Mining Mobile Users' Activities Based on Search Query Text and Context

Bingyue Peng[1,⋆], Yujing Wang[2,*], and Jian-Tao Sun[3]

[1] Beihang University, Beijing 100191, China
[2] Key Laboratory of Machine Perception, Peking University, Beijing 100871, China
[3] Microsoft Research Asia, Beijing 100080, China

Abstract. Mobile search market is growing very fast. Mining mobile search activities is helpful for understanding user preference, interest and even regular patterns. In previous works, text information contained by either search queries or web pages visited by users is well studied to mine search activities. Since rich context information (e.g., time, location and other sensor inputs) is contained in the mobile search data, it has also been leveraged by researchers for mining user activities. However, the two types of information were used separately. In this paper, we propose a graphical model approach, namely the Text and Context-based User Activity Model (TCUAM), which mines user activity patterns by utilizing query text and context simultaneously. The model is developed based on Latent Dirichlet Allocation (LDA) by regarding users' activities as latent topics. In order to guide the activity mining process, we borrow some external knowledge of topic-word relationship to build a constrained TCUAM model. The experimental results indicate that the TCUAM model yields better results compared with text-only and context-only approaches. We also find that the constrained TCUAM model is more effective than the unconstrained TCUAM model.

Keywords: mobile user modeling, user's activity mining, Latent Dirichlet Allocation.

1 Introduction

With the prosperity of mobile market, more and more web search activities go to mobile devices. This raises the requirement of mining mobile search data, which is important for understanding user preferences, interests and activity patterns. Compared with web search from PC, mobile search data contains rich context information, e.g., time, location, surrounding business and other signals captured by sensors of mobile devices. Previous works of mining search activities focus on analyzing the content of search query, web pages, etc., with limited attentions of mining context information [3]. According to our analysis of mobile search log

⋆ This work was done when the first two authors conducted internship at Microsoft Research Asia.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 109–120, 2012.

Table 1. Three main types of user search activity data

User Behavior	Query Text	Search Context
Text-Dominated	**restaurant**	Time = 08:00∼09:00 Day = Monday SurroundingType = Amusement Equipment
Context-Dominated	facebook	**Time = 14:00∼ 15:00** **Day = Sunday** **SurroundingType = Baseball Clubs & Parks**
Both-Dependent	**Samsung focus price Amazon.com**	**Time = 15:00∼ 16:00** **Day = Saturday** **SurroundingType = Downtown**

data, we discover that both search query text and context information can help understand user activities. Table 1 gives examples of three major types of user search activities.

- **Text-Dominated** activities can be fully understood by query content, without considering context information. For example, query "restaurant" indicates that the user wants to find a restaurant.
- **Context-Dominated** activities can be explained by the context information. E.g., a user issues several queries with the following context: "Time = 14:00∼15:00", "Day = Sunday" and "SurroundingType = Baseball Clubs & Parks". We can infer that the user's activity may be related to "Playing Baseball".
- **Both-Dependent** activities require both text and context information to explain the user's activities. For instance, the user's context is "Time = 15:00∼16:00", "Day = Saturday", "SurroundingType = Downtown", and the query is "Samsung focus price Amazon.com". We can infer that the user's activity is likely to be "Shopping".

We can see that both text and context information can help understand the activity of mobile users. However, as far as we know, currently there are few approaches which can model user activities based on text and context information simultaneously.

In this paper, we propose a graphical model approach, namely the Text and Context-based User Activity Model (TCUAM) to mine user activity patterns using both query text and search context information. The TCUAM model is developed based on Latent Dirichlet Allocation (LDA), by regarding user activities as latent topics. As there are many noises in mobile log data, TCUAM has difficulty in discovering meaningful patterns. Therefore, we leverage human knowledge to help. We borrow external knowledge of topic-word relationship to build a constrained TCUAM model. The experiments on real mobile log indicates that the TCUAM model yields better results compared with text-only and context-only approaches. We also find that the constrained TCUAM model behaves more effectively than the unconstrained TCUAM model.

The rest of this paper is organized as follows. Section 2 briefly introduces the related work. The TCUAM model is defined in Section 3. We describe experiments and results in Section 4. Conclusions are given in Section 5.

2 Related Work

There are mainly two groups of research works which are related to ours. The first is about feature modeling used in mining user activity patterns. Understanding user intent from text information such as past queries and user profiles is a common technique for web search personalization. Sieg *et al.* [2] analyzed user profiles and assigned implicitly derived interest scores to existing concepts in a domain ontology for personalization. Noll *et al.* [4] implemented personalization using social bookmarking and tagging. Teevan *et al.* [5] utilized a personalization technique to leverage implicit information about the users' interests and activities, including previously issued queries, previously visited web pages and the documents a user has read or created. Besides text information mentioned above, context information is also adopted by researchers to mine user activity patterns. Arias *et al.* [6] found that it was beneficial to understand user intents and complete desired queries by context information such as time and location. Hattori *et al.* [7] improved the performance of query refinement by incorporating user context information. Church *et al.* [8] proposed a novel interface to support multi-dimensional and context-sensitive mobile search, combining context features such as location, time, and community preferences to offer better search experiences.

The other group of related works is about learning models. The models of utilizing context information can be divided into three stages. In the first stage, context information is manually processed in a certain domain, especially in a geographical system [9,10,11]. In the second stage, researchers begin to use traditional text learning model to tackle context learning problems. Algorithms like Bayesian Network, Hidden Markov Model (HMM), Support Vector Machine (SVM) and Conditional Random Field (CRF) have been adopted to model user behaviors [12,13]. In the third stage, unsupervised models are used for learning tasks of large-scale data. Topic model related approaches are adopted in this stage for user activity mining. E.g., Bao *et al.* [3] tried to model context information by an unsupervised approach based on latent dirichlet allocation (LDA) to discover users' activities.

3 Methodology

3.1 Data and Preprocessing

In this work, we mine user activities using mobile search log. The log contains a set of records $R = \{r_1, r_2 ... r_n\}$, where $r_i =< q_i, c_i >$. q_i is the query issued by the

user and c_i is the context when the search event happens. $c_i = \{< f_i, v_i > | 1 \leq i \leq N_p\}$ where $< f_i, v_i >$ is a feature-value pair, f_i stands for the feature name while v_i is the value of feature f_i. As we know, there are various noises in search log and it is difficult to obtain satisfactory results without data preprocessing. In traditional approaches, queries with low frequencies are usually regarded as noises and excluded from query log. In our work, we regard the queries which are not very related to context as noises. We introduce a scoring function to calculate the possibility of a query to be noise:

$$\xi(q) = -\frac{\sum_{i=1}^{N_p} p(< f_i, v_i >) log(p(< f_i, v_i >))}{|f_q - \tilde{f_{75\%}}|} \tag{1}$$

where $p(< f_i, v_i >)$ stands for the probability of feature f_i to take the value v_i. The smaller value $-\sum_{i=1}^{N_f} p(< f_i, v_i >) log(p(< f_i, v_i >))$ takes, the more irrelevant feature f_i is with the query text, thus the more likely query q will be a noisy query. f_q denotes the frequency of query q, $\tilde{f_{75\%}}$ stands for the average frequency of 75% queries whose values lie in the middle of all the queries. The larger value $|f_q - \tilde{f_{75\%}}|$ takes, the more extreme the query frequency is, thus the more likely the query is considered to be noise.

The smaller value $\xi(q)$ takes, the more likely that query q is considered as noise. In practice, an appropriate threshold is chosen to filter out noisy queries.

3.2 Text and Context-Based User Activity Model

In this section, we will introduce the Text and Context-based User Activity Model (TCUAM) for mining users' activities based on Latent Dirichlet Allocation (LDA). Bao *et al.* [3] has proposed an LDA-based approach to mine user activities using context information. However, context information itself can only explain part of user activities. Our model is designed to utilize text and context information simultaneously for user activity mining.

Given a set of records $R = \{r_1, r_2 ... r_n\}$, we split them into sessions according to time information. Each session contains data records within 30-minutes time span. Given a collection of M sessions $S = \{s_1, s_2 ..., s_M\}$, we assume that each session is generated by a collection of topics, which follow dirichlet distributions. Suppose there are totally K topics, $r_{m,n} =< q_{m,n}, c_{m,n} >$ denotes the n^{th} observation of record in the m^{th} session. $q_{m,n} = \{w_{m,n,1}, w_{m,n,2}, ...\}$ stands for the n^{th} observation of query text in the m^{th} session where $w_{m,n,i}$ is the i^{th} word of query $q_{m,n}$, $c_{m,n} = \{< f_i, v_i >\}$ stands for the n^{th} observation of context in the m^{th} session. $< f_i, v_i >$ represents a feature-value pair where f_i denotes the feature name and v_i denotes the value of feature f_i.

The process of generating text and context information for all the sessions can be expressed as follows. Firstly, draw a query word distribution φ_k for each topic k from dirichlet distribution with parameter β. Secondly, for each topic k and feature f, draw a feature-value pair distribution $\omega_{k \cdot f}$ from dirichlet

distribution with parameter τ. Thirdly, for each session s_m, draw a topic distribution θ_m and a feature distributions λ_m from dirichlet distribution with parameter α and γ respectively. Then for each session s_m, records are generated repeatedly based on the model. For each record $r_{m,n}$ in session s_m, we first choose a topic $z_{m,n}$ according to the topic distribution Multi(θ_m). Afterwards, query text and context information are generated respectively according to their distributions on topic $z_{m,n}$. Each word $w_{m,n,i}$ in the query text is generated directly from word distribution $Multi(\varphi_{z_{i,j}})$. For the context $c_{m,n} = \{< f_i, v_i >\}$, each feature f_i is generated from the feature distribution $Multi(\lambda_m)$ while the corresponding value v_i is generated from the feature-value pair distribution $Multi(\omega_{z_{i,j},f_i})$. Table 2 summarizes the generative process of TCUAM.

In practice, we take the whole query as a single word. That is, we use $q_{m,n} = w_{m,n}$, instead of $q_{m,n} = \{w_{m,n,1}, w_{m,n,2}, ...\}$. The graphical representation of the model is shown in Figure 1.

Table 2. Generative process of TCUAM

1. For each topic k
   Draw word distribution $\varphi_k \sim Dir(\beta)$
2. For each topic k and feature f
   Draw feature-value pair distribution $\omega_{k \cdot f} \sim Dir(\tau)$
3. For each session s_m
   (a) Draw topic distribution $\theta_m \sim Dir(\alpha)$
   (b) Draw feature distribution $\lambda_m \sim Dir(\gamma)$
   (c) For each record observation $r_{m,n}$ in session s_m
      (1) Choose a topic $z_{m,n} \sim$ Multi(θ_m)
      (2) Generate *query text* $q_{m,n}$:
         For each word in $q_{m,n}$
            Choose a word $w_{m,n,i} \sim$ Multi($\varphi_{z_{i,j}}$)
      (3) Generate *context information* $c_{m,n}$:
         For each feature value pair $< f_i, v_i >$ in $c_{m,n}$,
            (i) Choose a feature $f_i \sim$ Multi($\lambda_{m,z_{i,j}}$)
            (ii) Choose a feature value $v_i \sim$ Multi($\omega_{z_{i,j},f_i}$)

3.3 Inference of Model

To simplify the equations, we define the following symbols. The hyper-parameters in the TCUAM model are denoted as Θ, which include α, β, γ and τ. The observations are denoted by Γ, which consist of N_s sessions. $r_{m,n} =< q_{m,n}, c_{m,n} >$ denotes the n^{th} record in the m^{th} session. $q_{m,n} = w_{m,n}$ stands for the n^{th} observation of query text in the m^{th} session, and $c_{m,n} = \{< f_i, v_i > | 1 \leq i \leq N_p\}$ stands for the n^{th} observation of context in the m^{th} session. $< f_i, v_i >$ represents a feature-value pair where f_i denotes the feature name and v_i denotes the value of feature f_i. The parameters are represented by Δ, including θ, φ, λ and ω. The latent variables of topics are denoted by z. We define $\underline{\Phi} = \{\varphi_k\}_{k=1}^{K}$, $\underline{\Lambda} = \{\lambda_k\}_{k=1}^{K}$ and $\underline{\Omega} = \{\omega_p\}_{p=1}^{K \cdot F}$, where K is the total number of topics and F is the total

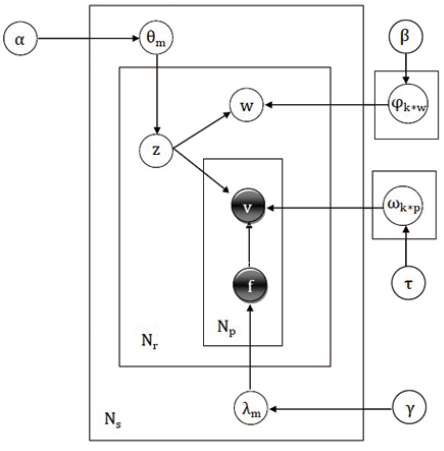

Fig. 1. Graphical Representation of TCUAM

number of features. Thus, given hyper-parameters, the joint distribution of all observations and hidden variables can be calculated as follows:

$$p(\Gamma, \Delta, z|\Theta) = \prod_{m=1}^{N_s} \prod_{n=1}^{N_r} p(w_{m,n}|\varphi_{z_{m,n}}) \prod_{j=1}^{N_p} p(v_p|\omega_{z_{m,n},f_j}) \\ \times p(f_p|\lambda_{m,z_{m,n}}) p(z_{m,n}|\theta_m) p(\theta_m|\alpha) p(\underline{\Phi}|\beta) p(\underline{\Lambda}|\gamma) p(\underline{\Omega}|\tau) \tag{2}$$

where N_s is the number of sessions, N_r is the number of records in each session, and N_p is the number of feature-value pairs in the context.

We obtain the joint probability for all observations by integrating over the parameters and latent variables:

$$p(\Gamma|\Theta) = \int\int\int\int p(\theta_m|\alpha) p(\underline{\Phi}|\beta) p(\underline{\Lambda}|\gamma) p(\underline{\Omega}|\tau) \\ \times \prod_{m=1}^{N_s} \prod_{n=1}^{N_r} \sum_{z_{m,n}} (w_{m,n}|\varphi_{z_{m,n}}) \prod_{j=1}^{N_p} p(v_p|\omega_{z_{m,n},f_p}) \\ \times p(f_p|\lambda_{m,z_{m,n}}) p(z_{m,n}|\theta_m) d\underline{\Phi} d\underline{\Lambda} d\underline{\Omega} d\theta_m \tag{3}$$

We use Gibbs sampling to get the approximate estimation of parameters. In Gibbs sampling, each record is assigned to a certain topic under the condition that other records have been labeled. We assume that *Text Information* and *Context Information* are generated by activity topics independently and obtain the following equation:

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}, \boldsymbol{c}) = p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w},) p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{c}) \tag{4}$$

where $\boldsymbol{w}$ stands for the vector of words; $\boldsymbol{c}$ stands for the vector of feature-value pairs in the context; z_i represents the topic of the i^{th} record whereas $\boldsymbol{z}_{\neg i}$ denotes the vector of topics for all records after excluding the i^{th} record.

The two conditional probabilities on the right side of Eq.(4) can be calculated by [14]:

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}) = \frac{p(\boldsymbol{w}|\boldsymbol{z})}{p(\boldsymbol{w}|\boldsymbol{z}_{\neg i})} \cdot \frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\neg \boldsymbol{i}})} \propto \frac{n^w_{k,\neg i} + \beta_w}{\Sigma^W_{w'=1} n^{w'}_k + \beta_w} \cdot \frac{n^k_{m,\neg i} + \alpha_k}{\Sigma^K_{k'=1} n^{k'}_m + \alpha_k} \tag{5}$$

$$\begin{aligned} p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{c}) &\propto p(v_p|z_i = k, \boldsymbol{z}_{\neg i}, F, V_{\neg i}) p(z_i = k|\boldsymbol{z}_{\neg i}) \\ &= \prod_p^{N_p} \frac{n^{f_p,v_p}_{k,\neg i} + \omega_{v_p}}{\sum_{v'} n^{f_p,v'}_{k,\neg i} + \sum_{v' \epsilon V_{f_p}} \omega_{v'}} \cdot \frac{n^k_{m,\neg i} + \alpha_k}{\sum^K_{k'=1} n^{k'}_{m,\neg i} + \sum^K_{k'=1} \alpha_k} \end{aligned} \tag{6}$$

Where α_k denotes the hyper-parameter of dirichlet distribution for topic k, and β_w denotes the hyper-parameter of dirichlet distribution for word w. $i =< m, n >$ stands for the index of the n^{th} observation in the m^{th} session, $n^w_{k,\neg i}$ stands for the times of word w being observed with topic k after excluding the i^{th} record, $n^{f,v}_{k,\neg i}$ stands for the times of feature-value pair $< f, v >$ being observed with topic k after excluding the i^{th} record, and $n^k_{m,\neg i}$ stands for the times of topic k being observed in session m after excluding the i^{th} record.

After the convergence of Gibbs sampling iteration, each observation will be assigned a final topic label. Eventually, the parameters can be inferred as below:

$$p(w|z_k) = \varphi_{k,w} = \frac{n^w_k + \beta_w}{\Sigma^W_{w'=1} n^{w'}_k + \beta_w} \tag{7}$$

$$p(f_p, v_p|z_k) = p(v_p|z_k, f_p) p(f_p) \tag{8}$$

$$p(v_p|z_k, f_p) = \frac{n^{f_p,v_p}_k + \omega_{v_p}}{\sum_{v'} n^{f_p,v'}_k + n^{f_p,v_p}_k + \omega_{v' \epsilon V_{f_p}} \omega_{v'}} \tag{9}$$

$$p(f_p) = \frac{\sum^K_{k'=1} \sum_{v'} n^{f_p,v'}_{k'} + \lambda_{f_p}}{\sum_{f'} \sum^K_{k'=1} \sum_{v'} n^{f',v'}_{k'} + \sum_{f'} \lambda_{f'}} \tag{10}$$

3.4 Constrained TCUAM Model

In practice, the unconstrained TCUAM model is unable to achieve satisfactory results due to massive noises in mobile log. Therefore, we borrow some external knowledge about topic-word relationship to help. We leverage a set of websites, which are organized into a list of topics. For each topic, we can use search log to associate queries with websites using the follow-click information. Given a set of topics $K = \{k_1, k_2...\}$, assume that the set of websites for topic k_i is $url(k_i) = \{u_1, u_2...\}$. Suppose there is a set of queries $Q = \{q_1, q_2...\}$ and the

set of follow-click URLs for each query q_j is $fclick(q_j) = \{u_1, u_2...\}$. We split each query into word sequences. Thus, the relevant score of topic k_i and word w_j can be calculated by:

$$Score(k_i, w_j) = \sum_{w_j \epsilon q_t, u' \epsilon fclick(q_t), u' \epsilon url(k_i)} tfidf_{w_j} \tag{11}$$

We choose top 50 words with the highest relevance scores for each topic and map them to $\eta[50 \sim 1]$ linearly. For example, if *airline* is the third relevant word to topic *Travel*, $\eta(Travel, airline) = 48$. Thus, we obtain 50 representative words for each topic as external knowledge to guide the activity mining model. The core idea of using this knowledge is to increase the weight of an unlabeled word in Gibbs Sampling if it is known to be a representative word for a specific topic. The whole procedure of Gibbs sampling is displayed in Algorithm 1, where the variables are defined in Table 3.

Table 3. Variables in Algorithm 1

N_s	number of sessions to generate
N_r	number of records in a certain session
N_p	number of feature-value pairs in a certain record
η	relevance scores between words and topics
n_k^w	the times of word w being observed with topic k
n_k^p	the times of feature-value pair p being observed with topic k
n_k^f	the times of feature f being observed with topic k
n_m^k	the times of records being observed with topic k in document m
n_m	the times of records being observed in session m
n_k	the times of records being observed with topic k
n_p	the times of records being observed with feature-value pair p

4 Experiment

4.1 Data Set

In this paper, we carry out our experiments on real mobile logs from a commercial search engine. The data set consists of half a year's mobile logs in California State, USA. Table 4 shows the feature list of text and context information used in our experiment. For *period* feature, we define its values according to the time range of search activity. We remove the users whose query numbers are less than 50 in half a year time span. The preprocessing procedure described in section 3.1 is applied to clean the dataset. In our experiment, we set the threshold as 0.25.

The external knowledge of topic-word relationship for the constrained TCUAM model is illustrated in Table 5. We have 15 kinds of activity topics which mobile users are specially interested in. For each activity topic, 50 words are selected to be the external knowledge. Because of space limitation, only top 5 words for each activity topic are listed in the table.

Algorithm 1. Gibbs Sampling For Constrained TCUAM

zero all counter, $n_m^k, n_k^w, n_k^p, n_k^f, n_m, n_k, n_p$
for *each session* $s_m \epsilon [1, N_s]$ **do**
  for *each record* $r_{m,n} \epsilon [1, N_r]$ *in session* s_m **do**
    if *pre-word* $\tilde{k}$ *exist* **then**
      $r_{m,n}.topic = \tilde{k}$
    end
    else
      $r_{m,n}.topic = z_{m,n} \sim Mult(1/K)$
    end
    $k = r_{m,n}.topic$
    $\mathbf{S} = \eta(k, r_{m,n}.w)$
    Increase counter: $n_m^k + \mathbf{S}$, $n_m + \mathbf{S}$, $n_k^w + \mathbf{S}$, $n_k + \mathbf{S}$
    for *each feature-value pair* $p \epsilon [1, N_p]$ *in record* $r_{m,n}$ **do**
      Increase counter: $n_k^p + \mathbf{S}$, $n_k^f + \mathbf{S}$, $n_p + \mathbf{S}$
    end
  end
end

while *not converged* **do**
  for *each session* $s_m \epsilon [1, N_s]$ **do**
    for *each record* $r_{m,n} \epsilon [1, N_r]$ *in session* s_m **do**
      for the current record $r_{m,n}$ assigned to topic k: $\mathbf{S} = \eta(k, r_{m,n}.w)$
      Decrease counter: $n_m^k - \mathbf{S}$, $n_m - \mathbf{S}$, $n_k^w - \mathbf{S}$, $n_k - \mathbf{S}$
      $\diamond$ sample a new topic $k' \sim p(z_i = k' | \boldsymbol{z}_{\neg i}, \boldsymbol{w}, \boldsymbol{c})$, $\mathbf{S} = \eta(k', r_{m,n}.w)$
      Increase counter: $n_m^{k'} + \mathbf{S}$, $n_m + \mathbf{S}$, $n_{k'}^w + \mathbf{S}$, $n_{k'} + \mathbf{S}$
      for *each feature-value pair* $p \epsilon [1, N_p]$ *in record* $r_{m,n}$ **do**
        Increase counter: $n_{k'}^p + \mathbf{S}$, $n_{k'}^f + \mathbf{S}$, $n_p + \mathbf{S}$
      end
    end
  end
end

4.2 Experimental Setup

In the experiment, we evaluate four models which are described below.

- **TM** (Text-based Model) is the baseline of our experiment. The text-based model utilizes query text to build a LDA model for mining users' activities.
- **CM** (Context-based Model) is proposed by Bao *et al.* [3] to mine mobile users' activity patterns based on context information.
- **TCUAM** (Text and Context-based User Activity Model) is an unconstrained approach proposed in this paper to model users' activities by using text and context information collaboratively.
- **CTCUAM** (Constrained TCUAM) is a constrained model which uses external knowledge of topic-word relationship to guide the TCUAM model.

In order to get a fair comparison of the models above, we adopt the same session segmentation method for all the models. Records are segmented into sessions by time information. Each session contains the records within a time span of 30 minutes. In addition, the number of topics in all the models is set to be 200 experimentally.

Table 4. Feature Information

Data type	Feature	Feature-Value Range
Text Information	N/A	free query, pre-assigned query
Time Information	Date	05/01/2010, 05/02/2010, 05/03/2010... 12/31/2010
	Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
	WorkdayOrWeekend	Workday, Weekend
	Period	Early_Morning, Morning, Noon, Afternoon, Evening Night1, Night2
	Time	$01:00 \sim 02:00$, $02:00 \sim 03:00$ ... $23:00 \sim 00:00$
Location Information	GPS	Longitude and Latitude
	CityName	Glendale, San Diego, Rosemead, Los Angeles, Dublin...
	PostalCode	92128, 92880, 91361, 92107, 94804, 91737 ...
	SurroundingType	None, Colleges & Universities, Natural Gas Services...
	PlaceType	Home, Workplace, Other

4.3 Evaluation

The goal of our experiment is to examine whether users' activity patterns can be mined correctly from mobile logs. Unfortunately, it is difficult to identify automatically whether the result patterns make sense or not. Therefore, we examine the result topics produced by each model manually and assign each topic a score. The score is given according to the following rules:

- **5**: It is a perfect pattern and indicates user activity clearly.
- **4**: It is a good pattern and can give an overall sense of user activity.
- **3**: It is a reasonable pattern and gives some clues of the user activity.
- **2**: It is a bad pattern and includes many noises.
- **1**: It is a non-sense pattern and difficult to be understood.

The average score (AS) of result topics is calculated after each topic is assigned a score manually. In our experiment, we use AS as the metric to evaluate the performances of different models.

4.4 Results

Table 6 shows the results of different models, evaluated by the average score (AS). We can find out that the worst way to mine user's activity is the Context-based Model (CM), whose AS value is 1.995. It indicates that only context information is not enough to determine users' actual activities. The Text-based Model (TM) achieves 2.295 for AS value, which shows that the text information is more informative than the context information. By using text and context information simultaneously, the TCUAM model achieves 2.545 for AS value (improving 27.6% from Context-based Model and 10.9% from Text-based Model). Therefore, the text information and context information can be utilized collaboratively to benefit the activity mining approaches. The Constrained TCUAM model enhances the performance further by 11.8%, achieving an AS value of 2.845. Moreover, the knowledge used in the constrained model is easy to be collected. Thus, the constrained model can mine user activity topics more precisely without taking too much human efforts.

Table 5. Knowledge of Text Information

Shopping	Legal & Finance	Education	Travel
tanger	bank	middle	airline
store	union	school	hotel
deb	stock	university	airtran
outlet	credit	college	cathay
hollister	financial	institution	hyatt
Arts & Entertainment	Automotive & Vehicles	Business to Business	Home & Family
cinemark	toyota	store	badcock
cinema	ford	hollister	arien
theater	tire	suntrust	dyson
imax	honda	levi	home
krikorian	dodge	graco	rug
Government	Sports & Recreation	Health & Beauty	Food & Dining
library	yankee	hospital	restaurant
court	coach	doctor	steakhouse
ccap	dodger	medical	pizza
park	sporting	alzheimer	burger
civil	dunham	sentara	chili
Professionals & Services	Computers & Technology	Real Estate & Construction	
oregonian	sony	apartment	
kinko	garmin	region	
fimserve	safelink	hotel	
kroger	logmein	blum	
train	gps	comerica	

Table 6. Comparison of Models

Model	**5**	**4**	**3**	**2**	**1**	**Sum**	**AS**
TM	2	16	58	87	37	200	2.295
CM	3	11	31	92	63	200	1.995
TCUAM	7	29	57	80	27	200	2.545
CTCUAM	10	42	74	55	19	200	2.845

Table 7. Examples of activity topics

Text Information	IsRelevant	Text Information	IsRelevant
f stock	Yes	cocktail lounges	Yes
ewbc stock	Yes	sports bars	Yes
culos de caseras	Yes	night clubs	Yes
caty stock	Yes	restaurants	No
twitter search	No	carnivals	Yes
coh stock	Yes	amusement places	Yes
cellufun	No	fairgrounds	Yes
games	No	taverns	Yes
monster tits	No	norwalk amc	No
dis stock	Yes	barbecue restaurants	No
Context Information	IsRelevant	Context Information	IsRelevant
WorkdayOrWeekend=Workday	Yes	Period=Evening	Yes
PlaceType=Home	Yes	WorkdayOrWeekend=Weekend	Yes
Day=Wendensday	Yes	Day=Saturday	Yes
Period=Morning	Yes	Day=Tuesday	No
Period=Early_Morning	Yes	PlaceType=Other	Yes
Day=Tuesday	Yes	SurroundingType=Food & Dining	Yes
SurroundingType=None	No	Period=Afternoon	Yes
Time= $07:00 \sim 08:00$	Yes	Time= $17:00 \sim 18:00$	Yes
Time= $06:00 \sim 07:00$	Yes	Time= $19:00 \sim 20:00$	Yes
CityName=Glendale	No	Time= $18:00 \sim 19:00$	Yes

4.5 Case Study

To get a further understanding of the Constrained TCUAM model, we select some examples to demonstrate the results produced by the model. Table 7 shows

two topics discovered by the model. The "IsRelevant" column gives the human judgement that whether the text or context is relevant to the topic. It is easy to infer that the user's activity is *"searching for stock information at home in the workday morning"* for the left case and it is *"searching for amusement place after dinner outside in the weekend"* for the right case.

5 Conclusion

In this paper, we propose a text and context-based user activity model to mine user's activity patterns from mobile logs. In addition, we introduce a small amount of external knowledge about topic-word relationship to build a constrained TCUAM model. The experiments were carried out on real mobile logs. The experimental results have indicated that the TCUAM model can yield better results, compared with text-only and context-only approaches. We can also conclude from the results that the constrained TCUAM model performs more effectively than the unconstrained TCUAM model.

References

1. Wagner, M., Balke, W.-T., Hirschfeld, R., Kellerer, W.: A Roadmap to Advanced Personalization of Mobile Services. In: Proceedings of the DOA/ODBASE/CoopIS, Industry Program (2002)
2. Sieg, A., Mobasher, B., Burke, R.: Web Search Personalization with Ontological User Profiles. In: Proceedings of CIKM 2007 (2007)
3. Bao, T., Cao, H., Chen, E., Tian, J., Xiong, H.: An Unsupervised Approach to Modeling Personalized Contexts of Mobile Users. In: Proceedings of ICDM 2010 (2010)
4. Noll, M.G., Meinel, C.: Web Search Personalization Via Social Bookmarking and Tagging. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 367–380. Springer, Heidelberg (2007)
5. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing Search via Automated Analysis of Interests and Activities. In: Proceedings of SIGIR 2005 (2005)
6. Arias, M., Cantera, J.M.: Context-based Personalization for Mobile Web Search. In: Proceedings of VLDB 2008 (2008)
7. Hattori, S., Tezuka, T., Tanaka, K.: Context-Aware Query Refinement for Mobile Web Search. In Proceedings of SAINT-W 2007 (2007)
8. Church, K., Smyth, B.: Who, What, Where & When: A New Approach to Mobile Search. In: Proceedings of IUI 2008 (2008)
9. Gregory, D.A., Atkeson, C.G., Hong, J., Long, S.: Kooper,R, Pinkerton, R.: Cyberguide: A Mobile Context-Aware Tour Guide. Wireless Networks (1997)
10. Ozturk, P., Aamodt, A.: Towards a Model of Context for Case-based Diagnostic Problem Solving. In: Proceedings of CONTEXT 1997 (1997)
11. Schilit, B., Adams, N., Want, R.: Context-Aware Computing Applications. In: Proceedings of the Workshop on Mobile Computing Systems and Applications (1994)
12. Liao, L., Patterson, D. J., Fox, D., Kautz, H.: Building Personal Maps from GPS Data. In: Proceedings of IJCAI Workshop on Modeling Others from Observation (2005)
13. Darnell, M.H.H., Moore, J., Essa, I.A.: Exploiting Human Actions and Object Context for Recognition Tasks. In: Proceedings of ICCV 1999 (1999)
14. Heinrich, G.: Parameter Estimation for Text Analysis. Technical Report (2004)