Mobile Phone Recommendation Based on Phone Interest

Bozhi Yuan[1,2], Bin Xu[1,2], Tonglee Chung[1,2], Kaiyan Shuai[3], and Yongbin Liu[1,2]

[1] Department of Computer Science and Technology, Tsinghua University, China
[2] Tsinghua National Laboratory for Information Science and Technology, China
[3] Computer School, Beijing Information Science and Technology University, China
{lawby1229,13636157238}@163.com, xubin@tsinghua.edu.cn,
{tongleechung86,yongbinliu03}@gmail.com

Abstract. As cellular users change mobile phone frequently, mobile phone recommendation system is of great importance for mobile operator to achieve business benefit. There are essential challenges for researchers to design such system. Among them, a critical one is how to obtain and model user's interest of mobile phone. So far, recommendation approaches based on phone's hardware features or personalized web behavior could not achieve satisfactory results. In this paper, we propose phone interest for mobile phone recommendation. Phone interest is a latent level concept which is extracted from a group of users' web log data, who have the same mobile phone. We propose a novel probabilistic model named "Phone Interest Model" only based on mobile web log data. All the log data are from cellular operators server, not from mobile phone's application. The model proves its effectiveness on large scale of station cellular data from real cellular operator. In experiments, we validated the model against 1.3 billion of mobile Web logs for 4 million distinct users in Beijing metropolitan areas, and show that the model achieves a good performance in the phone recommendation, also outperforms the baseline methods and offers significantly high fidelity.

Keywords: phone recommendation, phone interest, mobile data, cellular data, user behavior.

1 Introduction

Mobile phone recommendation system is of great importance for mobile operator to achieve business benefit. For example, mobile service providers always competed in the existing market space, strived for market share of products or services and customers. It's best to leverage recommendation of mobile phone as an opportunity to take the lead and push things forward. In a user-derived market, the number of contract user and contract phone is especially significant for mobile service operators.

So far, recommendation approaches based on phone's hardware features or personalized Web behavior could not achieve satisfactory results. Because each mobile phone is different from hardware, appearance, price etc., hence we have to model the mobile recommendation from the mobile phone's perspective. It's inevitable to lead to a specific mobile phone being suitable for some particular Websites or applications. The

B. Benatallah et al. (Eds.): WISE 2014, Part I, LNCS 8786, pp. 308–323, 2014.

character of mobile phone should be reflected from the crowd of users who have the same phone. And how to establish a model to extract the potential phone interest from the mobile phone has become a very intractable problem. The cellular data is the log data from the station of the operator, which describes the Web-access of all users who have used the cellular network. By using cellular network data, it is tempting to think that by analyzing the crowd of users' Web behavior and latent phone interest for mobile phone recommendation should be easy.

This paper proposes a modeling approach which takes as input cellular data. we explore mobile Web log analysis with user cellular data to build a comprehensive model describing phone interest. The model is named "Phone Interest Model" (PIM), which includes three parts: phone, Website behavior and phone interest. A particularly good data source comes from Event Detail Records (EDRs) maintained by a cellular network operator. EDRs contain information such as the IP address and time of each http link, network volume of uplink and downlink, as well as the identity of the cellular tower with which the phone was associated at that time. The learning process of "Phone Interest Model" (PIM) is as follows:

First, we manually labeled the access intent of the each application as "App-Behavior" via the "Useragent" field of the raw cellular data of the operator. Second, Websites are extracted from raw Web log for each user. Associated with access intent (App-Behavior) of each user access, the features of each mobile phone are generated. Third, a probabilistic topic model is proposed for extracting latent phone interest layer between mobile phone and Websites-behavior. We validate the PIM model against the EDRs from Beijing. The data lasts for 2 weeks, and covers 2 main urban districts, with an area of approximately 40 square kilometers. And more than 4 million user IDs appeared in the estimated 1.3 billion EDRs.

The contributions of the paper are: We propose a novel probabilistic model Phone Interest Model (PIM) to analyze phone interest, based only on the usage history of cellular data. And we proves its effectiveness for mobile phone recommendation on large scale of EDRs from real cellular data.

The paper is organized as follows.In Section 2, we describe recent related work on Web user behavior analysis and mobile recommendation.In Section 3, we give an overview of the data we use, and observe some important characteristics and definitions of the data in several aspects, and raise the remaining problem. Section 4 describes the discovery of "App-Behavior" from raw HTTP log, also introduces two baseline method "K-means Interest Learning" (KIL) and "K-SVM Interest Learning" and formally describes the "Phone Interest Model" (PIM) in detail. Section 5 describes the experimental results applying the framework on our dataset, and compares performance between the different method and gives detailed analysis on the result in several different angels. Finally, we conclude the paper in Section 6.

2 Related Work

User interest and behavior mining based on Web usage has long been a hot topic[1][2]. White, et, al.[3] consider 5 different contextual information to model user interest, and then do recommendation based on it. Nasraoui, et, al. [4] study user behavior of a particular Website based on tracking user profiles and their evolving. Some researchers use clustering methods to extract types of users.[5][6] But they either do clustering on the users' perspective and cluster user into different types, or on the Websites' perspective and make URL groups. Extracting user types as well as Website topics in a unified model with hierarchical clustering methods is still rarely seen.

The mobile recommendation and human behavior play an important role in many fields.Chen, Deng-Neng, et al.[7] built a recommendation system via the AHP. Soe-Tsyr Yuan, Y.W[8] presented a personalized contextualized mobile advertising infrastructure for the recommendation of advertisement. Fan.Y, Zhimei.W[9] built a scalable personalized mobile information pushing platform, which can recommend the location-based services to users. Tsao.Kowatsch T, Maass W.[10] investigated the use of mobile recommendation agents and they developed a model to better understand the impact of MRAs on usage intentions, product purchases and store preferences of consumers. Do, Gatica-Perez [11] mine user pattern using mobile phone app usage, including mobile Web usage on mobile phone.Zheng V W, Cao B, Zheng Y, et[12] mined useful knowledge from many users' GPS trajectories based on their partial location and activity annotations to provide targeted collaborative location and activity recommendations for each user.Huang K, Zhang C, Ma X, et[13] use a variety of contextual information, such as last used App, time, location, and the user profile, to predict the user's App whether will be open. Pinyapong S, Kato T.[14] proposed the relationship between 3 factors which are time, place and purpose. In consequence, they have summarized the basic rules to analyze essential data and algorithms to query processing. Ricci F[15].has done a nice survey of the mobile recommender systems, who has illustrated the overview of major techniques supported functions, and specific computational models.

Our method mines phone Interest patterns in mobile Web usage from the station cellular data of mobile operator's perspective, which is both comprehensive and large scale.

3 Problem Definition

We present required definitions and formulate the problem of the mobile phone recommendation based on the station cellular data of mobile operator. Without loss of generality, we assume there are two sets of the mobile users, source user set and the target user set. Our goal is to recommend one or more specific mobile phone to the target user from the model which is trained by the source user set. First of all, we give some formal definitions of the concepts used.

Table 1. Field details of the dataset

Filed name	Data type	Description
User Id	String	IMSI, the unique identifier of a user
Phone Id	String	IMEI, the unique identifier of a phone
Host	String	The domain name of the host requested
Content Type	String	The ContentType attribute in HTTP header
User agent	String	The app infomation in HTTP header

Table 2. Application Behavior

1	SEARCHING	11	PLAYING GAME
2	DO RECORD	12	MANAGING PHONE
3	BROWSING WEIBO	13	INTERNET SURFING
4	DO SHOPPING	14	READING
5	CHATTING	15	PHOTOGRAPHING
6	WATCHING STREAM	16	SENDING/RECIEVING MAIL
7	BROWSING RENREN	17	SEARCHING MAP
8	READ NEWS	18	LISTENING MUSIC
9	FLASHSLIGHTING	19	MESSAGING
10	BROWSING BLOG/ZONE	20	COMMUNICATION

Definition 1. ***Cellular data.*** *The cellular data is the log data from the station of the operator towers. Each tower records the cellular data, which can describes the Web-access of the users who are covered by a tower. One line in the dataset corresponds to a HTTP request/response pair occurred when using cellular network. All of follow framework and the method in this paper are based on cellular data. The main structure of the cellular data is shown in Table 1.*

Definition 2. ***User.*** *A user if uniquely identified by "IMSI" (the unique id of a SIM card) in the dataset. It corresponds to a real person using the mobile Web, regardless of what devices are used.*

Definition 3. ***Mobile Phone/Device.*** *A mobile device (phone or pad) which is uniquely identified by "IMEI" (the unique id of a phone) in the dataset. It corresponds to a real device (a cellphone serial number) using the mobile Web, regardless of its phone number.*

Definition 4. ***Useragent.*** *The useragent is the domain name of the HTTP request which acts as a client in a network protocol during communications within a client server distributed computing system. Some specific useragents can reflect the details of users' application.*

Definition 5. ***Website.*** *A Website is the domain name of the HTTP request. It may or may not be the address that is directly requested by the user / app. The Website reflects the host address of the HTTP request server.*

Definition 6. ***App-Behavior.*** *When the user uses web-applications, the applications have to send one or more request to the service host. Each request contains a intention from the application or user. For example, a request requires the game service, so the intention of the request is "gaming". We define the intention of each request as app-behavior. We can recognize the application's intention via the "Useragent" domain from the cellular data, and give the each app-behavior a label of natural language. The 20 types App-Behavior is shown in Table 2.*

Definition 7. ***Phone Interest.*** *Phone Interest is the distribution of the Website and App-Behavior, which reflects the interest of users who have device (mobile phone or pad)* *with the same type are interested in. The phone interest may contain several different Websites and App-Behaviors. Each tuple of Website and App-Behavior is associated with a weight, representing for the degree of users' fondness. On the other hand Website and App-Behavior may serve different phone interest.*

Problem 1. **Mobile Phone Recommendation**. (1) Given a complete cellular data in a period of time of an area. (2) Find out the latent phone interest of each type of phone from the cellular data, and build an interest model. (3) According to the "Phone Interest Model" (PIM) and the records of user, recommend the suitable phones to the user.

4 Phone-Interest Model (PIM)

In this section, we proposed a method to build a model, which can represent the real property and phone interest (defined in Definition 7) of different mobile phone. We attempt to extract latent phones' interests from their Web usage cellular data (defined in Definition 1); we propose two baseline methods K-means[16][17] and SVM (Support Vector Machine)[18][19] and also propose a probabilistic topic modeling method based on LDA (Latent Dirichlet Allocation)[20]. In our model, extracted latent layer represents the phone interest.

Let us briefly introduce notations below. T is the set of the phone interest; z is the index of phone interest; U is the set of user; u is the index of user; D is the set of the mobile devices; d is the index of mobile device; w is the index of Website of each request server address; b is the index of App-Behavior of each request want to express;

4.1 App-Behavior Discovery

Since raw HTTP log may not truly reflect user behavior. We do some additional work for the each HTTP request to express the intention of requester. We define the intention of requester as App-Behavior. First, we extract the Useragent domain from HTTP cellular data, and then we use natural language recognition to analyze the purpose of the request, and tag a formal label as the App-Behavior for this request. The label set contains 20 distinct labels, and each of the word can express a kind of attempt of the user. Based on the cellular data analysis above, the 20 different App-Behavior labels are given in Table 2.

4.2 Website and App-Behavior Represent of the Mobile Phone Feature

Since same Website address can appear in different request cases, and each host of those request may not have the same Web intention, we can't only use the Website address as the feature of the mobile phone. And we should add the user purpose part to the mobile phone feature. For each mobile device d, visited Website w, and each App-Behavior b, we define the transformation $f(w,b)$ as the feature of mobile phone. The set of mobile device D is represent as $D = \{\langle f(w,b)_{di}, n_{f(w,b)_{di}} \rangle\}$, where $f(w,b)_{di}$ is the i^{th} feature dimension $f(w,b)$ of the mobile device d, and $n_{f(w,b)_{di}}$ is the number of times that $f(w,b)_{di}$ is visited by mobile device d. According to feature expression above, it can describe the purpose of the user, and also can reflect the different Website. For simplicity,we use symbol F to represent the feature space of $f(w, b)$, and use f to represent dimension index of $f(w, b)$ in feature space F.

4.3 K-Means Interest Learning (KIL)

We can use an easy method to extract “phone interest”. We assume that the user always try to use or access the applications or websites overtime, even if the hardware or device is not perfectly suitable for those services. And most of the users can be divided into the limited number of groups, so we can use the tuple of host, App-Behavior as each phone’s feature field, and run a cluster method to get each centroid of all clusters, which can represent the favorite hosts and App-Behaviors of each specific mobile device. Then the features of user’s web behavior from the user log data is extracted, to match which kinds of mobile devices are similar to them.

For benchmark testing, we use K-means[16][17] method to cluster the mobile devices D into several clusters, with the tuple of host, App-Behavior F as the device feature. The vectors of the centroid which we calculate in the last step, are considered as “user interest”. We can use those vectors to judge the test user which cluster they belong to, and recommend mobile devices which exist in the cluster to user.

The specific steps of “K-means Interest Learning” (KIL) are shown in Algorithm 1. The input is the set of mobile device D and the interest number K. The algorithm returns the set of interest centroid C. The goal is to calculate the distance between the formal cellular data of user u and interest center c to decide which c_j a user u belongs to. Finally, with the user u who is clustered by centroid c_j, the mobile devices d which are assigned to c_j will be recommended to this user.

Algorithm 1. K-means Interest Learning

```
Require:
  The set of mobile device $D$
  The interest number $K$
Ensure:
  Initialize $\{c_1, c_2 ... c_K\}$ be the cluster centers of interest set $C$.
  while each $\{c_1, c_2 ... c_K\}$ are not convergence to a stable value do
      for each mobile device $d \in D$ do
          assign the device $d$ to the closest interest cluster $c_j$
      end for
      for each interest center $c_i \in C$ do
          update $c_i$ by averaging all of the device $d$ that have been assigned to it
          if $c_i$ don't contain any device $d$ then
              $c_i \leftarrow randomd$
          end if
      end for
  end while
  return $\{c_1, c_2 ... c_K\}$;
```

4.4 K-SVM Interest Learning (KSIL)

"K-SVM Interest Learning" (KSIL) is modified based on the "K-means Interest Learning" (KIL). KSIL retained all steps of the KIL, and introduced "Support Vector Machine" (SVM)[18][19] for training and predicting. In the training step, the mobile devices d which are assigned by interest centroid c from KIL should be trained as the instance of c. We use F as the device feature, and the identify of interest centroid as the label. In the prediction stage, we use the user u as the instance,and use formal cellular data f as the feature, then use the SVM model to predict which class could be the user u was belonged to. Finally, recommend the mobile devices d which exist in this class to the user u.

4.5 Phone-Interest Model

Topic Model is commonly used in text mining for discovering abstract "topics" in a set of documents. LDA (Latent Dirichlet Allocation) is a commonly used topic model currently, and it has also been applied in discovering user behavior patterns[21]. LDA is a unsupervised, generative model, which models the generation of a document into a two-step process: choosing a topic based on topics distribution over a document; and choosing a word based on words distribution over a topic. We use the specific mobile device type representation of all users who use this type as a document, and use each user's formal log data as the feature to express the words' vector space, and propose a probabilistic topic model for phone interest modeling. Table 3 summarizes the notations used in the PIM.

We also propose a generative model of record set of the mobile devices. Assume that the generating scheme of cellular records are as follows:

There are many hosts/app-behaviors $F = \{f_1, f_2 ...\}$ that belong to each mobile device d,and each f may belong to different latent phone interest $T = \{z_1, z_2 ...\}$ with different probability, so there is a probability vector $\theta_d = \{p_{f1}, p_{f2} ...\}$ to represent the probability of each host/app-behavior of each device which belong to different phone interest, meanwhile the probability is not unique. Accordingly, we can sample the phone interest z for each specific position of a specific device d by θ_d. And we can unite a probability matrix ϕ of z given f to generate the f for each position by Equation 1.

$$p(f) = \sum_z p(z/\theta_d)p(f/z, \phi) \tag{1}$$

z denotes the phone interest, θ_d denotes the probability distribution of a specific mobile device d over phone interest z and ϕ denotes the probability matrix z over all host/app-behavior f. In each device, there are n independent feature f, and the different θ_d can produce a same f, also f can be produced by different z, so the probability of a specific device d record can be generated by:

$$p(d) = \int_{\theta_d} p(\theta/\alpha) \prod_n \sum_z p(z/\theta_d)p(f/z, \phi) \tag{2}$$

α denotes the hyper parameter of the multinomial distributions and n denotes the host/app-behavior set of each device . If we want to generated all records of the device set we must accumulate the product of every device d , so the generation function of all record set is:

$$p = \prod_{d \in D} p(d) \tag{3}$$

Our goal is to calculate the 2 probability distributions θ and ϕ via phone interest model for clustering the mobile devices and recommending the mobile phones to user, also we have to find out the latent phone interest T from the set of mobile devices D. θ is a multinomial distribution over T specific to mobile device, and the dimensions of θ is $|D| \times |T|$. ϕ is a multinomial distribution over F specific to T, and the dimensions of ϕ is $|T| \times |F|$. We analyse phone interest z each mobile device d belongs to via the θ matrix. In the matrix θ, each row represents a probability distribution of a mobile device d to the latent phone interest T. We use the ϕ matrix and a specific user u formal log data F to predict which phone interest z the user u belong to, and recommend the suitable mobile devices to the user u, meanwhile we extract some phone features f with high probability value on the each topic z to represent the every $PhoneInterest$.

Table 3. Symbol description

T	the set of phone interests.
K	the number of phone interests.
D	the set of mobile devices.
B	the set of App-Behavior.
U	the set of users.
F	the set of formal cellular data feature.
z	the index of phone interest.
d	the index of mobile device.
u	the index of user.
w	the Website.
b	the App-Behavior.
$f(w,b)$	the formal cellular data feature, which is represented by tuple of Website and App-Behavior .
f	the index of formal cellular data feature.
x_{di}	the i^{th} $f(w,b)$ attribute (word) in mobile device d.
z_{di}	the phone interest(topic) assigned to attribute x_{di}.
s_{di}	if x_{di} is a word from a single domain or a cross domain.
θ_d	multinomial distribution over phone interest(topic) specific to mobile device d.
ϕ_z	multinomial distribution over tuple of host and app-behavior specific to phone interest(topic) z.
α, β	Dirichlet priors to multinomial distributions θ and ϕ.

We combine mobile device D, tuple of Website as well as App-Behavior F and phone interest T in a unified generative model PIM. The generative process of "Phone Interest Model" (PIM) is as follows Figure 1.

1. For each mobile device d, draw θ_d from Dirichlet prior α;
2. For each phone interest z, draw ϕ_z, from Dirichlet priors β_z;
3. For each feature i of d appearance $f(w,b)_{di}$ in mobile device d:
   - draw a phone interest (topic) z_{di} from a multinomial distribution θ_d.
   - draw a term for $f(w,b)_{di}$ from multinomial distribution ϕ_{zdi}.

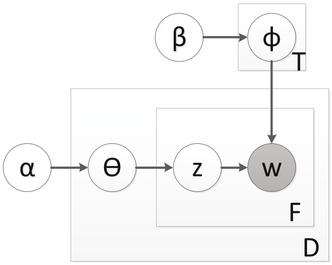

Fig. 1. Plate representation of Phone-Interest Model (PIM)

We use Gibbs sampling to estimate the model parameters, following [6]. For simplicity, we take fixed values for hyper parameters α and β (i.e. $\alpha = 50/|T|$, $\beta = 0.01$). We use Gibbs sampling to estimate the posterior distribution on $f(w,b)$ and z, then use the result to estimate θ and ϕ. The posterior probability can be calculated by Equation 4.

$$P(z_{di}|T_{-di},F,\alpha,\beta) \propto \frac{m_{d,z_{di}}^{-di}+\alpha_{z_{di}}}{\sum_{z=1}^{K}(m_{d_{di},z}^{-di}+\alpha_z)} \frac{n_{z_{di},f_{di}}^{-di}+\beta_{f_{di}}}{\sum_{f=1}^{F}(n_{z_{di},f}^{-di}+\beta_f)} \tag{4}$$

$m_{d,z}$ denotes the number of times that phone interest z occurs in mobile device d, and $n_{z,f}$ denotes the number of times that feature f of the mobile device generated by phone interest z. Superscript $-di$ means that the quantity is counted excluding the current instance by the i^{th} dimension of the d^{th} mobile device. After the sampling convergences, the multinomial distribution parameters θ and β can be estimated as follows:

$$\theta_{d,z} = \frac{m_{d,z}+\alpha_z}{\sum_{z'}(m_{d,z'}+\alpha_{z'})} \tag{5}$$

$$\phi_{z,f} = \frac{n_{z,f(w,b)}+\beta_{f(w,b)}}{\sum_{f(w,b)'}(n_{z,f(w,b)'}+\beta_{f(w,b)'})} \tag{6}$$

Also, if a user u with his formal cellular data feature f is given,and then we inference the probability distribution over the phone interest T of the user u by:

$$P(T_i = j|T_{-i}, f(w,b)_i, d_i, \cdot) \propto \frac{n_k^{(t)}+n_k^{(t)'}+\beta_t}{\sum_{f=1}^{F}(n_k^{(f)}+n_k^{(t)'}+\beta_f)} \cdot \frac{n_m^{(k)}+\alpha_k}{\sum_{z=1}^{K}(n_m^{z}+\alpha_z)} \tag{7}$$

Finally, we get the $\theta_{d,z}$ and $\phi_{z,f}$ cross the Equation 5 and 6. The characteristic of the phone interest z can be reflected by the f items which own the higher probability in the ϕ_z. So the phone interest z of each mobile device is expressed by several top f with maximum probability in the ϕ_z. For mobile phone recommendation, we have to infer which phone interest a new user's Web behavior most possibly belongs to using Equation 7 and the cellular data. Then according to θ_z, the user will be recommended the mobile device d, which already belongs to this phone interest z.

5 Experimental Results and Evaluations

In this section, we will demonstrate the experiment results of applying the framework on our dataset, and give analysis based on the experiment results.

5.1 Data Description

The data we use is the mobile Web usage log (HTTP request log) of the cellular network (2G, GPRS) of a mobile operator. The dataset covers the geographical range of Beijing, the capital of China. The data we use covers 2 urban districts, with an area of approximately 40 square kilometers. The area contains a major set of attractions, as well as one of the busiest business districts of the city.

Overview. The time range of the dataset is from June.8 to June.20, 2013. There are totally 1,344,654,257 complete records in the dataset. There are 4,084,230 distinct users in the raw dataset. We clean the raw data by 1) As we focus on tuple of Website, App-Behavior and phone interest, so we use Useragent field to filter out all supporting requests with inner joined by the App-Behavior table. 2) Records with meaningless (i.e. full of question marks) Host field are removed or null. After cleaning, there are 34.262% (460,713,102) records and 1,244,272 distinct users left.

We extracted the top 2,000 users, whose records are over 500 from the data, with 145 distinct phone types (IMEI). We retained the mobile phone records which keep over 10 distinct users in the data. Finally, we have got the data which hold a total of 18,908,715 records, 44 distinct mobile device types of the phone, 19,994 distinct hosts and 1,602 users as our source data. The distribution of mobile device in count of users is shown in Figure 2.

5.2 Phone-Interest Discovery

We run the PIM with the above source data, and use the Equation 6 to calculate a $T \times F$ ϕ parameter matrix, and follow Equation 5 to calculate a $|D| \times |T|$ θ matrix, where D is the number of the mobile devices in the source data. We use PIM-k to donate the PIM model which runs with k phone interests, and use ϕ_k, θ_k to donate the ϕ and θ which calculated by the PIM-k.

Evaluation metrics. We evaluate the PIM models based on the source data with mobile device $|D| = 44$, Website host $|W| = 19994$, App-Behavior $|B| = 20$, phone interest $|T| = \{5, 6, 7, 8\}$. To quantitatively evaluate the proposed methods, we use other cellular data as the testing data. All of the mobile devices in the testing data are known, to verify the accuracy of PIM model. In evaluation, since we have filtered the users whose records are more than others as the testing data, we consider those users

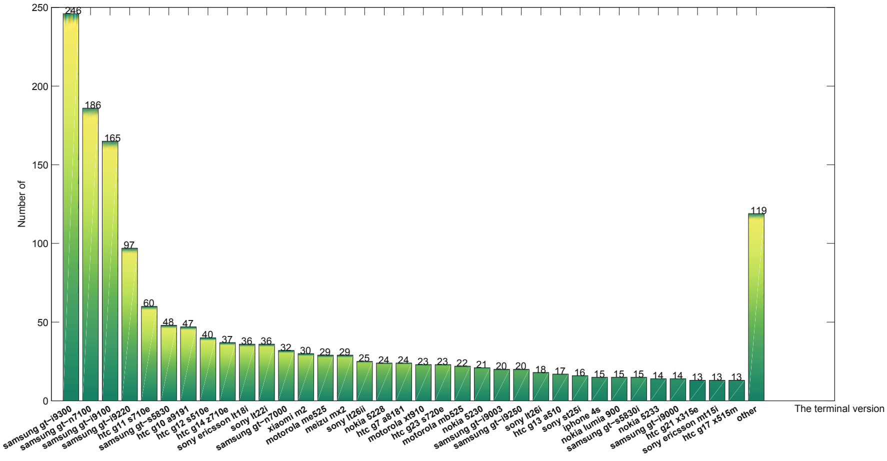

Fig. 2. Distribution of each mobile phone types

already have a suitable phone, which can perfect match their phone interest. According to the phone interest, the PIM model will predict some specific mobile device types for each user, if the result of prediction included the mobile device type which can perfect match the user's, and then we say the prediction is correct, otherwise we say the prediction is wrong. Based on this, we evaluate the prediction performance in terms of P@1 (Precision of the mobile phone recommendation based on the first prediction phone interest), P@2 (Precision of the mobile phone recommendation based on the both first and second prediction phone interest), Recall (Recall for the prediction phone interest) and MAP (Mean Average Precision).

All codes are implemented in Java, and all the experiments are conducted on an x64 server with E5-2609 2.4GHz Intel Xeon CPU and 32G RAM. The operation system is Microsoft Windows Sever 2008 R2 Enterprise. For training the PIM-6, PIM-7 and PIM-8 models, it takes about 2 hours respectively on the entire data set (18908715 records, 44 distinct mobile phone types). Recognizing the computation complexity of LDA style models, we are currently looking into developing more efficient computation mechanism to speed up the process.

5.3 Performance Analysis

Table 4 lists the performance of PIM on different test cases. In the table, MI0 means the first phone interest in the PIM, and each row means the sub-performance of the different phone interest. The last row shows us the average of the performance in the PIM. We also run the "K-means Interest Learning" KIL and "K-SVM Interest Learning" KSIL method on our dataset. The performance of those methods are shown in the Table 5. The proposed PIM method clearly gets an outstanding performance, and outperforms the baseline methods (KIL and KSIL) with different phone interests.

Table 4. The performance with diffrent phone interest

(a) PIM-5

	P@1	**P@2**	**Recall**	**MAP**
MI0	0.4939	0.5783	0.6507	0.8450
MI1	0.3212	0.8339	0.3886	0.9010
MI2	0.6855	0.7860	0.4523	0.8477
MI3	0.1396	0.1955	0.3472	0.7522
MI4	0.3750	0.3750	0.6774	0.8000
Average	**0.4496**	**0.6510**	**0.5032**	**0.8292**

(b) PIM-6

	P@1	**P@2**	**Recall**	**MAP**
MI0	0.5967	0.6290	0.925	0.8852
MI1	0.3571	0.3571	0.7407	0.8203
MI2	0.3181	0.4090	0.4375	0.7428
MI3	0.3939	0.4545	0.4482	0.8468
MI4	0.1923	0.8076	0.3333	0.9618
MI5	0.8070	0.8070	0.3458	0.8382
Average	**0.4878**	**0.5709**	**0.5384**	**0.8492**

(c) PIM-7

	P@1	**P@2**	**Recall**	**MAP**
MI0	0.3275	0.3448	0.6333	0.6966
MI1	0.1923	0.2307	0.2777	0.9389
MI2	0.3666	0.3666	0.3859	0.8414
MI3	0.125	0.125	0.3125	0.8527
MI4	0.5967	0.6290	0.9024	0.9016
MI5	0.8125	0.875	0.6842	0.6829
MI6	0.6923	0.6923	0.3082	0.8002
Average	**0.4464**	**0.4617**	**0.5006**	**0.8163**

(d) PIM-8

	P@1	**P@2**	**Recall**	**MAP**
MI0	0.5806	0.5806	0.7659	0.8965
MI1	0.2068	0.2241	0.1666	0.9520
MI2	0.3253	0.4216	0.1516	0.7875
MI3	0.1791	0.4701	0.4800	0.6772
MI4	0.6058	0.6569	0.2150	0.7832
MI5	0.1724	0.1896	0.2173	0.6933
MI6	0.8125	0.9375	0.5652	0.4864
MI7	0.2226	0.2226	0.7794	0.9237
Average	**0.3065**	**0.4119**	**0.4176**	**0.7750**

Phone Interest Topic Analysis. How many topics are enough for the phone recommendation? We perform an analysis by varying the number of phone interest topics in the proposed PIM method. Figure 3 shows its P@1, P@2, Recall and MAP performance with the number of phone interest topics varied. We see when the phone interest number is up to 6, decreasing the number often obtains a performance improvement. The precision trend becomes best when the number is at 6. This demonstrates the stability of the PIM method with respect to the number of topics. On the other hand, the phone interest number is defined by the how many mobile devices you want to recommend to the user. If is used k to express the number of mobile phone types will recommend to user, and then we roughly calculate the phone interest number $|T|$ via the expression $|T| = |D|/k$.

According to PIM-5, we displayed all 5 phone interests, also we listed top 5 frequent mobile devices with their weights from each phone interest. Their topic Websites/App-Behavior and mobile devices are listed in Table 6. The main mobile devices of the interests are listed, and top 5 tuple of Websites and App-Behavior of each interest are shown with their domains and weights within the interest.

It can be seen that the topic Websites of the interests are quite centralized. Analyst can draw the conclusions from the observation, even if give some meaningful tags for the users of different phone interest cluster, which can present the difference between the different crowd.

Table 5. Recommendation performance by different methods: KIL(K-means), KSIL (K-SVM), PIM

PI	Method	P@1	P@2	Recall	MAP
5 Interests	KIL	0.3995	**0.6777**	0.3752	0.4125
	KSIL	0.387	-	0.382	-
	PIM	**0.4496**	0.6510	**0.5032**	**0.8292**
6 Interests	KIL	0.3323	0.5195	0.3229	0.3427
	KSIL	0.344	-	0.334	-
	PIM	**0.4878**	**0.5709**	**0.5384**	**0.8492**
7 Interests	KIL	0.3430	**0.5254**	0.2905	0.3741
	KSIL	0.370	-	0.354	-
	PIM	**0.4464**	0.4617	**0.5006**	**0.8163**

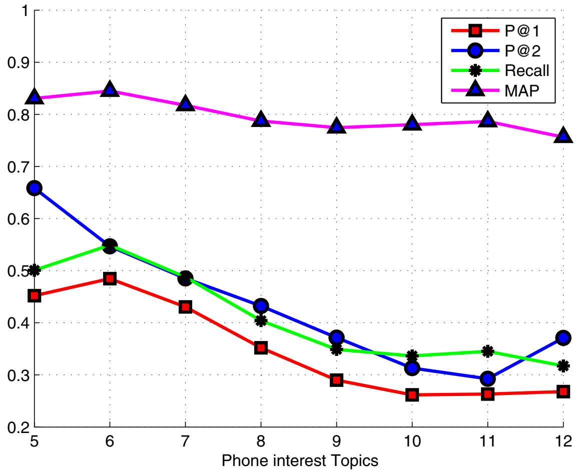

Fig. 3. Performance of PIM with different topic

Interest 0 is mostly about chatting, the top 2 Websites are related to live chat, but the chat clients are supported by different providers, and it is shown that Nokia phone is dominant in this interest, which is proven by the fact that all top 5 the most frequent mobile device are produced by Nokia phone.

Interest 1 is centralized by news reading, where about 35% of usage is used for browsing or reading news from the news application of Netease 163. In addition, mail service and browsing blog service also take up a significant part .

Interest 2 is great centralized by send/recieving mail service, since more than half of its usages are related to mail service. And we can observe that a lot of Samsung devices had very good support for this function.

Interest 3 is mostly about sending/recieving the mail service, also searching service and news service have played main roles in this interest. The listed devices indicate that sony ericsson lt18i is perfect suitable for those Web usages.

Table 6. Latent “phone interests” and clusters of mobile phone from the PIM5

	Mobile Phone Type	Weight of	Websites App-Behavior	Weight of
Phone Interest 0	nokia 5228	0.9993	wx.qlogo.cn CHATTING	0.3483
	nokia c7-00	0.9988	mmsns.qpic.cn CHATTING	0.3084
	nokia c5-03	0.9986	short.weixin.qq.com SENDING/RECIEVING MAIL	0.1638
	nokia 5230	0.9775	mobilemaps.clients.google.com SEARCHING	0.0692
	nokia 5233	0.9755	api.baiyue.baidu.com READ NEWS	0.0231
Phone Interest 1	htc g14 710e	0.9788	p.3g.163.com READ NEWS	0.3528
	motorola me525	0.9778	short.weixin.qq.com SENDING/RECIEVING MAIL	0.1803
	sony lt26ii	0.9619	m.qpic.cn BROWSING BLOG/ZONE	0.1010
	htc g13 a510	0.9464	ugc.qpic.cn BROWSING BLOG/ZONE	0.0464
	samsung gt-s5660	0.8645	qzonestyle.gtimg.cn BROWSING BLOG/ZONE	0.0387
Phone Interest 2	motorola xt910	0.9987	short.weixin.qq.com SENDING/RECIEVING MAIL	0.5143
	samsung gt-s7500	0.9914	m.qpic.cn BROWSING BLOG/ZONE	0.0795
	sony lt26i	0.9908	www.google-analytics.com MANAGING PHONE	0.0319
	samsung gt-n7000	0.9774	in1.feed.uu.cc SEARCHING	0.0251
	samsung gt-s5670	0.9224	api.mobile.360.cn MANAGING PHONE	0.0189
Phone Interest 3	sony ericsson lt18i	0.9930	short.weixin.qq.com SENDING/RECIEVING MAIL	0.4139
	htc g10 a9191	0.8894	p.3g.163.com READ NEWS	0.2918
	samsung gt-i9000	0.7955	m.api.dianping.com SEARCHING MAP	0.0913
	samsung gt-s5830i	0.7909	gomarket.goapk.com SEARCHING	0.0885
	nokia lumia 900	0.7778	218.205.179.22:8002 SEARCHING	0.0340
Phone Interest 4	iphone 4s	0.9106	api.weibo.cn BROWSING WEIBO	0.3194
	htc g12 s510e	0.8197	ww3.sinaimg.cn BROWSING WEIBO	0.0867
			ww1.sinaimg.cn BROWSING WEIBO	0.0775
			ww2.sinaimg.cn BROWSING WEIBO	0.0684
			wbapp.mobile.sina.cn BROWSING WEIBO	0.0580

Interest 4 is centralized by browsing Weibo (English name micro blog), a popular micro blog app in China, many users use Iphone 4s or HTC g12 s510e for Weibo.

6 Conclusion

In this paper, we have presented a framework for recommending the mobile device to user. We have proposed a modeling method that models the Web cellular data of the operator. We have also presented a probabilistic topic modeling method to extract latent phone interest from mobile Web usage cellular data. We have applied the proposed framework to a real world large scale dataset from Beijing, capital of China, with more than 18 million data records, and the output shows outstanding mobile phone recommendation accuracy. We have analyzed city-level collective behavior patterns in mobile Web usage based on the model output, and discussed mobile phone clustering and performance using phone interest distribution. For future work, one possibility to extend the work is to take social connections between users in consideration. Mobile Web usage behaviors can spread over social networks, thus establishing relationships between users according to social connections would help enrich context information of a user and improve accuracy of phone interest modeling.

Acknowledgement. This work is supported by China National Science Foundation under grant No.61170212, China National High-Tech Project (863) under grant No. SS2013AA010307, and Ministry of Education-China Mobile Research Fund under grant No.MCM20130381. Beijing Key Lab of Networked Multimedia also supports our research work.

References

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl. 1(2), 12–23 (2000)
2. Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explor. Newsl. 2(1), 1–15 (2000)
3. White, R.W., Bailey, P., Chen, L.: Predicting user interests from contextual information. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 363–370. ACM, New York (2009)
4. Nasraoui, O., Soliman, M., Saka, E., Badia, A., Germain, R.: A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Transactions on Knowledge and Data Engineering 20(2), 202–215 (2008)
5. Xu, J., Liu, H.: Web user clustering analysis based on kmeans algorithm. In: 2010 International Conference on Information Networking and Automation (ICINA), vol. 2, pp. V2-6 –V2-9 (October 2010)
6. Mobasher, B., Cooley, R., Srivastava, J.: Creating adaptive web sites through usage-based clustering of urls. In: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange (KDEX 1999), pp. 19–25 (1999)
7. Chen, D.N., Hu, P.J.H., Kuo, Y.R., Liang, T.P.: A web-based personalized recommendation system for mobile phone selection: Design, implementation, and evaluation. Expert Systems with Applications 37(12), 8201–8210 (2010)
8. Yuan, S.T., Tsao, Y.W.: A recommendation mechanism for contextualized mobile advertising. Expert Systems with Applications 24(4), 399–414 (2003)
9. Yang, F., Wang, Z.: A mobile location-based information recommendation system based on gps and web 2.0 services. Database 7, 8 (2009)
10. Kowatsch, T., Maass, W.: In-store consumer behavior: How mobile recommendation agents influence usage intentions, product purchases, and store preferences. Computers in Human Behavior 26(4), 697–704 (2010)
11. Do, T.M.T., Gatica-Perez, D.: By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia, MUM 2010, pp. 27:1–27:10. ACM, New York (2010)
12. Zheng, V.W., Cao, B., Zheng, Y., Xie, X., Yang, Q.: Collaborative filtering meets mobile recommendation: A user-centered approach. In: AAAI, vol. 10, pp. 236–241 (2010)
13. Huang, K., Zhang, C., Ma, X., Chen, G.: Predicting mobile application usage using contextual information. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1059–1065. ACM (2012)
14. Pinyapong, S., Kato, T.: Query processing algorithms for time, place, purpose and personal profile sensitive mobile recommendation. In: 2004 International Conference on Cyberworlds, pp. 423–430. IEEE (2004)
15. Horozov, T., Narasimhan, N., Vasudevan, V.: Using location for personalized poi recommendations in mobile environments. In: International Symposium on Applications and the Internet, SAINT 2006, p. 6. IEEE (2006)

16. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Applied Statistics, 100–108 (1979)
17. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al.: Constrained k-means clustering with background knowledge. In: ICML, vol. 1, pp. 577–584 (2001)
18. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9(3), 293–300 (1999)
19. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research 2, 45–66 (2002)
20. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
21. Farrahi, K., Gatica-Perez, D.: Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2(1), 3:1–3:27 (2011)