Mobile Web User Behavior Modeling

Bozhi Yuan[1,2], Bin Xu[1,2], Chao Wu[1,2], and Yuanchao Ma[1,2]

[1] Department of Computer Science and Technology, Tsinghua University, China
[2] Tsinghua National Laboratory for Information Science and Technology, China
lawby1229@163.com, xubin@tsinghua.edu.cn, ariesnix93@gmail.com, myccs@outlook.com

Abstract. Models of mobile web user behavior have broad applicability in fields such as mobile network optimization, mobile web content recommendation, collective behavior analysis, and human dynamics. This paper proposes and evaluates URI model, a novel approach to analyze user mobile Web usage behavior, which combines user interest modeling with location analysis. The URI model takes as input mobile user web logs associated with coarse-grained location drawn from real data, such as Event Detail Records(EDRs) from a cellular telephone network. We use probabilistic topic modeling to discover latent *user interest* from user mobile Web usage log. We validated the URI model against billions of mobile web logs for millions of cellular phones in Beijing metropolitan areas. Experiments show that the URI model achieves a good performance, and offers significantly high fidelity.

Keywords: Behavior modeling, Location mining, Behavior pattern analysis, Web Behavior, Latent Web Interest.

1 Introduction

Mobile web user behavior reflects human mobility, and has broad uses in mobile computing research and other fields of study. Models of mobile web user behavior can help answer questions in area like web user content recommendation, mobile network optimization, collective behavior analysis, and human dynamics.

Our work aims to produce accurate model of how people access mobile web and move in a city. To achieve this general aim, we define a number of more specific goals. The first goal is to discover the geographical region when users access mobile web, such as living, work, business, and way. Different users have the habits to access mobile web in different locations. The second goal is to discover user's interests on mobile web, such as news, sports, and entertainment etc. Different user likes access different kinds of website, which reflects his/her interests on mobile web. We use probability distribution to represent user's interest. The third goal is to discover the correlation among user, region and interest. Through analyzing users' regions and interests on accessing mobile web, we can find the behavior pattern of users.

The contributions of the paper are: *1*) We propose a novel probabilistic model to analyze mobile web user behavior, only based on the usage history of mobile

B. Benatallah et al. (Eds.): WISE 2014, Part I, LNCS 8786, pp. 388–397, 2014.

Web. The model proves its effectiveness on large scale of EDRs from real cellular operator. The model can be used to do city-level collective behavior analysis, as well as mobile Web usage prediction/service recommendation. *2*) A new approach is presented to discover regions by leveraging geographical feature and Web usage traffic history. Raw web log like EDRs can only give coarse information about user location, while our approach can give semantic information of user location such as living/work/business/way, which is important to make the model applicable to different scenarios.

The rest of the paper is organized as follows. In Section 2, we describe recent related work on Web user behavior analysis and location analysis. In Section 3, we give an overview of the data we use, and observe some important characteristics of the data in several aspects. Section 4 describes the discovery of *Region* from raw HTTP log. Section 5 formally describes the probabilistic topic model we use for user behavior modeling in detail. Section 6 describes the experimental results applying the framework on our dataset, and gives detailed analysis on the result in several different angels. Finally, we conclude the paper in Section 7.

2 Related Work

User interest and behavior mining based on Web log data has been a hot topic[1][2]. Some researchers use clustering methods to extract types of users [3][4]. They either do clustering on the users' perspective and cluster user into different types, or on the websites' perspective and make URL groups. *Nasraoui, et, al.* [5] study user behavior of a particular website based on tracking user profiles and their evolving.

As the mobile Web takes more and more proportions of people's total Web usage, study of mobile Web user behavior also gains a great attention recently. *Cui and Roto* [6] describe how people use the web on mobile devices by contextual inquiries, and analyze contextual factor as well as user activity patterns. *Tseng and Lin* [7] mine user behavior patterns in mobile web systems based on location trace, and do experiments using simulation. *Phatak and Mulvaney* [8] propose a fuzzy clustering method on URLs and users based on a distance matrix, and further do user profiling and recommendation. *Do and Gatica-Perez*[9] mine user pattern using mobile phone app usage, including mobile Web usage on mobile phone. *Verkasalo* [10] analyzes contextual patterns in mobile service usage statistically, using handset-based data, which includes location and Web usage data. Most of the studies use mobile phone collected data, which can hardly scale up due to the deployment limitation of their applications. Our method mine user behavior patterns in mobile Web usage from the mobile network service provider's perspective, which is both comprehensive and large scale.

3 Data Description

The data we use is the mobile Web usage log (HTTP request log) of the cellular network (including 2G and 3G) of a mobile operator. The dataset covers

the geographical range of Beijing, the capital of China. The time range is from Oct. 24 to Nov. 14, 2012. One line in the dataset corresponds to a HTTP request/response pair occurred using cellular network. The structure of the data is shown in Table 1.

Table 1. Field details of the dataset

Name	Data Type	Description
User Id	String	Imsi Id, which is unique identifier of a SIM card
Latitude	Float	Latitude of the cellular tower
Longitude	Float	Longitude of the cellular tower
Request Time	DateTime	Time when the request occurs
Host	String	The domain name of the host
Content Type	String	The ContentType in HTTP

Overview. There are totally **578,134,225** complete records in the dataset. We clean the raw data by record with meaningless field or empty identification. After cleaning, there are **66.823% (386,332,325)** records left.

Users. There are **3,524,929** distinct users appears in the dataset. **40%** of all users make less than 10 requests in a whole week; and over **95%** of all users make less than 1,000 requests in a week. The average number of requests per user is **137.27**.

Hosts. There are totally **363,841** hosts that appear in the dataset. However, only **40.12%** of these hosts are visited by more than 1 user. Top 990 hosts take 97.877% of all records, and top 10 hosts take 58.1% of all records. It can be observed that the websites / services people use on mobile Web is much more concentrated than traditional desktop Web.

Locations. 856 distinct cellulars appear in the dataset, represented by the latitude and longitude of cellular towers. Raw cellulars are less meaningful for behavior modeling, so we cluster them into regions (See Section 4).

4 Geographical Region Discovery

Definition 1. ***Region.*** *A function region of the city is a minimal geographical region that serves a particular set of functions. The function set of the region is the same for most of the citizens.*

As a function region of a city is usually larger and contains several cellular towers, we have to cluster cellular towers into larger function regions, and find characteristics of each region.

4.1 Clustering Algorithm

We use an improved DBSCAN algorithm for cluster cellular towers. The definition of a cluster in DBSCAN is based on the notion of density reachability.

There are two parameters in DBSCAN: ϵ and MinPts. The $\epsilon - neighborhood$ of point p, denoted by $N_\epsilon(p)$, is defined by $N_\epsilon(p) = \{q \in D | dist(p,q) \leq \epsilon\}$. A native approach could require for each point in a cluster that there are at least a minimum number ($MinPts$) of points in a $\epsilon - neighborhood$ of that point. A cluster, which is a subset of the points of the database, satisfies two properties: all points within the cluster are mutually density-connected, if a point is density-connected to any point of the cluster, it is part of the cluster as well[11].

We change the definition of $\epsilon - neighborhood$. In our algorithm, the definition of $\epsilon - neighborhood$ contains three part of restrictions: (A and B as two cellular towers)

- Record similarity sim_r, which represents the similarity of A and B on record amount. We believe that cellulars in the same region would have similar record distributions over different time of day.
- Record migration similarity sim_m. If there is a record of user u at A, followed by another record of u at B in a short time, we call this a *migrating record*.
- Geographical distance dg. $dg_{(A,B)}$ donates the geographical distance (in meters) between A and B.

4.2 Clustering Result

Using our algorithm and the parameter we set, 717 cellular towers are clustered into 126 regions, with the rest 139 as single-cellular regions marked by different colors. In total we get 265 regions. The regions are shown in Figure 1.

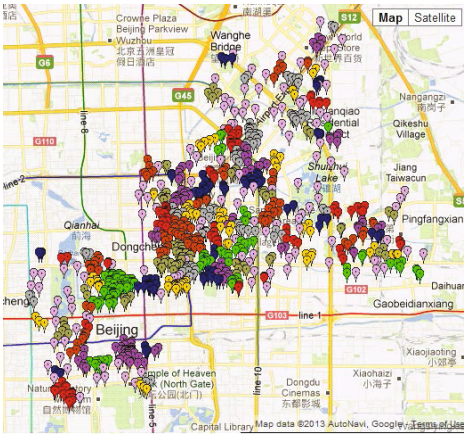


Fig. 1. Regions on map marked by cellular towers in the region

5 Model for User Interest Discovery

First of all, we give the definition of "user interest".

Definition 2. ***User Interest.*** *Interest of a user is a specific type of service / material that the user is interested in. A user may have several different interests, and each interest is associated with a weight, standing for the degree of fondness. On the other hand, user interest types are also website style types, and a website can also serves different user interest types.*

In order to extract latent user interests from their Web usage log, we propose a probabilistic topic modeling method based on LDA (Latent Dirichlet Allocation). In our model, the extracted latent layer represents a *User Interest* defined in Definition 2.

Topic Model is commonly used in text mining for discovering abstract "topics" in a set of documents. LDA (Latent Dirichlet Allocation)[12] is a commonly used topic model currently, and it has also been applied in discovering user behavior patterns[13]. 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 bag-of-website representation of a user as a document, and propose a probabilistic topic model for user behavior modeling.

5.1 Website Discovery

Firstly, we will discuss the details about how we extract website from raw HTTP log of mobile Web. There are some formal definitions of the concepts used.

Definition 3. ***User:*** *A user if uniquely identified by a UserId in the dataset. It corresponds to a real person (Imsi number of the SIM card) using the mobile Web, regardless of what devices are used.*

Definition 4. ***Host:*** *A host is the domain name of the HTTP request. It may or may not be the address that is directly requested by the user / app.*

Definition 5. ***Website.*** *A website stands for a unit that provides material/ services to users as a independent entity, using several different hosts.*

Websites and hosts do not have strong corresponding relationships. We use a simple strategy to cluster hosts into websites. We treat each host as a vector $H_i :< C_{h_i,u_1}, C_{h_i,u_2}, ..., C_{h_i,u_n} >$, where C_{h_i,u_j} is the number of requests user j makes to host i. Then we calculate pairwise cosine-similarities between all hosts pairs using the vectors. Then we merge the websites with higher similarity.

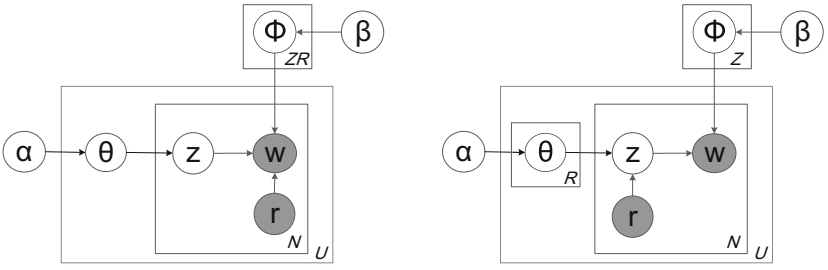
5.2 Bag-of-Website Representation of User

We treat a user as a document, and websites as words. Each occurrence of a website in a user is associated with a region, corresponding to the location where the user visits the website.

Since raw HTTP log may not truly reflect user behavior, we do some transformation for forming the bag-of-websites of a user. we firstly divide a day into 48 half-hour time slot and treat each slot as a period of user behavior[13]. Then for each user u and each time slot t, the top 3 most visited websites are treated as been "visited". A user is then represented as $U = \{< w_{ui}, r_{ui}, c_{u,w_{ui},r_{ui}} >\}$, where $c_{u,w_{ui},r_{ui}}$ is the number of times that $< w_{ui}, r_{ui} >$ is visited by u.

5.3 User-Region-Interest (URI) Model

We combine users, geographical regions, user interests and websites in unified generative models. Distinguished by the practical interpretation of the "interest" layer and the strategies to engage *region*, we propose two different generative processes and two corresponding URI models.



(a) URI1 (b) URI2

Fig. 2. Plate representation of User-Region-Interest (URI) model

URI Model 1. In the first model (URI1), the latent "interest" layer is treated as frequent user Web-using patterns, which is more related with users. Each user has a "interest" distribution, regardless of location; and within each "interest", there is a website distribution specific to each region. The generative process of URI1 is as follows (Figure 2(a)):

1. For each user u, draw θ_u from Dirichlet prior α;
2. For each interest z, draw $\phi_{(z,r)}$ for all $r \in R$ from Dirichlet priors β_z;
3. For each website appearance w_{ui} in user u:
   - draw a topic z_{ui} from a multinomial distribution θ_u;
   - draw a term for w_{ui} from multinomial distribution $\phi_{(}z_{ui}, r_{ui})$.

We use Gibbs sampling to estimate the model parameters, following [14]. For simplicity, we take fixed values for hyperparameters α and β (i.e. $\alpha = 50/T$, $\beta = 0.01$). We use Gibbs sampling to estimate the posterior distribution on w, z and r, then use the result to estimate θ and ϕ. The posterior probability can be calculated by :

$$P(z_{ui}|\mathbf{z_{-ui}}, \mathbf{w}, \mathbf{r}, \alpha, \beta) \propto \frac{m_{u,z_{ui}}^{-ui} + \alpha_{z_{ui}}}{\sum_z (m_{u,z}^{-ui} + \alpha_z)} \frac{n_{z_{ui},r_{ui},w_{ui}}^{-ui} + \beta_{w_{ui}}}{\sum_w (n_{z_{ui},r_{ui},w}^{-ui} + \beta_w)} \tag{1}$$

$m_{u,z}$ denotes the times that interest z is assigned to user u, and $n_{z,r,w}$ denotes the times that interest z is assigned to user website w in region r. Superscript $-ui$ means that the quantity is counted excluding the current instance.

After the sampling convergences, the multinominal distribution parameters θ and β can be estimated as follows:

$$\theta_{u,z} = \frac{m_{u,z} + \alpha_z}{\sum_{z'} (m_{u,z'} + \alpha_{z'})} \tag{2}$$

$$\phi_{z,r,w} = \frac{n_{z,r,w} + \beta_w}{\sum_{w'} (n_{z,r,w'} + \beta_{w'})} \tag{3}$$

URI Model 2. In the second model (URI2), the "interest" layer is treated as commonly appeared website clusters, which is more related with websites. Each "interest" has a unified website distribution regardless of location, while each user has a "interest" distribution specific to each region. The generative process of URI2 is as follows (Figure 2(b)):

1. For each user u, draw $\theta_{(u,r)}$ from Dirichlet prior α_u for all $r \in R$;
2. For each interest z, draw ϕ_z from Dirichlet priors β;
3. For each website appearance w_{ui} in user u:
   - draw a topic z_{ui} from a multinomial distribution $\theta_{(u,r_{ui})}$;
   - draw a term for w_{ui} from multinomial distribution $\phi_{z_{ui}}$.

Like in URI1, we use Gibbs sampling to estimate the model parameters, and the posterior probability can be calculated by :

$$P(z_{ui}|\mathbf{z_{-ui}}, \mathbf{w}, \mathbf{r}, \alpha, \beta) \propto \frac{m_{u,r_{ui},z_{ui}}^{-ui} + \alpha_{z_{ui}}}{\sum_z (m_{u,r_{ui},z}^{-ui} + \alpha_z)} \frac{n_{z_{ui},w_{ui}}^{-ui} + \beta_{w_{ui}}}{\sum_w (n_{z_{ui},w}^{-ui} + \beta_w)} \tag{4}$$

θ and β can be estimated as follows:

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

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

6 Experiments and Discussion

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

6.1 User Interest Discovery

We use $U = 469,297$ users, $W = 924$ websites, and $R = 265$ regions to evaluate the models. We have run both URI model and LDA model (using traditional LDA generative process) on same datasets. For comparison, we keep track of the corresponding region of a word when it is assigned a interest during the LDA sampling process, and use Equation 3 to calculate a $K \times R \times W$ Φ parameter matrix, and follow Equation 5 to calculate a $M \times R \times K$ Θ matrix, where M is number of users in the training set. The original Φ and Θ in LDA is called $\Phi 0$ and $\Theta 0$ here. We use LDA0 to donate the LDA model using $\Phi 0$ and $\Theta 0$, and LDA1 using Φ and $\Theta 0$, LDA2 donates the model using $\Phi 0$ and Θ.

Metrics. We use average Jensen Shannon divergence among topics to evaluate the quality of latent "interests". Jensen Shannon divergence (JSD) is commonly used for measuring similarity between probabilistic distribution describing the same random variable. Using different $\mathbf{\Phi}$ and $\mathbf{\Theta}$ matrix as parameters, we can calculate divergence for each model. For LDA1 and URI, in which website distribution within "interest" is specific to regions, JSD is calculated as:

$$JSD = \frac{\sum_{i=0}^{T} \sum_{j=i+1}^{T} \sum_{r=1}^{R} jsd(\mathbf{\Phi_{i,r}}, \mathbf{\Phi_{j,r}})}{\frac{T(T+1)}{2} \times R} \tag{7}$$

For the rest of the models (LDA0, LDA2 and URI2), JSD is calculated as:

$$JSD = \frac{\sum_{i=0}^{T} \sum_{j=i+1}^{T} jsd(\mathbf{\Phi 0_i}, \mathbf{\Phi 0_j})}{\frac{T(T+1)}{2}} \tag{8}$$

The estimated distribution of websites for users can be used to represent their (potential) preference about websites, which can be used for understanding their behavior pattern, and for recommendation. We use perplexity to evaluate the performance of these proposed model probability distributions. Perplexity is commonly used how well a proposed distribution can predict samples from the target distribution. It is defined as

$$perp(\widetilde{\mathbf{p}}, \mathbf{q}) = 2^{H(\widetilde{\mathbf{p}}, \mathbf{q})},$$

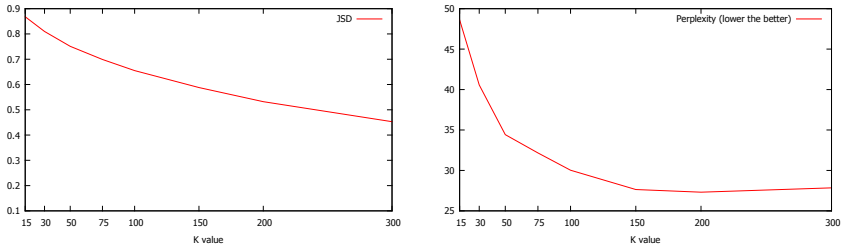
where the exponent $H(\widetilde{p}, q)$ is the cross-entropy:

$$H(\widetilde{\mathbf{p}}, \mathbf{q}) = -\sum_{x} \widetilde{p}(x) log_2^{q(x)}$$

Smaller perplexity means better prediction. The averaged perplexity over all users is defined as the perplexity of the proposed distribution.

Topic number. We have tried different numbers of user interests (topics) K. Generally, reproduction performance improves(perplexity drops) when increase K, and stays stable after K reaches a certain value. For topic divergence, JSD decreases with K. The experiment result is shown in Figure 3(a) and 3(b).

Since , Considering both generalization and reproduction ability of the model, JSD and $Perplexity$ are our major concern, we set $K = 50$ in our model.



(a) JSD with different K value (URI model)

(b) Perplexity of $\mathbf{p}$ with different K value (URI model)

Fig. 3. The experiment results

Table 2. Performance of LDA and URI models

	D0		D1		D2	
Model Type	JSD	Perplexity	JSD	Perplexity	JSD	Perplexity
LDA0	0.040	203.805	0.079	183.021	0.064	305.282
LDA1	0.211	181.824	0.307	89.910	0.247	158.881
LDA2	0.211	181.824	**0.990**	69.853	**0.992**	64.621
URI1	0.925	**98.229**	0.751	**34.423**	0.818	**54.079**
URI2	**0.979**	103.498	0.981	73.668	0.974	121.978

The performance of different methods on the the whole dataset and two random-generated subsets are shown in Table 2.

It can be seen that, generally, the URI model out-performs LDA significantly. URI Model 2 and LDA 2 discover better latent "interest", reflecting website clusters, while URI Model 1 can better estimate Web using behavior for each user.

7 Conclusion

In this paper, we have presented a framework for mining user behavior pattern in mobile Web usage. A method is proposed to cluster cellular towers into regions according to their mobile Web traffic log. We have also presented a probabilistic topic modeling method to extract latent "user interests" from mobile Web usage log with location. We have applied the proposed framework to a real-world large scale dataset from a Beijing, capital of China, covering more than 3 million users.

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. Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explor. Newsl. 2(1), 1–15 (2000)
2. 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)
3. 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)
4. 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)
5. 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)
6. Cui, Y., Roto, V.: How people use the web on mobile devices. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, pp. 905–914. ACM, New York (2008)
7. Tseng, V.S., Lin, K.W.: Efficient mining and prediction of user behavior patterns in mobile web systems. Information and Software Technology 48(6), 357–369 (2006), WAMIS 2005 Workshop
8. Phatak, D., Mulvaney, R.: Clustering for personalized mobile web usage. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2002, vol. 1, pp. 705–710 (2002)
9. 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. 1–27. ACM, New York (2010)
10. Verkasalo, H.: Contextual patterns in mobile service usage. Personal and Ubiquitous Computing 13, 331–342 (2009)
11. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press (1996)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
13. 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)
14. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)