Data driven user feature prediction in mobile applications based on multi-channel CNN

Yuanbang Li*, Chi Xu, Shi Dong, Laihang Yu
College of Computer Science and Technology, Zhoukou Normal University, Zhoukou, China

ABSTRACT

With the continuous development of mobile positioning technology and smart phones, users can use smart phones to obtain and share location information of themselves and various surrounding points of interest (POI) anytime, anywhere, and share their own activity information, thus forming a location-based social network (LBSN). A large amount of user data is generated in LBSN. How to quantitative analyze the effects of context to manager's view, how to extract the hidden user feature through the analysis of user data is of important research significance for the intelligent analysis of user characteristics. In this paper, first, a muti-dimensional user feature construction method is proposed, which extracts user feature from different influencing factors. Second, the fitness of user to a feature is analyzed. Third, a unified model is used to characterize this applicability. The method can promote the transformation from user data to user feature and help solve the problem of "explosive data but poor knowledge". Experimental verification shows that the method is feasible and can realize the mining of muti-dimensional feature of users.

Keywords: Impact analysis, multiple user feature, applicable analysis, muti-channel CNN

1. INTRODUCTION

With the popularization of mobile phones, it is convenient for users to obtain and share various data about themselves and their surroundings anytime and anywhere, and obtain various intelligent services, which has promoted the formation of LBSN. Mobile phones are widely used in daily lives and works.

How to obtain user feature is important to analyse user characteristics and provide intelligent services for them. However, in LBSN, the acquisition of user feature faces huge challenges, mainly in:

- User openness

The LBSN system is usually open to use in practical applications. Users can download and install freely on their mobile terminals. In most cases, the system administrator does not even know whom the user is. Therefore, it is difficult to directly communicate with users to obtain their feature.

- Limitations of user description ability

Blomberg et al. pointed out that users cannot clearly describe their requirements for situations that are not directly accessible[1]. In LBSN, the values of context dimensions such as the location and time of users are diverse, and this diversity is not completely predictable, so it is very difficult for users to express their contextual features clearly[2].

Fortunately, a large amount of user check-in data will be generated in LBSN. These data include not only the context, but also semantic information related to POI in the activity. Therefore, user features that users cannot clearly describe themselves can be reflected in the data. Lots of works have focused on the construction of user's feature models[3]. Different users are suit to different feature models. For example, some users usually visit occupational and surrounding POI on weekdays, and often visit store services, art and entertainment POIs on weekends[4]. while other users usually visit different types of POI at different distances from home.

Based on the analysis above, a data driven user feature construction and analysis method is proposed in this study. The contributions in this paper including

1. The effect of context on manager' views is quantifying analyzed.

*lybang@whu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122601M · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637606

2. Diversified user feature models are constructed based on their check-in data.

3. Fitness of the feature model to user is determined.

4. A multi-channel CNN is designed to depict the fitness of features to a user.

In this manuscript, Section 2 gives the related works. Section 3 described the proposed method. Section 4 introduces the experimental evaluation, and Section 5 draws a conclusion.

2. RELATED WORKS

User feature construction is an important research direction in LBSNs. These construction methods including two categories: explicit or implicit methods.

Explicit methods construct user features through direct communication with users[5], which is intuitive and with a high accuracy. However, when the environment around the users is complex, it is difficult for user to describe their features clearly.

Implicit methods use multiple analyses methods to construct user features automatically. These methods including four categories: content-based, geographical-based, temporal-based, and social-based feature extraction methods[6].

The first one constructs user features through related content of users, such as position or comments[7]. Geographical-based method aims to reveal users' feature related to distance. Such as naive Bayesian network is used to depict the relationship between the venue that a user has been visited and the distance[8]. Temporal-based method aims to reveal user's temporal related feature using data mining or machine learning methods[9]. Social-based methods assuming that users' feature may be similar to their friend. Therefore, a set of the features shared by their friends are used to infer current user's features[10].

Of course, more than one of these influence factors may combined to describe user features[11].

Although many works devote to construct users' features, few of these works pay attention to the applicable of feature models to a user. In this study, in the basis of the construction of multiple user feature models, an algorithm is proposed to distinguish the user set suit to a specific feature model, and a multi-channel CNN is designed to describe the suitable of different users to multiple feature models.

3. METHOD

3.1. Definitions

Definition 1: UCS

USC (User Context Set) is an attributes set to describe the context related to user activities.

$$UCS=\{UC^{i}\}$$

Definition 2: VS

VS (View Set) is a perspectives set that analysts are interested in.

$$VS=\{V^{j}\}$$

Definition 3: User context view feature set (UFS)

UFS (User context view Feature Set) is a feature set to describe user's characteristic generated from user's check-in data.

$$\mathrm{UFS}=\{\mathrm{UCVF}^{\mathrm{ij}}||0=<|\mathrm{i}|<|\mathrm{UC}|,0=<|\mathrm{j}|<|\mathrm{VS}|\}$$

and:

$$UCVF^{\mathrm{ij}}=\begin{pmatrix} \mathrm{ucvf}^{\mathrm{ij}}{}_{11} & \dots & \mathrm{ucvf}^{\mathrm{ij}}{}_{1|V^{j}|} \\ \vdots & \ddots & \vdots \\ \mathrm{ucvf}^{\mathrm{ij}}{}_{|C^{i}|1} & \cdots & \mathrm{ucvf}^{\mathrm{ij}}{}_{|C^{i}||V^{j}|} \end{pmatrix}$$

3.2. Overview of the method

The overall framework of the method is shown in Figure 1. The method is based on user check-in data. First, the set of view and user context is determined by the analyst, and the effect of the impact of context to the observer's view is quantify analysed.

Secondly, vectored user features are constructed from user check-in data. Thirdly, applicability analysis is conducted through difference value. Finally, a muti-channel CNN is designed to characterize the applicability of users to different features and used to predict user features.

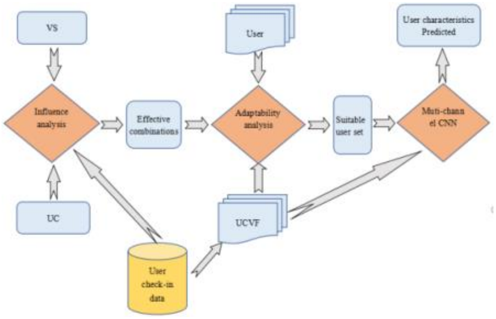

Figure 1. Overview of the method.

3.3. Information entropy gain-based influence analysis

The cardinal number of the influence combination |ES| is the product of |UCS| and |VS|. When |UCS| and |VS| increases, |ES| will increase rapidly, which is disadvantageous to storage and analysis. And in real life, not all user context set have a significant impact on each view. Therefore, the quantitative analysis of the impact of user context on the view is of great significance.

The concept of entropy can be used to describe the degree of chaos within the data, as described in equation (1), n describe number of categories and p_i means the probability of belonging to category i.

Entropy gain can be used to measure the effectiveness of data partitioning. The information gain and the gain rate are calculated as shown in equations (2) and (3), where A represents the attribute that divides the sample set S, Values represents the value set of the attribute A, v represents a value in the Values set, and Sv represents the sample set corresponding to the value v after the division.

$$Entropy(S) = -\sum_{i=1}^{n} p_i * log\, p_i \tag{1}$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values}(A) \frac{S_v}{S} Entropy(S_v) \tag{2}$$

$$Gain\_ratio(S,A) = \frac{Gain(S,A)}{Entropy(S)} \tag{3}$$

Based on the analysis above, this study uses information gain rate to find the UC^i that has a significant impact on V^j based on the data set.

3.4. UFS construction

According to Definition 3, each $UCVF^{ij}$ in UFS is a matrix of $|UC^i|$ times $|V^j|$ dimensions, which is constructed as following: first, $UCVF^{ij}$ is initialized; second, the check-in data is checked iteratively, the value of v of context value *uc* is found; third, modify the values in the matrix where the rows and columns corresponding to *uc* and *v*.

3.5. Applicability analysis

The process of the applicability analysis is described as following: first of all, a fixed time unit of user feature model is established; and then, difference value is generated for each feature of u; lastly, take the feature of the minimum difference value as the suit one, because when the value is the smallest, it is demonstrated that the user is most suitable for the feature.

3.6. Unified model—muti-channel CNN

Multichannel neural networks can effectively describe the local salience features of data, identify and analyse them, and then stack these different channels using a deep structure to support the fusion of multiple salient features. This feature is suitable for describing the user's adaptability to muti-perspective features; therefore, this study designs a muti-channel convolutional neural network to analyse the user's personalized features. The basic network structure is illustrated in Figure 2.

The network contains |ES| channels, and the input is user set US_UCVF^{ij}, which is suitable for the UCVF^{ij} model and the matrix UCVF^{ij} for these users. After learning the user and the user's matrix through different channels, the network can predict the user's activities during a given context.

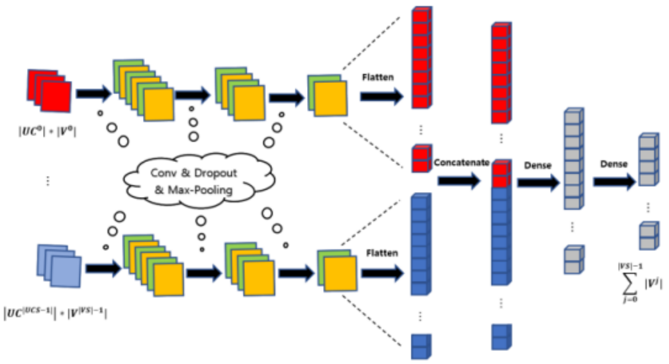

Figure 2. Muti-channel CNN structure diagram.

4. EXPERIMENT

3.7. Date sets of experiment

We use two datasets in the experiment to verify the effectiveness of the method, which are labelled as dataset1 and dataset2. The dataset1 is from reference[12]. The dataset2 was constructed by randomly selecting 5,000 users from the dataset used in reference[13]. Detailed information of the datasets is shown in Table 1.

Table 1. Information of the data sets.

	User number	POI number	Check-in times
Dataset1	1083	38333	227428
Dataset2	5000	359036	1472935

3.8. Research question and evaluation plan

RQ 1: Is the assumption established that user is applicable to a feature model when the difference value in UCVF^{ij} is small?

RQ 2: Does applicability analysis help to improve the accuracy to predict user behaviours.

RQ 3: Does the unified model improve the prediction accuracy? How does it compare with existing methods?

Accuracy of top K is used to evaluate the effectiveness of the method, which is shown in equation (4):

$$Accuracy@K = \frac{|\{u, l, t, a\}|a \in P_{u,l,t}(K), (u, l, t, a \in TS)|}{|TS|} \quad (4)$$

where *u* refers to user, *l* refers to location, *t* refers to time, *a* refers to an activity, and *TS* stands for test set.

4. EXPERIMENT RESULTS AND DISCUSSION

In this experiment, UCS = {UC^t, UC^d}, where *t* represents time and *d* represents distance. |UC^t|=24 because the time is depicted in hours. |UC^d|=4 because the distance is divided into four levels, which are within 1 kilometre, between 1 and 10 kilometres, between 10 and 30 kilometres, and more than 30 kilometres, so.

Based on the elements in the data, VS = {V^r, V^c}. The parameter *r* is a representation of root category, and the parameter c is a representation of category. | V^r | =9 and | V^c | =65 because the number of root categories and categories of POI in the experiment is 9 and 65. Therefore, UFS = {$UCVF^{\text{time-root category}}$, $UCVF^{\text{time-category}}$, $UCVF^{\text{distance-root category}}$, $UCVF^{\text{distance-category}}$}, which correspond to 24*9 matrix, 24*65 matrix, 4*9 matrix, and 4*65 matrix.

After UFS is constructed, the user set suit for each feature is analyzed, as shown in Table 2.

A multi-channel CNN is designed to depict the adaptability of a $UCVF^{ij}$ to a user, and the network is used to predict user activities. Experimental results were presented and discussed as follows.

RQ 1: Is the assumption established that user is applicable to a feature model when the difference value in $UCVF^{ij}$ is small?

The difference value was divided at intervals of 10, and the TOP K accuracy rates was analysed. The results are shown in Figures 3 and 4.

Table 2. Users suit to different $UCVF^{ij}$.

	User number	**$US\_UCVF^{\text{time-root category}}$**	**$US\_UCVF^{\text{time-category}}$**	**$US\_UCVF^{\text{distance-root category}}$**	**$US\_UCVF^{\text{distance-category}}$**
Dataset1	1083	526	82	429	46
Dataset2	5000	1396	1413	1410	781

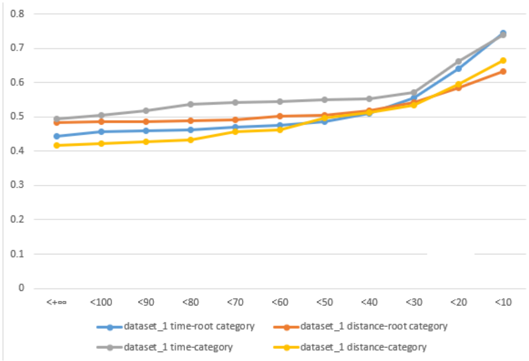

Figure 3. Accuracy rate of dataset1.

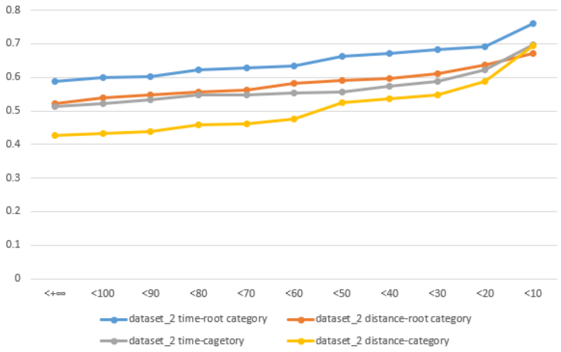

Figure 4. Accuracy rate of dataset2.

These figures show that in both data sets, when the difference value decrease, the accuracy increases, which means the assumption is established.

RQ 2: Does applicability analysis help to improve the accuracy to predict user behaviours.

To answer this question, $UCVF^{\text{time-root category}}$, $UCVF^{\text{time-category}}$, $UCVF^{\text{distance-root category}}$ and $UCVF^{\text{distance-category}}$ were used separately to predict user's activity, and the accuracy rate is generated. After that, uses the suit $UCVF^{ij}$ to describe corresponding users, and the accuracy rate is generated using equation (4), results are depicted in Figure 5.

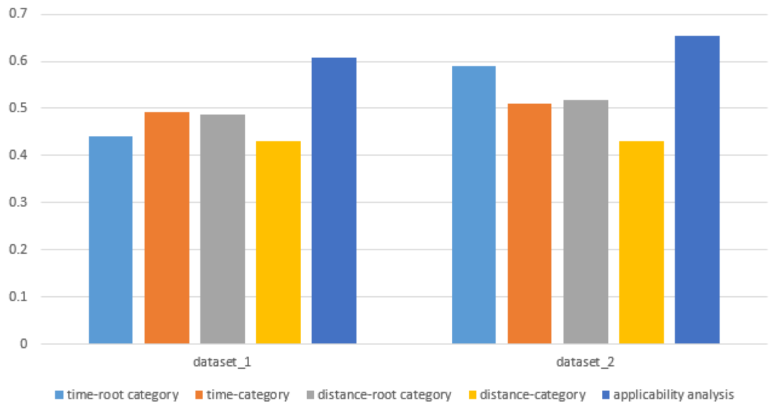

Figure 5. Effectiveness of applicability analysis.

The result shows that in both data sets, the predication of user's suited UCVF^{ij} is of higher to using each UVCF^{ij} separately.

RQ 3: Does the unified model improve the prediction accuracy? How does it compare with existing methods?

- Baseline

User activity prediction can be divided into the next prediction and any time prediction in terms of timeliness, coarse-grained and fine-grained predictions in terms of the granularity[14].

This research belongs to any time and coarse prediction in terms of timeliness and granularity. We adopt the HOSVD method, PFR method, PCLR method and STAP method as baselines for comparison[12, 15-17].

- Results

The proposed method was compared with baseline methods according to Top-K accuracy. The value of K in the Top-K accuracy rate formula is 1 and 5. the results are as follows in Figures 6 and 7.

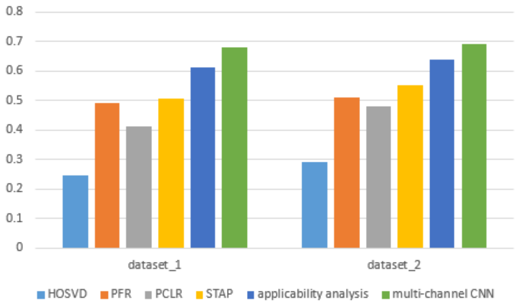

Figure 6. Accuracy comparison of Top 1.

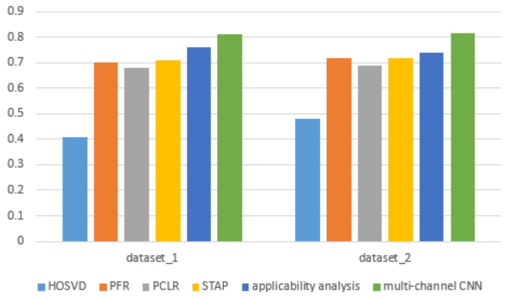

Figure 7. Accuracy comparison of Top 5.

From the graph, we can see that the adaptive analysis method and the muti-channel CNN method outperforms the baseline methods in both dataset1 and dataset2.

5. CONCLUSIONS

Multiple user feature was constructed, and whether these features are suit to user is quantitative analysed. Furthermore, a multi-channel CNN is designed to depict user's applicability to the feature models. The experiments result shows the effectiveness of the method.

Although the method is effectiveness in the experiment, it should be further evaluated in a more realistically data and scenarios. In addition, the feature model constructed now is data-sensitive, how to generated a more stable feature of the users in an abstract level is an interesting research direction.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. U1504602.

REFERENCES

[1] Blomberg, J., Burrell, M. and Guest, G., "An ethnographic approach to design," in: Jacko, J. A. and Sears A. (eds.), [The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications], Lawrence Erlbaum Associates, Mahwah, (2002).

[2] Anastassova, M., Mégard, C. and Burkhardt, J. M., "Prototype evaluation and user-needs analysis in the early design of emerging technologies," Inter. Conf. on Human-computer Interaction: Interaction Design & Usability, (2007).

[3] Xu, T., Ma, Y. and Wang, Q., "Cross-urban point-of-interest recommendation for non-natives," International Journal of Web Services Research (IJWSR), 15(3), 82-102 (2018).

[4] Cai, G., Lee, K. and Lee, I., "Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos," Expert Systems with Applications, 94, 32-40 (2018).

[5] Böhmer, M., Bauer, G. and Krüger, A., "Exploring the design space of context-aware recommender systems that suggest mobile applications," 2nd Work. on Context-Aware Recommender Systems, (2010).

[6] Hess, A., Hummel, K. A., Gansterer, W. N., et al., "Data-driven human mobility modeling: A survey and engineering guidance for mobile networking," ACM Computing Surveys (CSUR), 48(3), 38 (2016).

[7] Sun, X., Huang, Z., Peng, X., et al., "Building a model-based personalised recommendation approach for tourist attractions from geotagged social media data," International Journal of Digital Earth, 12(6), 661-678 (2019).

[8] Wai, K. P. and New, N., "Measuring the distance of moving objects from big trajectory data," 2017 IEEE/ACIS 16th Inter. Conf. on Computer and Information Science (ICIS), 137-142 (2017).

[9] Hsueh, Y. L. and Huang, H. M., "Personalized itinerary recommendation with time constraints using GPS datasets," Knowledge and Information Systems, 1-22 (2018).

[10] Ding, Y., Liu, J., Jiang, C., et al., "A study of friends recommendation algorithm considering users' preference of making friends in the LBSN," Systems Engineering—Theory & Practice, (11), 22 (2017).

[11] Zhu, Z., Cao, J. and Weng, C., "Location-time-sociality aware personalized tourist attraction recommendation in LBSN," 2018 IEEE 22nd Inter. Conf. on Computer Supported Cooperative Work in Design (CSCWD), 636-641 (2018).

[12] Yang, D., Zhang, D., Zheng, V. W., et al., "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(1), 129-142 (2015).

[13] Yang, D., Zhang, D. and Qu, B., "Participatory cultural mapping based on collective behavior data in location-based social networks," ACM Transactions on Intelligent Systems and Technology (TIST), 7(3), 1-23 (2016).

[14] Xu, S., Fu, X., Cao, J., et al., "Survey on user location prediction based on geo-social networking data," World Wide Web, 23, 1621-1664 (2020).

[15] Yang, D., Zhang, D., Zheng, V. W., Yu, Z. and Wang, Z., "A sentiment-enhanced personalized location recommendation system," Proc. HT, 119-128 (2013).

[16] Lathauwer, L. D., Moor, B. D. and Vandewalle, J., "Multilinear singular value tensor decompositions," SIAM Journal on Matrix Analysis and Applications, 24(4), 1253-1278 (2000).

[17] Rahimi, S. M. and Wang, X., "Location recommendation based on periodicity of human activities and location categories," Pacific-Asia Conf. on Knowledge Discovery and Data Mining, in: Pei, J., Tseng, V. S., Cao, L., Motoda, H. and Xu, G. (eds.), Lecture Notes in Computer Science, Springer, Berlin, 7819 (2013).