# Mining Mobile Users' Interests Through Cellular Network Browsing Profiles

Fan Yan[1], Yunpeng Ding[1], and Wenzhong Li[1,2(✉)]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{151220139,dingyunpeng}@smail.nju.edu.cn, lwz@nju.edu.cn
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China

**Abstract.** Mining mobile users' interest is very important for numerous of commercial applications such as product recommendation, personalized advertisement, precision marketing, etc. In this paper, we proposed a novel clustering approach for semantic mining from cellular network browsing profiles based on the topic model. We treat each URL as a word and the user's browsing history as a document, and adopt the Latent Dirichlet Allocation (LDA) model to represent the web browsing interest of mobile users. We further used K-means to cluster the users into several groups according to their topic similarities, and apply a feature ranking approach to explain the sematic meaning of the clustering results. The performance of the proposed approach is verified on a dataset from a telecom operator, which explains users' interests well in the clusters.

**Keywords:** LDA · Clustering · User persona · Interest mining

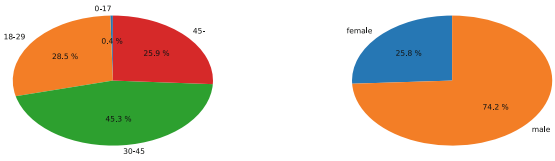
## 1 Introduction

With the rapid development of mobile network, people are getting used to do more daily works and entertainments on their smartphones. Thus a growing number of analyses in the recent years have sought to explore mobile users related data to analyze their behavior and preference.

App using trace is one of the most popular source materials for researchers. Zhao et al. [1] collected one month of app usage from 106,762 Android users. They found that users are heterogeneous in smartphone usage, and proposed a 2-step clustering method to cluster users into 382 distinct types. Xu et al. [2] got app usage at a national level and analyzed how, when and where they are used. Some interesting patterns of using preference were revealed in this paper.

Comparing with app using trace, users' browsing profiles made up with URLs provide more semantic for researchers to mine, since URLs show what users really get in detail. Traditional method to mine user interests from surfing profiles is


S. Chellappan et al. (Eds.): WASA 2018, LNCS 10874, pp. 806–812, 2018.
https://doi.org/10.1007/978-3-319-94268-1_71




(a) age (b) gender

**Fig. 1.** Age and gender distribution of users.

keywords mining. Researchers try to find specific keywords in URLs to classify a record to a certain category. In that way, the keywords vocabulary should be large enough to deal with the complexity of reality, and building such vocabulary is labor-intensive and time-consuming.

To overcome those shortcomings, Giri et al. [3] proposed the unsupervised topic model (LDA) to mine the interests of the cellular users based upon their browsing profiles. However, they did not dig more from the results of LDA nor mining from time information of clicks. Extracting click stream is also an useful method to solve this problem, such as Zhang et al. [4] and Wang et al. [5].

In our research, we cooperated with a telecommunication company to get anonymous surfing profiles. To mining abundant semantic meanings, we proposed a novel clustering approach for semantic mining based on the topic model, which is quite useful to discover the hidden semantic structures in a text body. Specifically, we treat each URL as a word and each user's browsing history as a document, and adopt a famous topic model, Latent Dirichlet Allocation (LDA), to represent the web browsing interest of mobile users. To better understand the behaviors of similar users, we used K-means to cluster the users into several groups. Then we apply a feature ranking approach to explain the sematic meaning of the clustering results.

## 2 Dataset Overview

The dataset used in this paper is from a telecom operator of China, which contains 945 users' browsing profiles from March 15th to March 21st in 2016, together with some basic information, such as their genders and ages. The users are randomly draw from the Jiangsu province. As shown in Fig. 1, users' age varied from 10 to 79, and most of them are between 18 and 45. The gender ratio is not balanced in our dataset as the male users account for 3 fourths. Take a unique URL request as a record, the whole dataset contains about 3148110 records, about 3310 records per user in average. The data is totally anonymous with the virtual number to identify users. Those records contain 12792 different URLs. Except for URL and identifier, session start time, session end time, user agent, target IP and network traffic-related information are also included.

## 3 LDA Model and Implementation

LDA is short for Latent Dirichlet Allocation, which is an unsupervised topic model presented by Blei et al. in 2003 [7]. It assumes each document in the corpus as a mixture of topics and find out the possibility of the documents to be in each topic. In our case, LDA regards each users' browsing profile as a document. The model will find the hidden topic of each URL automatically.

We reviewed the work of Giri et al. [3], and used the same LDA tool–Mr. LDA [6]. It's an open-source package for scalable, multilingual topic modeling using variational inference in MapReduce.

LDA treats a document as a bag of words. Thus the order of words will not be taken into analyzing process. To capture time information, we propose to divide the users' profiles into four intervals according to its sending time. We divide the whole day into four stages: 0:00 to 5:59 is early morning, 6:00 to 11:59 is morning, 12:00 to 17:59 is afternoon, and 18:00 to 23:59 is evening. We assume that users' interests varied in different periods of time.

**Stop Words and Informed Prior.** Stop words are those URLs which are nearly visited by everyone. Those URLs make topics less distinctive as they may appears in every topic. We sorted all the URLs by frequency and then picked the top 100 URLs as stop URLs out. Ignoring the stop words in corpus indeed helps to get a better result, but we have to admit that we sacrificed some semantic.

After removing the stop words, how to determine the semantic meaning of each topic is still challenging. Mr. LDA provides an extension block named Informed Prior, which helps to force some similar words to be grouped into the same topic. For example, if we want URLs related to Reading/News to be grouped into the same topic, we collect some famous websites in this area and picked their URLs for topic one, like [hupu.com, sina.com, ifeng.com, baidunews.com, toutiao.com, ucweb.com ...]. When a URL $\beta$ belongs to the list, we set a higher initialization value as its informed prior to this topic. In that case, we abandon the default symmetric prior provided by the model and get a better initialization.

**Results of LDA.** Now we have divided users' browsing profiles into four parts according to time, cleaned them up by removing stop words, and added enough common URLs for each topic. The only parameter we need to assign is the number of topics, which we chose to set as 7 after comparing the likelihood with some other parameters. From topmost URLs of each topic, we can find the semantic meanings and tick labels for users. Here is a table shows these top URLs with the score representing the probability of locating in this topic (Table 1).

As each user's profile is divided into 4 parts on different periods, and each part has the probability of locating in different topics, we finally got four 7-dimension vectors for each user. We combined them as the user feature vector. We also normalized the vector to make it easier for clustering.

**Table 1.** The top URLs in different topics.

| Topic | Top URLs in the topic |
|---|---|
| Education/working | Imail.qq (6.92), kmail.com (6.91), yuansouti (6.90), dsp-impr2.youdao (6.91), log.yex.youdao (6.90) |
| Ecommecial | Talaris2-lbseleme (7.47), m.360buy (6.95), wp.360buyimg (6.93), wq.jd (6.92), pic.alipay.objects (6.92) |
| Entertainment | Video.qpic (6.96), vhot.hdnion.videocdn.qq (6.95), cm.passport.iqiyi (6.93), p1.music126 (6.92), ccstream.qqmusic.qq (6.91) |
| Game | Report.game.center (7.27), api.coolmart.net (7.11), games.qq (6.93), game11h5 (6.93), gift.gamecenter.qq (6.93) |
| Reading/news | Zzm.sohu (7.09), p2p.statp (7.06), d.ifengimg (7.03), stadig.ifeng (6.99), vs3.wxctu3.ucweb (6.98) |
| Social communication | Cn.battleofballs (7.00), ossweb-img.qq (6.91), oth.strmdt.qq (6.91), api.place.weibo (6.91), cgi.connect.qq (6.91) |
| Junk/ad | Router.g0.push.leancloud (6.92), hispace.clthicloud (6.91), api-webrp-analytics.cloudtoast (6.91), icsnssdk (6.91), imagebox.xiaomi (6.91) |

## 4 User Clustering

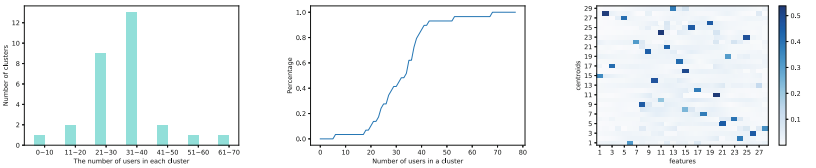
The LDA model analyzes the original user data and provides us with normalized user vectors. To depict the similarities and personalities of users better, we choose to use K-means, a famous clustering method which based on measuring distances between samples, to cluster users into different clusters.

We combined several metrics to measure the performance of different value of K. The following equation is used to calculate the score of a certain K's clustering result.

$$Score = SE + DI + \frac{UserNum - MaxClusterNum}{UserNum} \tag{1}$$

In the equation, $SE$ means Shannon Entropy and $DI$ means Dunn index. We take the number of users in each cluster and divide it by the total number of users as the possibility of each cluster to calculate Shannon Entropy. Dunn index is a typical internal index used to evaluate clustering performance. Notice that the last item of the equation is not a typical index, we use the total number of users minus the number of users in the biggest cluster to penalize results which most users are grouped into few clusters.

According to our experiments, 29 was chosen as the best K for clustering. Figure 2(a) and (b) show the distribution of users in clusters. Centroids of each cluster are the intuitive materials to describe clusters' features. We visualize 29 centroids in the heat map to see how we distinguish one cluster to another. The




(a) Numbers of users in clusters.

(b) Cumulative Distribution Function(CDF) of the number of users.

(c) Centroid heatmap.

**Fig. 2.** The distribution of users in clusters. (Color figure online)

result is showed in Fig. 2(c). We can easily tell the difference by the highest dimension of each centroid.

## 5 Analysis of Clustering Results

In this part, we will explain our approach to select the most notable features and take the biggest cluster as an example to show how we give a cluster a label as personas. The following features we defined will help us to analyze the result.

**Salient Features.** As shown in the heat map Fig. 2(c), each centroid contains one or two dimension shown as deep blue, which means the value is significantly larger than other dimensions. Thus we chose those features with highest values as salient features.

**Idiosyncratic Features.** We got inspiration from the work of Zhao et al. [1], especially the part used to distinguish clusters. They call it idiosyncratic features.

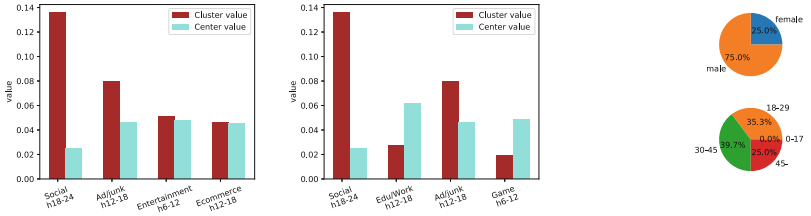
$$r_{i,j} = \frac{|c_{i,j} - u_j|}{\max_{j \in J(|c_{i,j} - u_j|)}} \tag{2}$$

In this equation, $c_{i,j}$ is the value of $j^{th}$ dimension of $i^{th}$ cluster's centroid, $u_j$ means $j^{th}$ dimension of the "average" user. The largest $r_{i,j}$ means $j^{th}$ feature (or dimension of the vector), is the main contributor for the cluster to be distinctive.

**Example of Cluster Results:** The 12th cluster is the largest cluster of the result, which contains 68 users. By visualizing the age and gender distribution of this cluster (Fig. 3(c)), we found the gender distribution is similar to the whole dataset. While the ratio of users aged from 18 to 29 in this cluster is significantly larger than average. That is to say, users are younger than average age in this cluster. The top 4 salient features for this cluster is Social/h18–24, Ad/junk/h12–18, Entertainment/h6–12 and E-commerce/h12–18. Only the first feature, Social/h18–24, is significantly larger than the average value. And

top 4 Idiosyncratic Features are Social/h18–24, education/working/h12–18, Ad/junk/h12–18 and Game/h6–12.

The salient features only depict they love using social applications as other dimensions are similar to the average user. According to the idiosyncratic features, we can see these users are not fond of games and education/working applications since they are the main features distinguish them from other clusters, and values of these features are low. We can stick the label "social lovers" to this cluster.




(a) Salient features (b) Idiosyncratic features (c) Age and gender distribution

**Fig. 3.** Analysis of the 12th cluster (68 users).

## 6 Conclusion

In this paper, we proposed a method to mine users' interests from cellular network browsing profiles. We used the topic model (LDA) to extract the interest and find the semantic meaning hidden behind the URLs. Using LDA, the interest of each user can be represented by a vector, which is used for clustering similar users using the K-means algorithm. The methods we provide in this paper is useful for telecommunication companies to understand their users and build personas, which is helpful for product recommendation and personalized advertisement.

**Acknowledgements.** This work was partially supported by the National Key R&D Program of China (Grant No. 2017YFB1001801), the National Natural Science Foundation of China (Grant Nos. 61672278, 61373128, 61321491), the science and technology project from State Grid Corporation of China (Contract No. SGSNXT00YJJS1800031), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Sino-German Institutes of Social Computing.

## References

1. Zhao, S., Ramos, J., Tao, J., Jiang, Z., Li, S., Wu, Z., Pan, G., Dey, A.K.: Discovering different kinds of smartphone users through their application usage behaviors. In: Proceedings of UbiComp 2016, pp. 498–509. ACM (2016)
2. Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., Venkataraman, S.: Identifying diverse usage behaviors of smartphone apps. In: Proceedings of IMC 2011, pp. 329–344. ACM (2011)

3. Giri, R., Choi, H., Hoo, K.S., Rao, B.D.: User behavior modeling in a cellular network using latent Dirichlet allocation. In: Corchado, E., Lozano, J.A., Quintián, H., Yin, H. (eds.) IDEAL 2014. LNCS, vol. 8669, pp. 36–44. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10840-7_5
4. Zhang, X., Brown, H.F., Shankar, A.: Data-driven personas: constructing archetypal users with clickstreams and user telemetry. In: Proceedings of CHI 2016, pp. 5350–5359. ACM (2016)
5. Wang, G., Zhang, X., Tang, S., Zheng, H., Zhao, B.Y.: Unsupervised clickstream clustering for user behavior analysis. In: Proceedings of CHI 2016, pp. 225–236. ACM (2016)
6. Zhai, K., Boyd-Graber, J., Asadi, N., Alkhouja, M.L.: Mr. LDA: a flexible large scale topic modeling package using variational inference in mapreduce. In: Proceedings of WWW 2012, pp. 879–888. ACM (2012)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)