomial distribution over topics
θ^w	multinomial distribution over query terms (TUM)
θ^u	multinomial distribution over URLs
τ	Beta distribution over timestamps
α	Dirichlet prior vector for θ^z
β	Dirichlet prior vector for θ^w
γ	Dirichlet prior vector for θ^u
ρ	binomial distribution over clickthrough
X	binomial indicator of clickthrough
z_i	the topic of the ith section
$\mathbf{z}_{-i}$	the topics of the sessions except the ith
$\mathbf{w}$	query-term list representation of the corpus
$\mathbf{u}$	URL list representation of the corpus
$C_{k,d}$	the number of sessions that are assigned to topic k in document d
$C_{w,k,d}$	the number of times that query term w is assigned to topic k in document d
$C_{u,k,d}$	the number of times that URL u is assigned to topic k in document d
$C_{q,z,u}$	the number of clicks q upon URL u of topic z
S_d	the number of sessions in document d
$W_{s,d}$	the number of unique meta-words / query terms
$T_{s,d}$	the number of unique timestamps in session s of document d
$N_{w,s,d}$	the number of query term w in session s of document d
$N_{u,s,d}$	the number of ULR u in session s of document d
$N_{t,s,d}$	the number of timestamp t in session s of document d
$N_{w,s}$	the number of query term w in session s
$N_{u,s}$	the number of URL u in session s

When there is clickthrough in the session, the conditional probability of the kth topic for the ith session is

$$P(z_i=k|X_i=1,\mathbf{z_{-i}},\mathbf{w},\mathbf{u},\alpha,\beta,\gamma)\propto\frac{C_{k,d}+\alpha_k}{\sum_{k'=1}^{K}(C_{k',d}+\alpha_{k'})}$$
$$\frac{\Gamma(\sum_{w=1}^{W}(C_{k,w,d}+\beta_{w,k}))}{\Gamma(\sum_{w=1}^{W}(C_{k,w,d}+\beta_{w}+N_{i,w}))}\prod_{w=1}^{W}\frac{\Gamma(C_{k,w,d}+\beta_{w,k}+N_{i,w})}{\Gamma(C_{k,w,d}+\beta_{w,k})}$$
$$\frac{\Gamma(\sum_{u=1}^{U}(C_{k,u,d}+\gamma_{u,k}))}{\Gamma(\sum_{u=1}^{U}(C_{k,u,d}+\gamma_{u,k}+N_{i,u}))}\prod_{u=1}^{U}\frac{\Gamma(C_{k,u,d}+\gamma_{u,k}+N_{i,u})}{\Gamma(C_{k,u,d}+\gamma_{u,k})}. \tag{5}$$

4.2 Topical Dependence Model (TDM)

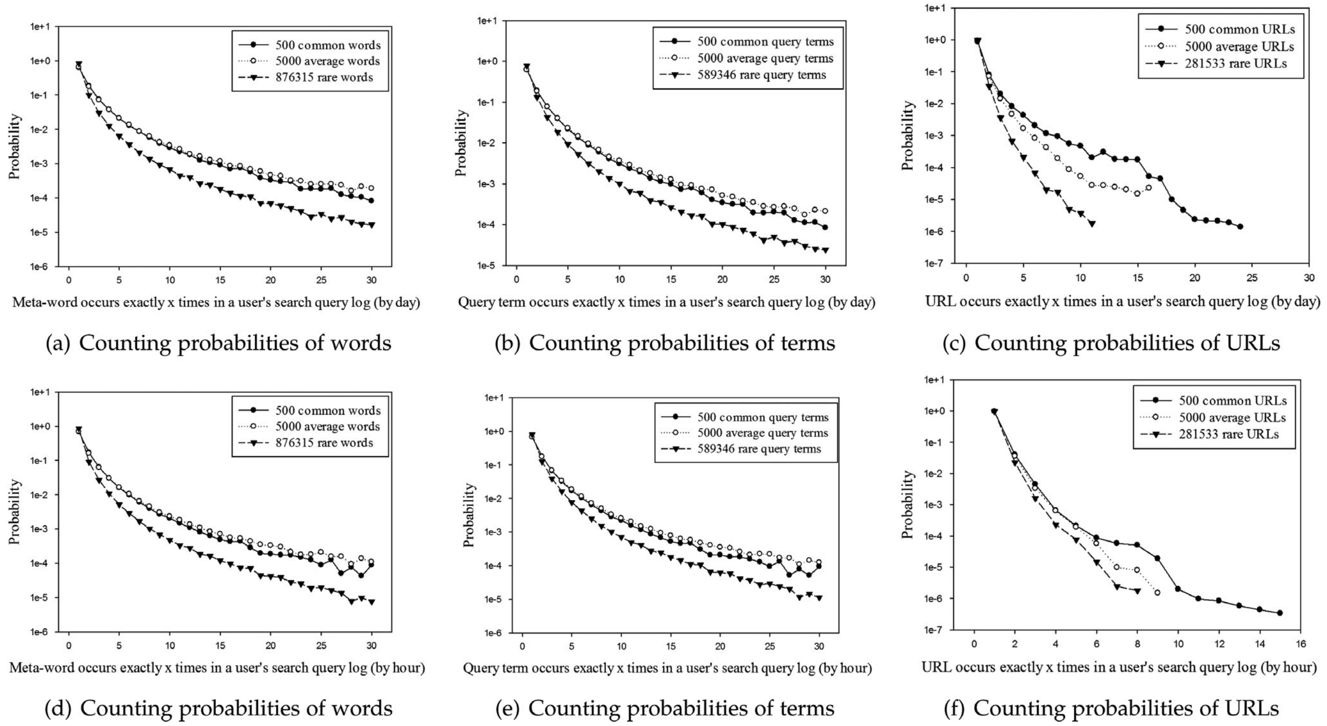
What makes the problem more complicated is the fact that query terms and URLs are closely coupled via the search engine, clicked URLs is raised by the corresponding query terms. In the case of web search, the URLs are the results of submitting queries to the search engine. Thus, URLs and search queries are coupled. Consequently, we introduce the variable $\Delta_{q,k,u}$ to represent the query-URL multinomial whose prior is denoted by γ, and depict the relation between query and URL. Since the query-URL multinomial 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.

Algorithm 2 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(Line 12 13).

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

$$P(\mathbf{w},\mathbf{u},\mathbf{z}|\alpha,\beta,\gamma)=P(\mathbf{u}|\mathbf{z},\mathbf{w},\gamma)P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{z}|\alpha). \tag{6}$$

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.



(a) Counting probabilities of words
(b) Counting probabilities of terms
(c) Counting probabilities of URLs
(d) Counting probabilities of words
(e) Counting probabilities of terms
(f) Counting probabilities of URLs

Fig. 3. The phenomenon of the user's time-based web search burstiness.

$$
\begin{aligned}
P(\mathbf{u}|\mathbf{z},\mathbf{w},\gamma) &= \int \prod_{d=1}^{D}\prod_{i=1}^{N_d} P(u_{d,i}|\Delta_{w_{d,i}z_{d,i}}) \prod_{w=1}^{W}\prod_{z=1}^{K} p(\Delta_{w,z}|\gamma) d\Delta \\
&= \int \prod_{z=1}^{K}\prod_{w=1}^{W}\prod_{u=1}^{U} \Delta_{w,z,u}^{N_{w,z,u}} \prod_{w=1}^{W}\prod_{z=1}^{K} \left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{w,u})}{\prod_{u=1}^{U}\Gamma(\gamma_{w,u})} \prod_{u=1}^{U} \Delta_{w,z,u}^{\gamma_{w,u}-1} \right) d\Delta \\
&= \prod_{w=1}^{W} \left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{w,u})}{\prod_{u=1}^{U}\Gamma(\gamma_{w,u})} \right)^{K} \times \prod_{z=1}^{K}\prod_{w=1}^{W} \frac{\prod_{u=1}^{U}\Gamma(N_{w,z,u}+\gamma_{w,u})}{\Gamma(\sum_{u=1}^{U}(N_{w,z,u}+\gamma_{w,u}))}.
\end{aligned}
\tag{7}
$$

When $X = 1$, the conditional probability of the kth topic for the ith session is the same with TIM. But if there exists clickthrough the conditional probability is defined as follows:

$$
\begin{aligned}
& P(z_i = k|X_i = 1, \mathbf{z}_{-\mathbf{i}}, \mathbf{w}, \mathbf{t}, \mathbf{u}, \alpha, \beta, \gamma) \propto \frac{C_{k,d} + \alpha_k}{\sum_{k'=1}^{K}(C_{k',d} + \alpha_{k'})} \\
& \frac{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k}))}{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k} + N_{i,w}))} \prod_{w=1}^{W} \frac{\Gamma(C_{k,w,d} + \beta_{w,k} + N_{i,w})}{\Gamma(C_{k,w,d} + \beta_{w})} \\
& \prod_{q\in s_i} \frac{\Gamma(\sum_{u=1}^{U}(C_{q,z,u} + \gamma_{q,u}))}{\Gamma(\sum_{u=1}^{U}(C_{q,z,u} + \gamma_{q,u} + N_{i,u}))} \\
& \prod_{u\leftarrow q} \frac{\Gamma(C_{q,z,u} + \gamma_{q,u} + N_{i,u})}{\Gamma(C_{q,z,u} + \gamma_{u})}.
\end{aligned}
\tag{8}
$$

4.3 Including Temporal Information

Another phenomenon in web search is the temporal burstiness. Fig. 3 presents the phenomenon of the user's time-based web search burstiness. A user tends 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. In this case, we can make each topic's temporal prominence on the corpus level (which is denoted as X-TG) as well as user-based (which is denoted as X-TU).

Algorithm 2. Generative Process of TDM

```
for document $d \in 1, \dots, D$ do
  draw $d$'s topic distribution $\theta_d^z \sim$ Dirichlet($\alpha$);
  for topic $k \in 1, \dots, K$ do
    draw a word distribution $\theta_{kd}^w \sim$ Dirichlet($\beta_k$);
    draw a URL distribution $\theta_{kd}^u \sim$ Dirichlet($\delta_k$);
  end for
  for each session $s$ in $d$ do
    choose a topic $z \sim$ Multinomial($\theta_d^z$);
    generate words $w \sim$ Multinomial($\theta_{zd}^w$);
    for all $q \in s$ do
    draw $X \sim$ Binomial($\rho$);
    if $X = 1$ then
      generate URL $u \sim$ Multinomial($\theta_{qz}^u$);
    end if
  end for
end for
end for
```

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, Algorithm 3 gives the main generative process of temporal information. Within a session, for the temporal prominence on the corpus level, the timestamps are drown from a Beta distribution τ_z(X-TG) based on the search topic z, and for the temporal prominence on the user-based level, the timestamps are drown from a Beta distribution τ_{dz}(X-TU) based on the session topic z and the document d (Line 4).

Algorithm 3. Including Temporal Information

```
1: the same as the original model
2: for each session $s$ in $d$ do
3:     choose a topic $z \sim$ Multinomial($\theta_d^z$);
4:     generate timestamps $t \sim$ Beta($\tau_z$) (X-TG) or $t \sim$ Beta($\tau_{dz}$)
       (X-TU)
5:     the same as the original model
6: end for
```

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

$$P(\mathbf{w}, \mathbf{u}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \gamma, \tau) = P(\mathbf{t}|\mathbf{z}, \tau)P(\mathbf{u}|\mathbf{z}, \gamma)P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{z}|\alpha). \tag{9}$$

When the temporal prominence on the corpus level, the generative process of timestamp and the temporal parameters update rules are the same as TOT. But if the temporal prominence on the user-based level, the generative process of timestamp will be different. For TIM and TDM, when there exists no clickthrough, after combining terms, the conditional probability of the kth topic for the ith session is defined as (10)

$$\begin{aligned}&P(z_i = k|X_i = 0, \mathbf{z_{-i}}, \mathbf{w}, \mathbf{u}, \alpha, \beta, \gamma, \tau) \propto\\ &\prod_{j=1}^{T} \frac{(1-t_j)^{\tau_{dk1}-1} t_j^{\tau_{dk2}-1}}{B(\tau_{dk1}, \tau_{dk2})} \frac{C_{k,d} + \alpha_k}{\sum_{k'=1}^{K}(C_{k',d} + \alpha_{k'})}\\ &\frac{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k}))}{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k} + N_{i,w}))} \prod_{w=1}^{W} \frac{\Gamma(C_{k,w,d} + \beta_{w,k} + N_{i,w})}{\Gamma(C_{k,w,d} + \beta_{w})}.\end{aligned} \tag{10}$$

For TIM when there exists clickthrough, after combining terms, the conditional probability of the kth topic for the ith session is defined as follows:

$$\begin{aligned}&P(z_i = k|X_i = 1, \mathbf{z_{-i}}, \mathbf{w}, \mathbf{u}, \alpha, \beta, \gamma, \tau) \propto\\ &\prod_{j=1}^{T} \frac{(1-t_j)^{\tau_{dk1}-1} t_j^{\tau_{dk2}-1}}{B(\tau_{dk1}, \tau_{dk2})} \frac{C_{k,d} + \alpha_k}{\sum_{k'=1}^{K}(C_{k',d} + \alpha_{k'})}\\ &\frac{\Gamma(\sum_{w=1}^{W}(C_{k,w,d} + \beta_{w,k}))}{\Gamma(\sum_{w=1}^{W}(C_{k,w} + \beta_{w} + N_{i,w}))} \prod_{w=1}^{W} \frac{\Gamma(C_{k,w,d} + \beta_{w,k} + N_{i,w})}{\Gamma(C_{k,w,d} + \beta_{w,k})}\\ &\frac{\Gamma(\sum_{u=1}^{U}(C_{k,u,d} + \gamma_{u,k}))}{\Gamma(\sum_{u=1}^{U}(C_{k,u,d} + \gamma_{u,k} + N_{i,u}))} \prod_{u=1}^{U} \frac{\Gamma(C_{k,u,d} + \gamma_{u,k} + N_{i,u})}{\Gamma(C_{k,u,d} + \gamma_{u,k})}.\end{aligned} \tag{11}$$

For TDM when there exists clickthrough, after combining terms, the conditional probability of the kth topic for the ith session is defined as follows:

$$\begin{aligned}&P(z_i = k|X_i = 1, \mathbf{z_{-i}}, \mathbf{w}, \mathbf{t}, \mathbf{u}, \alpha, \beta, \gamma, \tau) \propto\\ &\prod_{j=1}^{T} \frac{(1-t_j)^{\tau_{dk1}-1} t_j^{\tau_{dk2}-1}}{B(\tau_{dk1}, \tau_{dk2})} \frac{C_{k,d} + \alpha_k}{\sum_{k'=1}^{K}(C_{k',d} + \alpha_{k'})}\\ &\frac{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k}))}{\Gamma(\sum_{t=1}^{W}(C_{k,w,d} + \beta_{w,k} + N_{i,w}))} \prod_{w=1}^{W} \frac{\Gamma(C_{k,w,d} + \beta_{w,k} + N_{i,w})}{\Gamma(C_{k,w,d} + \beta_{w})}\\ &\prod_{q \in s_i} \frac{\Gamma(\sum_{u=1}^{U}(C_{q,z,u} + \gamma_{q,u}))}{\Gamma(\sum_{u=1}^{U}(C_{q,z,u} + \gamma_{q,u} + N_{i,u}))} \prod_{u \leftarrow q} \frac{\Gamma(C_{q,z,u} + \gamma_{q,u} + N_{i,u})}{\Gamma(C_{q,z,u} + \gamma_{u})}.\end{aligned} \tag{12}$$

We utilize follow equations to update temporal parameters:

$$\tau_{dk1} = t^{-}_{d,k}\left(\frac{t^{-}_{d,k}(1 - t^{-}_{d,k})}{s^2_{d,k}} - 1\right), \tag{13}$$

$$\tau_{dk2} = (1 - t^{-}_{d,k})\left(\frac{t^{-}_{d,k}(1 - t^{-}_{d,k})}{s^2_{d,k}} - 1\right), \tag{14}$$

where $t^{-}_{d,k}$ is the mean of sampling, and $s^2_{d,k}$ denote the biased sample variance of topic z's timestamps in document d.

5 Burstiness-Aware Search Topic Rank

In this section, we introduce our burstiness-aware search topic rank with TIM and TDM. Search Topic Rank is an effective method to verify the importance of discovered topic. We successfully apply KL-divergence to compare their topic distances of TIM and TDM with some commonly recognized baselines, and then rank the significance of discovered search topics. Notice that the hyper parameters of TIM and TDM need to be learned. Similar to DCMLDA, θ^w and θ^u in TIM and TDM relates to the values β and γ in the models, respectively.

5.1 Parameter Estimation For TIM and TDM

The complete likelihood of TIM $p(w, u, z|\alpha, \beta, \gamma)$ is computed as follows,

$$\begin{aligned}P(\mathbf{w}, \mathbf{u}, \mathbf{z}|\alpha, \beta, \gamma) = &\prod_{d}\left(\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{d,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K}(m_{d,k} + \alpha_k))}\right.\\ &\prod_{d,k}\left(\left(\frac{\Gamma(\sum_{w=1}^{W}\beta_{w,k})}{\prod_{w=1}^{W}\Gamma(\beta_{w,k})}\right)\frac{\prod_{w=1}^{W}\Gamma(N_{k,w,d} + \beta_{w,k})}{\Gamma(\sum_{w=1}^{W}(N_{k,w,d} + \beta_{w,k}))}\right.\\ &\left.\left(\frac{\Gamma(\sum_{u=1}^{U}\gamma_{u,k})}{\prod_{u=1}^{U}\Gamma(\gamma_{u,k})}\right)\frac{\prod_{u=1}^{U}\Gamma(N_{k,u,d} + \gamma_{u,k})}{\Gamma(\sum_{u=1}^{U}(N_{k,u,d} + \gamma_{u,k}))}\right).\end{aligned} \tag{15}$$

The complete likelihood of TDM $p(w, u, z|\alpha, \beta, \gamma)$ is computed as follows,

$$\begin{aligned}P(\mathbf{w}, \mathbf{u}, \mathbf{z}|\alpha, \beta, \gamma) = &\prod_{d}\left(\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{d,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K}(m_{d,k} + \alpha_k))}\right.\\ &\prod_{d,k}\left(\left(\frac{\Gamma(\sum_{w=1}^{W}\beta_{w,k})}{\prod_{w=1}^{W}\Gamma(\beta_{w,k})}\right)\frac{\prod_{w=1}^{W}\Gamma(N_{k,w,d} + \beta_{w,k})}{\Gamma(\sum_{w=1}^{W}(N_{k,w,d} + \beta_{w,k}))}\right)\\ &\prod_{k=1}^{K}\prod_{q=1}^{Q}\left(\left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{q,u})}{\prod_{u=1}^{U}\Gamma(\gamma_{q,u})}\right)\frac{\prod_{u=1}^{U}\Gamma(N_{q,k,u} + \gamma_{q,u})}{\Gamma(\sum_{u=1}^{U}(N_{q,k,u} + \gamma_{q,u}))}\right).\end{aligned} \tag{16}$$

The complete likelihood of TIM-T $p(w, u, t, z|\alpha, \beta, \gamma, \tau)$ is computed as follows,

$$\begin{aligned}&P(\mathbf{w}, \mathbf{u}, \mathbf{z}|\alpha, \beta, \gamma) = \prod_{d} \Bigg(\Bigg(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \Bigg) \frac{\prod_{k=1}^{T} \Gamma(N_{k,d} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (N_{k,d} + \alpha_k))} \\ &\prod_{k,d} \Bigg(\Bigg(\frac{\Gamma(\sum_{w=1}^{W} \beta_{w,k})}{\prod_{w=1}^{W} \Gamma(\beta_{w,k})} \Bigg) \frac{\prod_{w=1}^{W} \Gamma(N_{w,k,d} + \beta_{w,k})}{\Gamma(\sum_{w=1}^{W} (N_{w,k,d} + \beta_{w,k}))} \\ &\Bigg(\frac{\Gamma(\sum_{u=1}^{U} \gamma_{u,k})}{\prod_{u=1}^{U} \Gamma(\gamma_{u,k})} \Big) \frac{\prod_{u=1}^{U} \Gamma(N_{u,k,d} + \gamma_{u,k})}{\Gamma(\sum_{u=1}^{U} (N_{u,k,d} + \gamma_{u,k}))} \Bigg) \\ &\prod_{d,s,i} p(t_{dsi}|\tau_{dk_s}).\end{aligned} \tag{17}$$

The complete likelihood of TDM-T $p(w, u, t, z|\alpha, \beta, \gamma, \tau)$ is computed as follows,

$$\begin{aligned}&P(\mathbf{w}, \mathbf{u}, \mathbf{z}|\alpha, \beta, \gamma, \tau) = \prod_{d} \Bigg(\Bigg(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \Bigg) \frac{\prod_{k=1}^{T} \Gamma(m_{d,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (m_{d,k} + \alpha_k))} \\ &\prod_{d,k} \Bigg(\Bigg(\frac{\Gamma(\sum_{w=1}^{W} \beta_{w,k})}{\prod_{w=1}^{W} \Gamma(\beta_{w,k})} \Bigg) \frac{\prod_{w=1}^{W} \Gamma(N_{k,w,d} + \beta_{w,k})}{\Gamma(\sum_{w=1}^{W} (N_{k,w,d} + \beta_{w,k}))} \Bigg) \\ &\prod_{z=1}^{K} \prod_{q=1}^{Q} \Bigg(\Bigg(\frac{\Gamma(\sum_{u=1}^{U} \delta_{q,u})}{\prod_{u=1}^{U} \Gamma(\gamma_{q,u})} \Bigg) \frac{\prod_{u=1}^{U} \Gamma(N_{q,k,u} + \gamma_{q,u})}{\Gamma(\sum_{u=1}^{U} (N_{q,k,u} + \gamma_{q,u}))} \Bigg) \\ &\prod_{d,s,i} p(t_{dsi}|\tau_{dk_s}).\end{aligned} \tag{18}$$

It can be further converted to log-likelihood as follows,

$$\begin{aligned}\alpha'_{.} &= \sum_{d,k} (\log \Gamma(N_{.k,d} + \alpha_k) - \log \Gamma(\alpha_k)) \\ &\quad + \sum_{d} (\log \Gamma\Bigg(\sum_{k} \alpha_k \Bigg) - \log \Gamma\Bigg(\sum_{k} N_{.k,d} + \alpha_k) \Bigg)\end{aligned} \tag{19}$$

$$\begin{aligned}\beta'_{.k} &= \sum_{d,k,w} (\log \Gamma(N_{w,k,d} + \beta_{w,k}) - \log \Gamma(\beta_{w,k})) \\ &\quad + \sum_{d,k} (\log \Gamma\Bigg(\sum_{w} \beta_{w,k} \Bigg) - \log \Gamma\Bigg(\sum_{w} N_{w,k,d} + \beta_{w,k}) \Bigg).\end{aligned} \tag{20}$$

For the non-coupling formula,

$$\begin{aligned}\gamma'_{.k} &= \sum_{d,k,u} (\log \Gamma(N_{u,k,d} + \gamma_{u,k}) - \log \Gamma(\gamma_{u,k})) \\ &\quad + \sum_{d,k} (\log \Gamma\Bigg(\sum_{u} \gamma_{u,k} \Bigg) - \log \Gamma\Bigg(\sum_{u} N_{u,k,d} + \gamma_{u,k}) \Bigg).\end{aligned} \tag{21}$$

For the coupling formula,

$$\begin{aligned}\gamma'_{.q} &= \sum_{q,k,u} (\log \Gamma(N_{q,k,u} + \gamma_{q,u}) - \log \Gamma(\gamma_{q,u})) \\ &\quad + \sum_{q,k} (\log \Gamma\Bigg(\sum_{u} \gamma_{q,u} \Bigg) - \log \Gamma\Bigg(\sum_{u} N_{q,k,u} + \gamma_{q,u}) \Bigg).\end{aligned} \tag{22}$$

The above Equations (19), (20), (21) and (22) define vectors, $\alpha_{.}$, $\beta_{.k}$, $\gamma_{.k}$ and $\gamma_{.q}$ respectively. The maximization is implemented by BFGS with limited memory, as detailed in Algorithm 4.

Algorithm 4. Single-Sample Monte Carlo EM

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

5.2 Burstiness-Aware Search Topic Rank

We define a background distribution as the insignificant topics. Loulwah [29] proposed that a topic can not have a clear and certain identity, if it is used to generate words in an extensive documents, or the documents in excessive case. The beat choice of the background terms are the topics that are irrelevant to the document. For each document, the probability of background topic is equal. So, under the background topic $\theta^{u,B}$, Equation (23) gives the probability of each document d_m.

$$p(d_m|\theta^{u,B}) = \frac{1}{D}, \text{ for } m \in \{1, 2, \ldots, D\}. \tag{23}$$

We get a topic vector $(\theta_{i1}^{z}, \theta_{i2}^{z}, \ldots, \theta_{in}^{z})$. Thus, the probability of the kth search topic for the ith user can be described like this

$$\theta_{ki}^{z} = \frac{\theta_{ik}^{z}}{\sum_{i=1}^{D} \theta_{ik}^{z}}. \tag{24}$$

The distance between the discovered topic and the background topic represents the significance of the discovered topic. Following [30], in our experiments, we utilize the KL-divergence to measure the distance between background topic and the discovered topics from proposed models. If a document carries more discovered topics and less "background" topics, the value of KL-divergence will be high. The average KL-distance of search topics is shown in Fig. 5. We find that in any different number of search topics, the KL-divergence of TDM and TIM are both higher than the three baselines LDA, DCMLDA and TOT. This result is coherent with our expectation that TIM and TDM are useful for finding some burstiness and temporal issues.

Next, we should define the junk topic distribution. Following [29], a junk topic means all the terms of the dictionary has equal probability. So we can take a uniform distribution over the corpus as the junk topic distribution.

$$P(w_i|\theta^{u}) = \frac{1}{W}, \text{ for } i \in \{1, 2, \ldots, W\}. \tag{25}$$

Then, we can compute the distance between the words belong to discovered search topic $\theta^{q,k}$ and the uniform junk word distribution by utilizing KL-divergence. The lower

TABLE 3
Topic Rank(# Represents Labeled Topics)

Rank	Title	Search Item	Score
1	Employment$^{\#}$	employment company job	2.714
2	Food$^{\#}$	hungry fish hamburg cheese	2.659
3	Computer$^{\#}$	online myspace internet java	2.326
4	Health$^{\#}$	healthcare insurance welfare	2.302
5	Art$^{\#}$	gallery studio paris artists	2.287
46	23	ro ny wifi drudge	0.024
47	16	wtue map clay match koa	0.017
48	44	ohio jack cum nurse	0.016
49	18	foose sexy tony noaa	0.012
50	39	tv air fl al	0.009

distance corresponds to distribution with probability mixture models that represent insignificant topics.

Finally, we calculate the score of each topic, the simplest formula of score is that

$$final\_score_k = \Psi * (S_{k,d} + S_{k,w}), \tag{26}$$

where Ψ is the weight of the insignificance standard in the total score, the S_{kd} is the distance between the kth estimate topic and the insignificance background topic, and S_{kw} is the distance between the kth estimate topic and the uniform junk topic. The ranklist of the topic is given by sorting the topics based on their final score. In order to verify our strategy, we manually labeled a title for a part of discovered search topics by experts. Table 3 shows a part of the topics in the result. The result shows that the top ranked topics are all labeled, while the low score topics are ranked at the end of the table, which indicates that the labeled topics are consistent with the final results.

6 Experiments

We present the experiment of proposed models. Section 6.1 prepare the data sets of this experiment. Section 6.2 reports some discovered search topics of the models and analyzes their temporal characters. In Section 6.3 we compare the goodness-of-fit the proposed models. Section 6.4 present an effective application of these models - rank topic.

6.1 Data Sets Preparation

In this experiment, we select a real world major commercial search query data set as the train set. It contains contain users and search queries and is sorted by anonymous user ID and sequentially arranged. 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 query log. In each document, we segment these query log into sessions by adopting methods proposed in [31]. After processing, we get about 6,500,000 sessions. In order to filter out meaningless queries and URLs, we follow the stopword lists that recommended in [32] to filter out queries. As for URLs, the websites such as 'www.google.com' and 'www.bing.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. If the timestamp is the earliest, we normalize it as 0, and the lasted timetamps is normalized as 1. In the end, a document associates with a user's search query log, each document concludes a number of search sessions and each search session has query terms, URLs(if clicked) and a timestamp.

6.2 Discovered Topics and Captured Burstiness

The plausibility of the discovered search topic is a significant measure to judge the success of the proposed model. In this experiment, 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 4. As a comparison, We also show the TIM-TU and TDM-TU search topics based on the user-based level in Table 5. The topic

TABLE 4
Four Topics Discovered by TIM-TG and TDM-TG for the Data Set on the Corpus Level

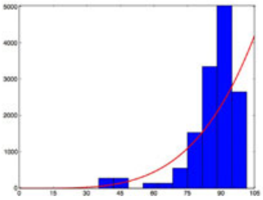
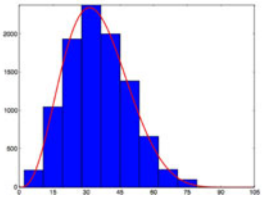
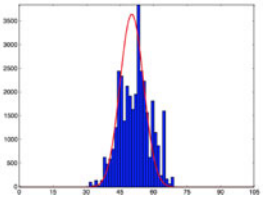
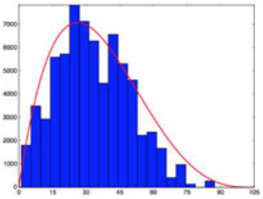
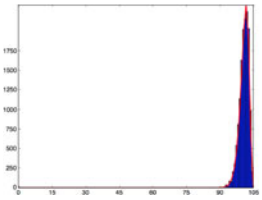
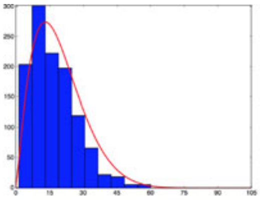
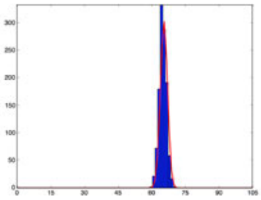
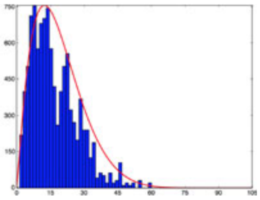
TIM-TG				TDM-TG			
War		Computer		Immigration		Health Care	
							
war	0.02864	online	0.02495	immigration	0.06181	care	0.03455
world	0.02766	ebay	0.02451	illegal	0.06067	insurance	0.03072
civil	0.02128	chat	0.02416	law	0.05879	welfare	0.02918
cold	0.01975	myspace	0.02336	us	0.05313	nutrition	0.02883
kill	0.01849	games	0.02092	senate	0.05086	medical	0.02649
national	0.01821	photoshop	0.01985	forbidding	0.05000	issue	0.02587
die	0.01738	network	0.01773	maxican	0.04865	community	0.02501
America	0.01502	java	0.01219	American	0.04518	ill	0.02486
veterans	0.01215	desktop	0.01131	reform	0.04457	public	0.02249
country	0.01193	internet	0.01004	Israel	0.04396	mental	0.02163

TABLE 5
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	spanlish	0.04081	calcium	0.02802
casualties	0.02867	internet	0.02004	senate	0.03883	walk	0.02659
veterans	0.01973	ps2	0.01435	usa	0.03702	oil	0.02384
wounded	0.01924	download	0.01201	protest	0.03302	medical	0.02319
soldier	0.01718	phone	0.01159	latina	0.03214	club	0.02237
blood	0.01502	play	0.01084	report	0.03029	health	0.02206

titles in the table are manually added by our own judgement. The histograms in the table are the topics' distribution over time, and the curve is the fitted beta PDF. Blow the histograms, we present top ten 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 4, he just concerns about Iraq war.

The second topic in Tables 4 and 5, *Computer*, shows the wave of the topic, and get to a peak on about the 30th day. On the corpus level, “online”, “ebay”, “chat” are the most frequent terms. All of these words belong to the topic *Computer*, and the fact that online chat and electronic commerce was very popular at that time. On the user-based level, The topic *Computer* is composed of the terms like “game”, “pogo” and “online”. “pogo” is a popular online game website, and these words are closely related to the computer games.

In the search topic *Immigration*, TDM-TG clearly shows that the immigration reform occurred between the 40th day and the 75th day, which is frequently associated with the immigration reform protests in April. The search topic “Immigration” primarily contains with “immigration”, “illegal”, “law”, “us” and “senate”, since we use the global bipartite. On the user-used level, the topic is correctly localized in time by TDM-TU. “Immigration”, “Mexican” and “illegal” are the critical words, which prove that the user focuses more on Mexican illegal immigrations.

The rightmost topic is *Health Care*. There's a quite difference between global and user-based query-URL bipartite. In Table 4, the words like “care”, “insurance” and “welfare” are all related to health care, but they emphasize public health care department. In Table 5, on the user-based level, the search topic *Health Care* contain with “care”, “mental”, nutrition” and “vitamin”, all of which are about private health care.

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

6.3 Quantitative Measure

In this section, we compare TIM and TDM with three baseline models by utilizing two metrics. 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 as the baselines, namely Latent Dirichlet Allocation (LDA) [7], DCMLDA [5], and Topic over Time(TOT) [33]. Then we perform experimental studies under two general evaluation metrics - held-out method and cross validation method.

1) *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 [12]. The perplexity is formally defined as follows:

$$Perplexity_{held-out}(M) = \left(\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(w_i|M)\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}. \tag{27}$$

In the Equation (27), N_d is the number of words in each document, M refers to the parameter of the

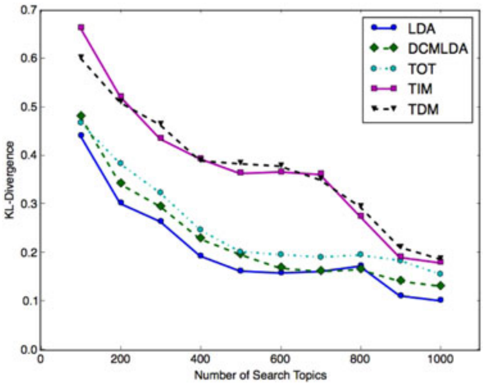
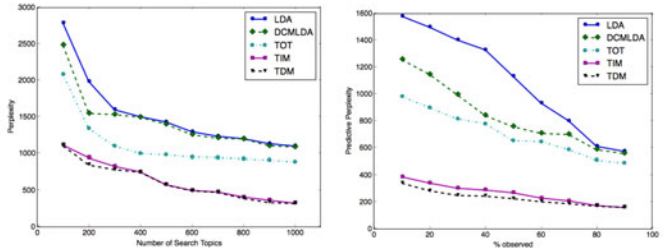


Fig. 4. The KL-divergence of different models.



(a) Perplexity of Models for Held-out Data (b) Perplexity of Models for a portion of Observed Data

Fig. 5. Perplexity of models for different metrics.

trained model. The lower perplexity is, the better generalization performance is. Fig. 4a shows that TIM and TDM has better performance than the three baselines in predicting further search query. As we can see, when the search topic number is 1,000, TIM and TDM achieve perplexity of 324.47 and 320.97, while the perplexity of LDA is 1100.02, that of DCMLDA is 1087.03 and that of TOT is 890.08. The perplexity of TIM and TDM are much lower than the three baselines, which suggests that TIM and TDM have a better capability to analyze web search topic.

2) *The Perplexity of Observed Data:* This approach aims to estimate each model's ability of predicting the rest search query, after observing a part of the search query log. Assume that we want to find a model with a more effective predictive distribution $p(w|w_{1:P})$, after observing the predictive distribution of the query $w_1 : P$. More specifically, we choose a portion of training data and the rest as the testing data. We can calculate the perplexity according to the Equation (28)

$$Perplexity_{portion}(M) = \left(\prod_{d=1}^{D}\prod_{i=P+1}^{N_d} p(w_i|M, W_{a:P})\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}. \tag{28}$$

The predictive perplexity for partially observed data was present in the Fig. 4b. We observe that when the percentage of the training dataset is 60%, LDA shows a perplexity of 934.90, the perplexity of DCMLDA is 707.92, and TOT demonstrates a perplexity of 647.52. The proposed models show significantly outperform than the three baselines. The perplexity of TIM is 230.76, and TDM achieves perplexity of 203.17. The experiment result shows that when we have observed a part of user's search query data, TIM and TDM are more suitable for predicting user's search query trail in the future.

7 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. Topic Independence Model(TIM) focuses on capture the burstiness of web search query, while Topic Dependence Model(TDM) analyzes the relationship between query terms and URLs. 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 series of experiments based on the real search query log, and get better experiment results than many baselines with respect to different metrics. For further plan, we intend to apply these models to targeted group-buying advertising via analyzing user's web search history.

Acknowledgments

Chen Zhang and Haodi Zhang have contributed equally to this work. The authors are grateful to the crowd workers for their efforts to complete the experiment. The work was partially supported by the National Natural Science Foundation of China under Grants 61806132, 61729201, U2001207, 61872248, Guangdong NSF 2017A030312008, Shenzhen Science and Technology Foundation ZDSYS20190902092853047, R2020A045, the Project of DEGP 2019KCXTD005, the Guangdong Pearl River Talent Recruitment Program 2019 ZT08X603, Hong Kong RGC GRF Project 16202218 CRF Projects C6030-18G, C1031- 18G, C5026-18G, AOE Project AoE/E-603/18, Guangdong Basic and Applied Basic Research Foundation 2019B151530001, Hong Kong ITC ITF grants ITS/044/18FX and ITS/470/18FX, Microsoft Research Asia Collaborative Research Grant, Didi-HKUST joint research lab project, and Wechat and Webank Research Grants.

References

[1] P. Sunehag, "Using two-stage conditional word frequency models to model word burstiness and motivating TF-IDF," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007.

[2] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 281–288.

[3] F. Ahmad and G. Kondrak, "Learning a spelling error model from search query logs," in *Proc. Conf. Hum. Lang. Technol. Empir. Methods Natural Lang. Process.*, 2005, pp. 955–962. [Online]. Available: http://dx.doi.org/10.3115/1220575.1220695

[4] W. Gao *et al.*, "Cross-lingual query suggestion using query logs of different languages," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 463–470.

[5] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the dirichlet distribution," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 545–552.

[6] C. Zhang, S. Zhang, C. Lei, and P. Lin, "Burstiness in query log: Web search analysis by combining global and local evidences," in *Proc. 34th IEEE Int. Conf. Data Eng.*, 2018, pp. 1388–1391.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] J. Vosecky, D. Jiang, and W. Ng, "Limosa: A system for geographic user interest analysis in Twitter," in *Proc. 16th Int. Conf. Extending Database Technol.*, 2013, pp. 709–712.

[9] J. Vosecky, D. Jiang, K. W. Leung, K. Xing, and W. Ng, "Integrating social and auxiliary semantics for multifaceted topic modeling in Twitter," *ACM Trans. Internet Technol.*, vol. 14, no. 4, pp. 27:1–27:24, 2014.

[10] J. Vosecky, D. Jiang, K. W. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in Twitter," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 879–884.

[11] D. Jiang, K. W. Leung, L. Yang, and W. Ng, "TEII: Topic enhanced inverted index for top-k document retrieval," *Knowl.-Based Syst.*, vol. 89, pp. 346–358, 2015.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, 2004, pp. 487–494.

[13] Y. Tong, C. C. Cao, and L. Chen, "TCS: Efficient topic discovery over crowd-oriented service data," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 861–870.

[14] E. Sadikov, J. Madhavan, L. Wang, and A. Halevy, "Clustering query refinements by user intent," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 841–850.

[15] D. Jiang, J. Vosecky, K. W. Leung, and W. Ng, "G-WSTD: A framework for geographic web search topic discovery," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1143–1152.

[16] D. Jiang and W. Ng, "Mining web search topics with diverse spatiotemporal patterns," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 881–884.

[17] 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.

[18] 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.

[19] C. Elkan, "Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 289–296.

[20] T. Lappas, B. Arai, M. Platakis, D. Kotsakos, and D. Gunopulos, "On burstiness-aware search for document sequences," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2009, pp. 477–486.

[21] I. Sato and H. Nakagawa, "Topic models with power-law using pitman-yor process," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 673–682.

[22] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The query-flow graph: Model and applications," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 609–618.

[23] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal, "Using the wisdom of the crowds for keyword generation," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 61–70.

[24] J. Vosecky, D. Jiang, K. W.-T. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in Twitter," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 879–884.

[25] X. Li, Y.-Y. Wang, D. Shen, and A. Acero, "Learning with click graph for query intent classification," *ACM Trans. Inf. Syst.*, vol. 28, no. 3, 2010, Art. no. 12.

[26] N. Matthijs and F. Radlinski, "Personalizing web search using long term browsing history," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 25–34.

[27] D. Jiang, K. W. Leung, and W. Ng, "Context-aware search personalization with concept preference," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 563–572.

[28] C. M. Carlo, "Markov chain monte carlo and gibbs sampling," *Notes,(April).*, 2004.

[29] L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi, "Topic significance ranking of LDA generative models," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2009, pp. 67–82.

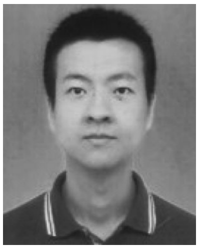
[30] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 247–256.

[31] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 77–86.

[32] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[33] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 424–433. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150450

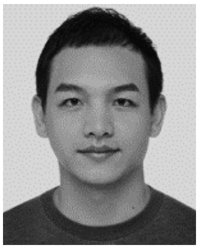
Chen Zhang (Member, IEEE) received the PhD degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2015. He is currently a research assistant professor of the Department of Computing, Hong Kong Polytechnic University, Hong Kong, SAR China. Before joining the Department, he worked as a senior manager of the Big Data Institute at the Hong Kong University of Science and Technology, Hong Kong. He is broadly interested in crowdsourcing, fintech, and machine learning.



Haodi Zhang received the PhD degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently a principal investigator in the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China, and an assistant professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Qifan Li received the bachelor's degree from the College of Mechanical Engineering, Dongguan University of Technology, China, in 2018. He is currently working toward the master's degree in the College of Computer Science and Software Engineering, Shenzhen University and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, China.

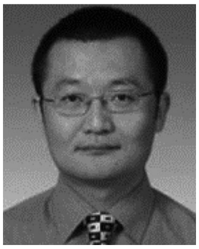
Kaishun Wu (Member, IEEE) received the PhD degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2011. He is currently distinguished professor with the College of Computer Science and Software Engineering, Shenzhen University and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, China.



Jiang Di received the PhD degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a senior scientist at WeBank AI, China. His research interests include information retrieval, natural language processing and massive data management.

Yuanfeng Song received the MPhil degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2012. He is currently a scientist at WeBank AI, China. His research interests include information retrieval, natural language processing, and speech recognition.

Peiguang Lin received the PhD degree from the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, in 2006. He is an associate professor with the School of Computer Science and Technology, Shandong University of Finance and Economics, Shandong, China. His current research interests include massive data processing and integrated, and network information processing.



Lei Chen (Fellow, IEEE) received the PhD degree in computer science from the University of Waterloo, Canada, in 2005. He is currently a professor with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. His research interests include crowdsourcing over social media, social media analysis, probabilistic and uncertain databases, and privacy-preserved data publishing.