Beyond Click Graph: Topic Modeling for Search Engine Query Log Analysis

Di Jiang, Kenneth Wai-Ting Leung, Wilfred Ng, and Hao Li

Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong, China
{dijiang,kwtleung,wilfred,hliaj}@cse.ust.hk

Abstract. Search engine query log is a valuable information source to analyze the users' interests and preferences. In existing work, click graph is intensively utilized to analyze the information in query log. However, click graph is usually plagued by low information coverage, failure of capturing the diverse types of co-occurrence and the incapability of discovering the latent semantics in data. In this paper, we go beyond click graph and analyze query log through the new perspective of probabilistic topic modeling. In order to systematically explore the potential assumptions of the latent structure of the log data, we propose three different topic models. The first model, the *Meta-word Model* (MWM), unifies the co-occurrence of query terms and URLs by the meta-word occurrence. The second model, the *Term-URL Model* (TUM), captures the characteristics of query terms and URLs separately. The third model, the *Clickthrough Model* (CTM), captures the clicking behavior explicitly and models the ternary relation between search queries, query terms and URLs. We evaluate the three proposed models against several strong baselines on a real-life query log. The experimental results show that the proposed models demonstrate significantly improved performance with respect to different quantitative metrics and also in applications such as date prediction, community discovery and URL annotation.

1 Introduction

Search engine query log provides a good window for understanding the users' underlying interests and preferences. Therefore, query log has also been serving as the basis of many functionalities of search engines, such as spelling correction [1], query suggestion [3] and search personalization [16][9]. The majority of existing work on query log analysis is conducted by analyzing click graph, which is essentially a bipartite graph built upon search queries and clicked URLs. While reasonably good performance has been achieved for some tasks, some inherent drawbacks of click graph have not been satisfactorily addressed so far. The following example illustrates the limitations of click graph.

Consider the query log sample in Table 1 and its corresponding click graph in Fig. 1. The following three limitations of click graph can be observed:

W. Meng et al. (Eds.): DASFAA 2013, Part I, LNCS 7825, pp. 209–223, 2013.

Table 1. Search Engine Query Log Sample

ID	User	Query	Clicked URL	Timestamp
q_1	u_1	flash player	get.adobe.com/.../...	2011-04-11 15:12:41
q_2	u_1	adobe		2011-04-12 11:13:44
q_3	u_1	flash	www.adobe.com/.../...	2011-04-12 11:14:21
q_4	u_1	photoshop	www.brothersoft.com/...	2011-04-13 07:13:01
q_5	u_1	flash		2011-04-15 19:13:01
q_6	u_2	blade series	en.wikipedia.org/.../...	2011-04-10 09:44:26
q_7	u_2	superhero	www.superherodb.com/	2011-04-14 14:35:14
q_8	u_2	flash	www.imdb.com/.../...	2011-04-14 14:36:26

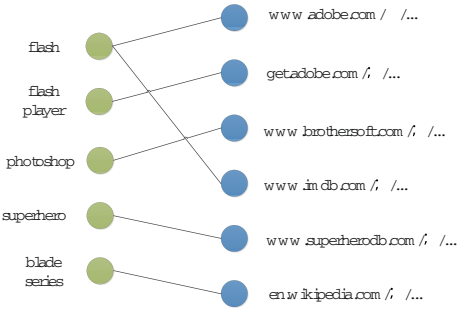


Fig. 1. An Example Click Graph of Table 1

1. *The click graph combines q_3, q_5 and q_8 as a single node "flash". However, these queries are submitted to satisfy different information needs. By manually analyzing the corresponding URLs, we find that u_1 submits q_3 to look for a software product of Adobe while u_2 submits q_8 to search a popular TV series. Due to the polysemy of the query terms, information confusion clearly exists in the example click graph.*
2. *The click graph ignores the user information, which is effective in disambiguating the meaning of queries. If the user information is taken into consideration, we can see that u_1 is interested in IT technology and u_2 is interested in superheroes. Thus, q_3 is more likely to be related to the software product and q_8 is more likely to be about the TV series of the superhero.*
3. *The click graph ignores the information of timestamps and abandoned queries, which are critical for inferring a query's real meaning. Before submitting q_3, u_1 searched q_2 ("adobe") within a minute. Thus, q_3 is likely to be related to "adobe" and to be interpreted as Adobe Flash. Right before q_8, u_2 searched q_7 ("superhero"). Therefore, q_8 is more likely to be related to the TV series of the superhero Flash. Another drawback of ignoring the timestamp is the lack of capturing the web dynamics. For example, if the TV series Flash is a hot topic during the period, then q_5 is also likely to be submitted by u_1 to satisfy the information need about the TV series.*

Due to the advantage of capturing complicated relations in a principled manner, we explore the paradigm of probabilistic topic modeling to tackle with the unsolved problems of click graph. However, this task is not trivial and the challenges are primarily twofold. First, as is shown in Table 1, each log entry contains different types of information. How to integrally utilize all the information to tackle the limitations of click graph is a challenging issue. Second, different from the scenario of document modeling which faces homogeneous *words*, query log is composed of two kinds of heterogeneous items, the query terms and the URLs. Thus, topic modeling on query log needs to handle the heterogeneous items and capture the complicated co-occurrence between them. The two challenges render conventional topic models inapplicable or they can only work suboptimally in the scenario of query log analysis.

To better handle with the aforementioned challenges, we first pre-process the raw query log, making it suitable for topic modeling. Then we propose three probabilistic topic models: the *Meta-word Model* (MWM), the *Term-URL Model* (TUM) and the *Clickthrough Model* (CTM), in order to systematically explore the potential assumptions of the relations between the query terms and URLs. The *Meta-word Model* (MWM) unifies the co-occurrence relations between query terms and URLs as the meta-word co-occurrence, and assumes these meta-words follow the same distribution given a topic. The *Term-URL Model* (TUM) models the clickthrough behavior explicitly and assumes that query terms and URLs follow different distributions given a topic. The *Clickthrough Model* (CTM) introduces the variable of search query and utilizes the ternary relationship between search queries, query terms and URLs. We quantitatively evaluate the proposed models with conventional topic models such as Latent Dirichlet Allocation (LDA) [2] and Topics-Over-Time (TOT) [21]. The proposed models demonstrate significantly better performance in the scenario of query log analysis. Furthermore, we also compare the proposed models with some click graph based approaches in the applications of community discovery and URL annotation. The three models also demonstrate superior performance. The contributions of this paper are summarized as follows:

1. *First, we identify the limitations of click graph and view search engine query log from a new perspective of probabilistic topic modeling.*
2. *Second, we formulate three probabilistic topic models to analyze query log. The three models can effectively integrate multiple types of information in query log and systematically explore different assumptions of the relations between query terms and URLs.*
3. *Third, we carry out extensive evaluations on the three proposed models with a real-life query log. The proposed models demonstrate significantly improved performance compared to several strong baselines with regard to different quantitative metrics. We also validate the usefulness of the models with applications such as date prediction, community discovery and URL annotation.*

The remainder of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we discuss the pre-processing procedure. In Section 4,

we formulate three probability topic models, the *Meta-word Model* (MWM), the *Term-URL Model* (TUM) and the *Clickthough Model* (CTM), in order to discover search topics from query log. The experimental results are presented in Section 5. Some discussion is presented in Section 6. Finally, the paper is concluded in Section 7.

2 Related Work

In recent years, probabilistic data analysis is gaining momentum in data mining [2] [19] [21]. Among them, the topic modeling approach demonstrates superior performance in exploring the latent knowledge of electronic archives. Griffiths *et al.* [4] applied Latent Dirichlet Allocation (LDA) to scientific articles and studied its effectiveness in finding scientific topics. As an extension of LDA, Wang *et al.* [21] presented a topic model that captures both the latent structure of data and how the structure changes over time. There follow more topic models that are proposed to handle the problems of document analysis that exist in specific domains, such as sentiment analysis [11] and geographical analysis [10]. Furthermore, Kang *et al.* [12] proposed a topic-concept cube which supports online multidimensional mining of query log. Mei *et al.* [17] proposed a novel probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously. Some recent work on query log analysis also studied the impact of temporal issues. Ha-Thuc *et al.* [5] proposed an approach for event tracking with emphasis on scalability and selectivity, and their experiments showed that the approach can extract important temporal patterns about the news events. To the best of our knowledge, our work is the first one to systemically explore different assumptions about the relations between query terms and URLs via probabilistic topic modeling. The experimental results show that topic modeling is an effective approach to discover the latent semantics in query log and outperforms several strong baselines with regard to both quantitative metrics and real applications.

3 Pre-processing

In the pre-processing procedure, we first organize the log entries of the ith search engine user as the document d_i and then group the consecutive queries that have semantic relations as *search sessions*. A search session refers to a series of queries which are submitted within a short time period to satisfy the same information need. In order to avoid the performance degradation that is caused by including irrelevant queries in the same session, we prioritize the semantic coherency across queries within the same session. The *query reformulation taxonomy* proposed in [8] consists of a series of rules that evaluate the lexical similarity between queries and demonstrates high precision in detecting semantically relevant search queries. Thus, we utilize it to evaluate the relevancy between two consecutive queries in the log. Finally, we use the stopword list provided in [15] to filter

out the non-informative terms from each search query. The timestamps are normalized to a real number between 0 and 1 based on the earliest and the latest timestamps in query log.

4 Topic Models

4.1 Meta-word Model (MWM)

The *Meta-word Model* (MWM) assumes that each user's query log (i.e., each document) has a unique distribution over a set of K search topics and each of which is represented as a multinomial distribution over all the meta-words in the vocabulary drawn from a symmetric Dirichlet prior β. The meta-words have two potential interpretations:

- The first interpretation only utilizes the query terms. This interpretation simply ignores the URLs and is denoted as MWM-T in the rest of the paper.
- The second interpretation considers both the query terms and the URLs as meta-words. This interpretation is denoted as MWM-TU in the rest of the paper. Note that this interpretation does not explicitly capture the click-through behavior, since the query terms and URLs are utilized without differentiation.

Although we may only use the URLs as the meta-words to derive topics that solely consist of URLs, this option is not included as an interpretation of MWM due to its lack of topic interpretability and the incapability of supporting downstream applications. The generative process of MWM is depicted in Algorithm 1. Each document is generated by first drawing a document-specific mix θ over topic 1 to topic K that is drawn from a symmetric Dirichlet prior α. Since the information within the same session serves the same information need, we assume that a session is relevant to the same search topic. This observation inspires us to use sessions rather than meta-words as the basic unit of topic assignment. Since we assign search topics on a session basis, a session-specific topic z is drawn from θ. Then within the session, some meta-words are drawn from a multinomial distribution based on the topic z. In MWM, the topic assignment of a session is not only subject to the co-occurrence of meta-words but also subject to the timestamps within the session. We utilize the continuous Beta distribution to capture the temporal prominence of each topic. The timestamps within a session are drawn from a Beta distribution ψ_z which is specific to the session topic z. Ultimately, each meta-word w is picked in proportion to how much the enclosing document prefers the topic z and how much the topic prefers the meta-word w. The timestamp is picked in proportion to how much the enclosing document prefers the topic z and how much the topic prefers the timestamp t.

We aim to find an efficient way to compute the joint likelihood of the observed meta-words and timestamps with the hyperparameters:

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

Algorithm 1. Generative Process of MWM

for topic $k \in 1, ..., K$ **do**
  draw a meta-word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
end for
for each document $d \in 1, ..., D$ **do**
  draw d's topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$
  for each session s in d **do**
    choose a topic $z \sim \text{Multinomial}(\theta_d)$
    generate meta-words $w \sim \text{Multinomial}(\phi_z)$
    draw timestamps $t \sim$ from $\text{Beta}(\Psi_z)$
  end for
end for

We will use this joint likelihood to derive efficient updates for the parameters Θ, Φ and Ψ. The right term $P(\mathbf{z}|\alpha) = \int P(\mathbf{z}|\Theta)P(\Theta|\alpha)d\Theta$ is the same as for the standard LDA and this term ultimately contributes the same terms to the full conditional as well as the sampling formula for updating individual topic assignments z_i. Thus, we use the same derivation as in [4]. Using the independence assumptions of the model, we consider the probability of the meta-words and the timestamps. The probability of the meta-words is given as follows:

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

The probability of the timestamps is listed as follows:

$$P(\mathbf{t}|\Psi, \mathbf{z}) = \prod_{d=1}^{D} \prod_{s=1}^{S_d} \prod_{j=1}^{T_{ds}} P(t_{dsj}|\psi_{z_{ds}})^{N_{dst_{dsj}}}. \tag{3}$$

After combining terms, applying Bayes rule and folding terms into the proportionality constant, the conditional probability of the kth topic for the ith session is defined as follows:

$$\begin{aligned} &P(z_i = k|\mathbf{z}_{-\mathbf{i}}, \mathbf{w}, \mathbf{t}, \alpha, \beta, \Psi) \propto \\ &\frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK} + \alpha_{k'})} \frac{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w + N_{iw}))} \\ &\prod_{w=1}^{W} \frac{\Gamma(C_{kw}^{KW} + \beta_w + N_{iw})}{\Gamma(C_{kw}^{KW} + \beta_w)} \prod_{j=1}^{T} \frac{(1-t_j)^{\psi_{k1}-1} t_j^{\psi_{k2}-1}}{B(\psi_{k1}, \psi_{k2})}. \end{aligned} \tag{4}$$

Gibbs sampling [20] is used to estimate the probability that a query belongs to a certain topic. For simplicity and efficiency, we estimate these Beta distribution ψ_z by the method of moments, once per iteration of Gibbs sampling. After each iteration, we update ψ_{k1} and ψ_{k2} for each topic as follows:

$$\psi_{k1} = \bar{t_k}(\frac{\bar{t_k}(1-\bar{t_k})}{s_k^2} - 1), \tag{5}$$

$$\psi_{k2} = (1 - \bar{t_k})(\frac{\bar{t_k}(1 - \bar{t_k})}{s_k^2} - 1), \tag{6}$$

where $\bar{t_k}$ and s_k^2 denote the sample mean and biased sample variance of topic k's timestamps.

4.2 Term-URL Model (TUM)

Since the semantics of the URLs is less ambiguous than query terms [13], the URLs are stronger endorsements than query terms in terms of the topical commonality. Considering them separately can better capture the importance of URL clicking, since the amount of query terms significantly outnumbers that of the URLs. Therefore, we further propose the *Term-URL Model* (TUM) to capture the topical distribution of query terms and URLs separately.

The generative process of TUM is presented in Algorithm 2. Similar to MWM, we constrain that the query terms and URLs in the same session share the same topic. The query term selection process is the same as the meta-word selection process in MWM and the timestamp selection process is the same as that of the MWM. Additionally, each URL is picked in proportion to how much the enclosing document prefers the topic z and how much the topic prefers the URL u. The joint probability of terms, URLs and timestamps is given as follows:

Algorithm 2. Generative Process of TUM

for topic $k \in 1, ..., K$ **do**
  draw a term 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 terms $t \sim$ Multinomial(ϕ_z)
    if $X_s = 1$ **then**
      generate URLs $u \sim$ Multinomial(Ω_z)
    end if
    draw timestamps t from Beta(Ψ_z)
  end for
end for

$$P(\mathbf{w}, \mathbf{t}, \mathbf{u}, \mathbf{z}|\alpha, \beta, \delta, \Psi, \mathbf{X}) = P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{u}|\mathbf{z}, \delta, \mathbf{X})P(\mathbf{t}|\Psi, \mathbf{z})P(\mathbf{z}|\alpha). \tag{7}$$

In Equation (7), the terms $P(\mathbf{z}|\alpha)$, $P(\mathbf{w}|\mathbf{z}, \beta)$ and $P(\mathbf{t}|\Psi, \mathbf{z})$ are the same as those in MWM. However, one issue that results from separate modeling the topical distributions of queries and URLs is that sometimes the session is *abandoned* and no clickthrough is raised. Since these queries are also complementary with

respect to the user's search interests [14], we introduce a variable X to indicate whether there exists clickthrough in the session.

$$P(\mathbf{u}|\mathbf{z},\delta,\mathbf{X}) = \int \prod_{d=1}^{D} \prod_{s=1}^{S_d} \{ \prod_{i=1}^{U_{ds}} P(u_{dsi}|\Omega_{z_{ds}})^{N_{dsu_{dsi}}} \}^{I(X_{ds}=1)} \prod_{z=1}^{K} P(\Omega_{z_{ds}}|\delta) d\Omega. \quad (8)$$

The generative process of TUM is further updated as follows. The user first decides the topic and then selects some query terms according to the chosen topic. For each session, the user needs to decide whether to click on some URLs. If $X = 1$, the user clicks on one or more URLs according to the chosen topic. Again, Gibbs sampling is used to estimate the probability that a query belongs to a certain topic. At each transition step of the Markov chain, the conditional probability of the topic of the queries in the ith session should be differentiated. Using a deduction process similar to MWM, we can obtain the update formulas for TUM. The topic of the queries in the ith session, z_i is drawn according to Equation (9) and the temporal parameters ψ_z are updated after each iteration according to Equations (5) and (6).

$$\begin{aligned} P(z_i = k|\mathbf{z_{-i}},\mathbf{w},\mathbf{t},\mathbf{u},\mathbf{X},\alpha,\beta,\delta,\Psi) \propto \\ \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK} + \alpha_{k'})} \prod_{j=1}^{T} \frac{(1-t_j)^{\psi_{k1}-1} t_j^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})} \\ \frac{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w + N_{iw}))} \prod_{w=1}^{W} \frac{\Gamma(C_{kw}^{KW} + \beta_w + N_{iw})}{\Gamma(C_{kw}^{KW} + \beta_w)} \\ \{ \frac{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u))}{\Gamma(\sum_{u=1}^{U}(C_{ku}^{KU} + \delta_u + N_{iu}))} \prod_{u=1}^{U} \frac{\Gamma(C_{ku}^{KU} + \delta_u + N_{iu})}{\Gamma(C_{ku}^{KU} + \delta_u)} \}^{I(X_i=1)}. \end{aligned} \quad (9)$$

4.3 Clickthrough Model (CTM)

TUM assumes that query terms and URLs have the topical independence, i.e., their generation processes are independent given the topic. A more sophisticated strategy is to assume that the two items are not independent given the topic. We propose CTM to model the dependence between query terms and the URLs through search queries. The generative process of CTM is presented in Algorithm 3. As we assume that search queries, query terms and URLs within a session share the same search topic, we also use search session as the basic unit for topic assignment. Similar to MWM and TUM, a session-specific topic z is drawn from θ. Within the session, some query terms are drawn from a multinomial distribution based on the topic z. These query terms are then composed as search queries. We also use an indicator X to indicate whether there exists clickthrough in a search session. If there exists clickthrough ($X = 1$), the URLs are drawn from a multinomial distribution, which is identified by the selected topic z and the corresponding search query q. The joint likelihood of generating the corpus is as follows:

$$P(\mathbf{w},\mathbf{u},\mathbf{t},\mathbf{q},\mathbf{z}|\alpha,\beta,\delta,\Psi,\mathbf{X}) = P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{u}|\delta,\mathbf{z},\mathbf{q},\mathbf{X})P(\mathbf{q}|\mathbf{w})P(\mathbf{t}|\Psi,\mathbf{z})P(\mathbf{z}|\alpha), \quad (10)$$

Algorithm 3. Generative Process of CTM

for topic $k \in 1, ..., K$ **do**
  draw a term distribution $\phi_k \sim \text{Dirichlet}(\beta)$
  for query $q \in 1, ..., Q$ **do**
    draw a URL distribution $\Omega_{qk} \sim \text{Dirichlet}(\delta)$
  end for
end for
for each meta document $d \in 1, ..., D$ **do**
  draw d's topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$
  for each session s in d **do**
    choose a topic $z \sim \text{Multinomial}(\theta_d)$
    generate terms $t \sim \text{Multinomial}(\phi_z)$
    for each query q in s **do**
      if $X_q = 1$ **then**
        generate URLs $u \sim \text{Multinomial}(\Omega_{qz})$
      end if
    end for
    draw timestamps t from Beta(Ψ_z)
  end for
end for

In CTM, the formula terms $P(\mathbf{z}|\alpha)$, $P(\mathbf{w}|\mathbf{z}, \beta)$ and $P(\mathbf{t}|\Psi, \mathbf{z})$ are the same as those in TUM. $P(\mathbf{q}|\mathbf{w})$ is constant and independent of the search topic. The major difference is that w and u are not independent anymore given the topic. The generation of u subjects to both the topic z and the corresponding search query q.

$$P(\mathbf{u}|\delta, \mathbf{z}, \mathbf{q}, \mathbf{X}) = \int \prod_{d=1}^{D} \prod_{s=1}^{S_d} \{ \prod_{i=1}^{U_{ds}} P(u_{dsi}|\Omega_{q_{u_{dsi}} z_{ds}})^{N_{dsu_{dsi}}} \}^{I(X_{ds}=1)} \prod_{q=1}^{Q} \prod_{z=1}^{K} P(\Omega_{qz_{ds}}|\delta) d\Omega. \tag{11}$$

The conditional probability of the kth topic for the ith session is defined in Equation (12). After each iteration, the temporal parameters ψ_z are updated according to Equations (5) and (6).

$$P(z_i = k|\mathbf{z}_{-\mathbf{i}}, \mathbf{w}, \mathbf{t}, \mathbf{u}, \mathbf{X}, \alpha, \beta, \delta, \Psi) \propto \frac{C_{dk}^{DK} + \alpha_k}{\sum_{k'=1}^{K} C_{dk'}^{DK} + \alpha_{k'}} \prod_{j=1}^{T} \frac{(1-t_j)^{\psi_{k1}-1} t_j^{\psi_{k2}-1}}{B(\psi_{k1}, \psi_{k2})}$$
$$\frac{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW} + \beta_w + N_{iw}))} \prod_{w=1}^{W} \frac{\Gamma(C_{kw}^{KW} + \beta_w + N_{iw})}{\Gamma(C_{kw}^{KW} + \beta_w)}$$
$$\{ \prod_{q=1}^{Q_i} \frac{\Gamma(\sum_{u=1}^{U}(C_{qku}^{QKU} + \delta_u))}{\Gamma(\sum_{u=1}^{U}(C_{qku}^{QKU} + \delta_u + N_{iqu}))} \prod_{u=1}^{U_{iq}} \frac{\Gamma(C_{qku}^{QKU} + \delta_u + N_{iqu})}{\Gamma(C_{qku}^{QKU} + \delta_u)} \}^{I(X_i=1)} \tag{12}$$

5 Experiments

In this section, we present the experimental results. We utilize a real-world query log from a major commercial search engine to conduct the experiments.

The raw query log is pre-processed according to the discussion in Section 3. The dataset records the search history of 2,417 users during 3 months. In Section 5.1, we quantitatively evaluate the proposed models against LDA and TOT by using three metrics: the perplexity of held-out data, the predictive perplexity of partially observed data. In Sections 5.2, 5.3 and 5.4, we demonstrate the effectiveness of the the proposed models in applications such as date prediction, community discovery and URL annotation.

5.1 Quantitative Evaluation

We now evaluate the effectiveness of the proposed models by two quantitative metrics. The pre-processing procedures such as document grouping and stopword preprocessing are the same for all the models under evaluation. We choose the following methods as the baselines:

- LDA-T: Latent Dirichlet Allocation [2] that only utilizes the query terms.
- LDA-TU: Latent Dirichlet Allocation that utilizes both the query terms and URLs as metawords.
- TOT-T: Topics-Over-Time model [21] that only utilizes the query terms.
- TOT-TU: Topics-Over-Time model that utilizes the query terms and URLs as metawords.

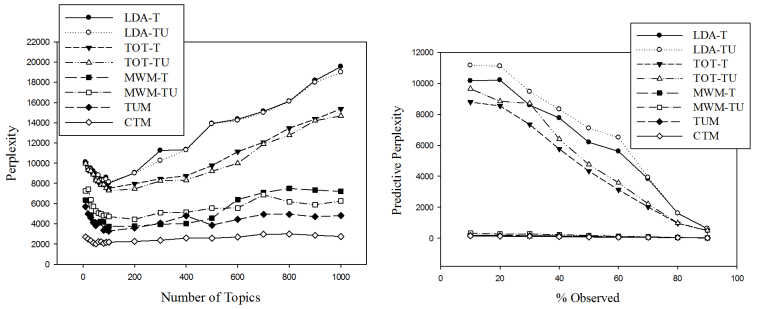
The first metric we use is the perplexity of heldout data. Perplexity is a measure of the ability of a model to generalize to unseen data. Better generalization performance is indicated by a lower perplexity. We compare the proposed models with LDA and TOT by a ten-fold cross validation. We use Equation (13) to calculate the perplexity for each model [18].

$$Perplexity_{heldout}(\mathcal{M}) = (\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(w_i|\mathcal{M}))^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}, \tag{13}$$

where $\mathcal{M}$ is the model learned from the training process.

Figure 2(a) illustrates the average perplexity for each model. MWM, TUM and CTM all provide significantly better fit than LDA and TOT. For example, when the number of topics is set to 100, the average perplexity of LDA-T is 8013, that of MWM-T is 3753, MWM-TU is 4713, TUM is 3287 and CTM achieves the lowest perplexity of 2189. We also observe that the performance of MWM-T and MTM-TU are comparable when the number of topics is small while MWM-TU has better performance than MWM-T when the number of topics is large. The result suggests that incorporating the URL information enables the model to support more topics.

Another metric for comparing the relative strengths of LDA and TOT with our proposed models is how well the models predict the remaining query terms after observing a portion of the user's search history. Suppose we observe the query terms $w_{1:P}$ from a user's query log and aim to find out which model provides a better predictive distribution $p(w|w_{1:P})$ of the remaining query terms. We use Equation 14 to calculate the perplexity of the remaining unseen data.



(a) Perplexity for held-out data (b) Predictive perplexity for partially observed data

Fig. 2. Perplexity Comparison

The results of the comparison are presented in Figure 2(b). We observe that the proposed models significantly outperform LDA. The average perplexity of MWM-T is 119, MWM-TU is 176, TUM is 112 and CTM demonstrates the best performance with an average perplexity of 83.

$$Perplexity_{portion}(\mathcal{M}) = (\prod_{d=1}^{D} \prod_{i=P+1}^{N_d} p(w_i|\mathcal{M}, w_{a:P}))^{\frac{-1}{\sum_{d=1}^{D}(N_d-P)}}. \quad (14)$$

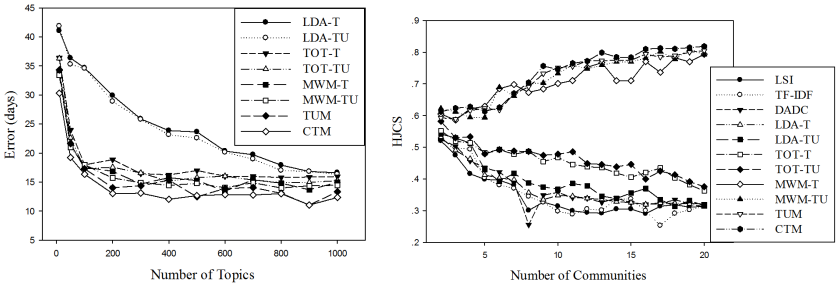
5.2 Data Prediction

We proceed to compare the accuracy of the timestamp prediction of our models given the query terms in a session. We use 6624 held-out search sessions as the evaluation data and then evaluate each model's ability to predict the date of a search session. The Beta distribution for each LDA topic is fitted in a post-hoc fashion. The results of the comparison are presented in Figure 3(a). The average date prediction error of LDA-T is 22.93 days and the average error of LDA-TU is 21.56 days. The average error of MWM-T is 15.14 days, that of MWM-TU is 14.94 days and TUM is 13.87 days and CTM demonstrate the highest date prediction accuracy with an average error of 11.26 days. The above three metrics indicate that the proposed models are better at capturing the temporal trends in web search and thus achieves better performance in date prediction.

5.3 Community Discovery

After processing each user's search history by the proposed models, the ith user's search interests are represented by a topic vector $(\theta_{i1}, \theta_{i2}, ..., \theta_{in})$ where θ_{ik} is a real number that indicates the ith user's endorsement for the kth search topic.

We prepare the ground truth with a small portion of the query log, including 500 users and their 114,400 queries. The queries are manually classified into



(a) Date prediction (b) Community Quality Analysis

Fig. 3. Performance Comparison

21 ODP[1] categories and thus each user is represented by a 21-element vector, where each element is the frequency of the user's queries that belong to the corresponding category. After normalization, the vector serves as the ground truth user profiles.

The first baseline for user profiling is the widely used TF-IDF text representation (denoted as TF-IDF), by which we represent a user's search history by his/her corresponding TF-IDF vector of query terms and URLs. The second baseline for user profiling is Latent Semantic Indexing (denoted as LSI) [7], which is able to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. In the LSI profiling method, we use the extracted "concept" vector as the user profile. The fourth baseline is the DADC algorithm [13], which utilizes a variant of click graph to analyze the relation between users, search queries and concepts.

We then apply K-means clustering algorithm to the ground truth user profiles and those from the baselines and the proposed models. Our objective is to evaluate whether the profiles obtained from the proposed models can generate communities that are closer to that of the ground truth comparing to the baselines. For the purpose of community quality evaluation, we check the resultant relation[2] between each pair of users against the results obtained from the ground truth. For the clustering result c_p that is obtained from the user profiling method p and two users i and j, if the relation of i and j in c_p is consistent with the result c_{truth} that is obtained from the ground truth, then we consider it as a positive judgment. We repeat this process for each pair of users and then normalize the final counting by the total number of user pairs. We call the normalized value the *Human Judgment Correlation Score* (HJCS) and formally define HJSC as follows:

$$HJCS_p = \frac{\sum_{i,j \neq i} 1(c_p(i,j) = c_{truth}(i,j))}{C_n^2}. \tag{15}$$

[1] `http://www.dmoz.org/`

[2] The relation means whether i and j belong to the same cluster or not.

The higher the HJCS, the closer the correlation between c_p and c_{truth}. The results shown in Figure 3 (b) demonstrate the effectiveness of the proposed models. With the increase of the number of communities, the HJCS of TF-IDF, LSI and DADC demonstrate decreasing trends while the HJCS of the proposed models gradually increase. Since high HJCS is trivial when the number of communities is small (e.g., in the extreme case, HJCS value would be all 100% if the number of communities is set to be 1) while high HJCS is challenging when the number of communities is large, the increasing trend of our models illustrates their high degree of correlation with human judgment and suggests that they are effective to discover small but coherent user communities.

5.4 URL Annotation

We now evaluate the quality of URL annotations that are generated by the proposed models. LDA-TU, TOT-TU, MWM-TU, TUM and CTM support the discovery of the semantic relation between query terms and URLs. Within each search topic, the top-ranked query terms can be considered as the annotation of the top-ranked URLs. In order to quantitatively evaluate the quality of URL annotation obtained from the models, we compare them with *M2* and *baseline+M2*, which are two click graph based methods and achieve the best performance in [6], in the task of URL classification. For a specific topic, we select the top 2 URLs and use the top 10 terms as their annotations. In total 500 URLs are selected for this experiment. Other experimental settings are the same as that discussed in [6] and are skipped here to save space. TUM and CTM achieve classification accuracies of 0.6397 and 0.6535, which significantly outperform M2's 0.5124 and baseline+M2's 0.5558. LDA-TU, TOT-TU and MWM-TU achieve accuracies of 0.4979, 0.5213 and 0.5444, respectively. The result suggests that the TUM and CTM are effective to capture the semantic relation between query terms and URLs. Thus, the resultant search topics are effective for interpreting the URL's content with higher accuracy.

6 Discussion

Based on the evaluation in the previous section, we find that query log analysis needs specialized probabilistic topic models due to its unique characteristics. Conventional topic models such as LDA and TOT can only work suboptimally in this task. We also observe that good probabilistic topic models can outperform the click graph (or its variant) based methods in the task of community discovery and URL annotation. The result indicates that utilizing the information missed in click graph can effectively boost the performance of some applications. Among all the three proposed models, the CTM model achieves the best performances in most cases. This result shows that the two heterogenous items, query terms and URLs follow distinct distributions and are strictly coupled via search queries. However, the performance superiority of CTM is gained with a price to pay. The space complexity of MWM is $O(DK + KW)$, where W is the number of

metawords. The space complexity of TUM is $O(DK + KW + KU)$, where W is the number of query terms and U is the number of URLs. The space complexity of CTM is $O(DK + KW + QKU)$, where W is the number of query terms, U is the number of URLs and Q is the number of queries. Therefore, CTM usually consumed more space than MWM and TUM. Thus, when the memory is limited, MWM or TUM can be used as good alternatives of CTM.

7 Conclusion

In this paper, we introduce three new probabilistic topic models, the *Metaword Model* (MWM), the *Term-URL Model* (TUM) and the *Clickthrough Model* (CTM), in order to analyze search engine query log. The three models explore different assumptions to discover search topics from the users' search history. Parameter inference approaches such as Gibbs sampling are further introduced to estimate the value of latent variables. Our findings demonstrate that probabilistic topic modeling has the advantage of seamlessly integrating different types of information in query log and effectively capturing the latent semantics. Empirical evaluations on a real-life query log unequivocally demonstrate the superiority of the proposed models and their utilities in different applications.

Acknowledgments. This work is partially supported by GRF under grant numbers HKUST 617610 and 618509. We also wish to thank the anonymous reviewers for their comments.

References

1. Ahmad, F., Kondrak, G.: Learning a spelling error model from search query logs. In: Proc. of the HLT- EMNLP Conference (2005)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research (2003)
3. Deng, H., King, I., Lyu, M.R.: Entropy-biased models for query representation on the click graph. In: Proc. of the ACM SIGIR Conference (2009)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. of the National Academy of Sciences of the United States of America (2004)
5. Ha-Thuc, V., Mejova, Y., Harris, C., Srinivasan, P.: A relevance-based topic model for news event tracking. In: Proc. of the ACM SIGIR Conference (2009)
6. Hinne, M., Kraaij, W., Raaijmakers, S., Verberne, S., van der Weide, T., Van Der Heijden, M.: Annotation of urls: more than the sum of parts. In: Proceedings of the ACM SIGIR Conference (2009)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of the ACM SIGIR Conference (1999)
8. Huang, J., Efthimiadis, E.N.: Analyzing and evaluating query reformulation strategies in web search logs. In: Proc. of the ACM CIKM Conference (2009)
9. Jiang, D., Leung, K.W.T., Ng, W.: Context-aware search personalization with concept preference. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management

10. Jiang, D., Vosecky, J., Leung, K.W.T., Ng, W.: G-wstd: A framework for geographic web search topic discovery. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management
11. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proc. of the Fourth ACM WSDM Conference (2011)
12. Kang, D., Jiang, D., Pei, J., Liao, Z., Sun, X., Choi, H.J.: Multidimensional mining of large-scale search logs: a topic-concept cube approach. In: Proc. of the ACM WSDM Conference (2011)
13. Leung, K.W.-T., Lee, D.L.: Dynamic agglomerative-divisive clustering of clickthrough data for collaborative web search. In: Kitagawa, H., Ishikawa, Y., Li, Q., Watanabe, C. (eds.) DASFAA 2010. LNCS, vol. 5981, pp. 635–642. Springer, Heidelberg (2010)
14. Li, J., Huffman, S., Tokuda, A.: Good abandonment in mobile and pc internet search. In: Proc. of the ACM SIGIR Conference (2009)
15. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to information retrieval. Cambridge University Press, Cambridge (2008)
16. Matthijs, N., Radlinski, F.: Personalizing web search using long term browsing history. In: Proc. of the ACM WSDM Conference (2011)
17. Mei, Q., Liu, C., Su, H., Zhai, C.X.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proc. of the WWW Conference (2006)
18. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proc. of the UAI Conference (2004)
19. Tong, Y., Chen, L., Ding, B.: Discovering threshold-based frequent closed itemsets over probabilistic data. In: IEEE 28th International Conference on Data Engineering (2012)
20. Walsh, B.: Markov chain monte carlo and gibbs sampling (2004)
21. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proc. of the ACM SIGKDD Conference (2006)