URDO&RG: A Tool for User Role Discovery and Optimization & Requirement Generation Based on User Check-in Data in LBSN

Yuanbang Li[1], Kai Wang[1], Ping Liu[2], Zhenzhu Fu[1], Chi Xu[1]*

[1] *School of Computer Science and Technology, Zhou Kou Normal University, China*
[2] *Information Section of the Medical Security Center, 988 Hospital of the United Logistics, China*
lybang@whu.edu.cn, 3101709158@qq.com, 20212017@zknu.edu.cn,
1431600697@qq.com, 20201067@zknu.edu.cn

Abstract

The regularities shared by a set of users in a long-term in Location-Based Social Network (LBSN) can be abstracted as user roles. Although methods are proposed to identify user roles, the quantitative analysis of the effectiveness of the method is insufficient, and as far as we know, tools to generate user roles automatically from user data are missing. Therefore, a tool named URDO&RG (User Role Discovery and Optimization & Requirements Generation) is realized. The tool includes four main functions: (1) visualization of the data from multiple dimensions in diverse display diagram, (2) discovery of user roles using various algorithms, (3) optimization of the method to achieve better user role discovery outcome. (4) generation of user requirements from user data. Data set is used to validate the effectiveness of the tool. Result shows that, the tool can discover user roles and generate user requirements effectively, and the optimization method outperforms the original methods in many evaluation metrics.

Keywords: Location based social network, User role discovery, Optimization, User requirement generation

1 Introduction

As of December 2022, the number of Chinese netizens reached 1.067 billion, of which more than 99 percent are mobile netizen. With the explosive use of mobile phones, users tend to share their experiences freely when they attend various social activities, which prompted the foundation of Location-based Social Network (LBSN) [1].

LBSN is a complex network with rich physical and social information. It includes various colorful functions and is capable of providing intelligent service to end users, which attracts many users to engage in various activities.

Massive data are accumulated by the activities that users conducted in LBSN, which include a rich user knowledge. The knowledge includes time related knowledge, location related knowledge, and social related knowledge. Time related knowledge describes the activity patterns of users in time, such as when are they active, in the morning, afternoon or late night. Location related knowledge describes activity patterns related to location, such as user often engage in activities in busy markets, suburbs, or close to home, social related knowledge such as user often engage in food related activity or job-related activity or entertainment related activity.

Although users' activities may be varied, these activities may take on regularities in the long run. These regularities, which are shared by a set of users can be identified as user roles. These user roles are important means to reveal user characteristics and generate user requirements [2].

The concept of role used in sociology is an abstraction of 'patterned human behavior' [3]. In traditional sociological studies, user role is either determined by profession, such as lawyer or actor, or assigned legally, such as commissioner or minister, or determined based on personal relationships such as brother or sister and so on. Those users share a set of behaviors as they are expected such as the minister supervising the staff of the department to implement various work instructions of the superior, check the work conditions of each post, and correct the work deviation in time.

The identification of user role is meaningful in online social networks [4-5]. However, it is a difficult task to identify user role in online scenario. The reasons are as follows: firstly, it is difficult to communicate with the user directly, the information that can be obtained directly in the online social network is the virtual accounts of the user without social attributes. Secondly, user's behavior in social networks is fragmentary and messy. Thirdly, Users are reluctant to provide their personal privacy information.

The use of social networks by users generates massive amounts of user data, which contains a wealth of user information. The identification and analysis of user role is helpful to understand the social attributes and living habits of users, so as to deduce the requirements of users and provide them with more intelligent services.

Although methods have been proposed to identify user roles in social network, these methods focus on using a single technology to discover user roles from regular data, and there is no tool to support data presentation, user role discovery and requirements generation.

Therefore, in this article, a tool named URDO&RG is developed to support the discovery and optimization of user role from user data and the generation of user requirements.

Contributions of this article including:

*Corresponding Author: Yuanbang Li; E-mail: lybang@whu.edu.cn
DOI: https://doi.org/10.70003/160792642024112506010

(1) Multiple display diagrams are implemented to exhibit data distribution from various perspectives, including heat map, correlation matrix, scatter matrix, histograms, density plots and boxplots.
(2) Five clustering algorithms are used to generate user roles, and their effectiveness are discussed.
(3) Two kinds of algorithms, named GBK-means and RDK-means are implemented to optimize user role generation.
(4) Fuzzy logic is used to generate user requirements based on user data, which can help managers to understand users' requirements more intuitively.
(5) A tool is implemented to integrate these functions mentioned above to provide managers with one-stop user role analysis and requirement generation services.

This paper is structured as follows:

Related works are summarized and analyzed in the next section. The framework and the function of the tool is detailed in section 3. Experiments is conducted in section 4 to show the effectiveness of the optimization of user role discovery. Then, the meaningful findings and limitations of the tool is discussed in section 5. Finally, the paper is concluded in section 6.

2 Related Works

2.1 Traditional User Role Identification Methods

Traditional user role identification methods include expert experience-based methods and survey-based methods. The former method identifies user roles based on domain experts, such as defined by senior practitioner, the roles in computational advertising were summarized [6]. And the latter method identifies user roles through a literature survey or users. Such as user roles in enterprise social networks are defined as initiator, coordinator, expert and so on based on literature review [7]. While the roles in digital library are identified investigate of a group of users [8].

Traditional methods tend to be inefficient, and the results are often difficult to generalize. Therefore, more and more automatic methods based on the analysis of user activity data are proposed, including preassigned role identification and self-assigned role identification based on whether the value set of user roles is determined [9].

2.2 Preassigned User Role Identification Methods

The preassigned role identification method focuses on when the application background and the value set of the role is known, how to determine which type of roles the users belong to. Preassigned role identification methods mainly including two kinds of methods, namely content analysis and social network analysis methods.

Content analysis methods identify user roles based on the analysis of user post content, user reply content, user comment content and user expertise level [10-12]. P. Roy identified the role of experts based on the historical question and answer behavior of users, combined with the current answer behavior and Posting time [10]. While T. Zhao infers the topic knowledge expertise of each user based on the measurement of the topic sensitivity of users and professional knowledge, and then constructs an appropriate recognition framework to identify user roles in each topic [11].

Social networks have great influence on human activities. A post published by a star can result in his fans to respond or even act crazy in his life. In social network analysis methods, users, posts and other parts in the network are constructed as nodes of the network, relationship between users, forwarding or comments activities between users and posts are regarded as the edges of the network. Then, based on the analysis of the network characteristics, such as outgoing, incoming, point centrality and so on, the roles in the network are identified [13-14], such as Y. Hu [13] identifies the role of opinion leaders according to network characteristics such as degree centrality, intermediate centrality and proximity centrality. Kayes [14] used the network centrality matrix for analysis and found that blog influence usually presents a "core-edge" network structure connection distribution. Therefore, the role of "Influencer" can be identified through the discovery of the core node in the network.

Some researchers try to combine content analysis method and social network analysis method to achieve better results in user role discovery [15-16]. For example, K. Song constructed a multi-topic user social network model based on explicit and implicit links, combined with the emotional analysis of the comments, the role of "Opinion leader" are identified effectively [15]. Tang [16] propose a framework to identify user roles in online healthcare forums, which constructs the social network of healthcare forum firstly, and use UserRank algorithm to combine the social network with link and content analysis to establish a comprehensive identification mechanism to identify user roles.

2.3 Self-assigned User Role Identification Methods

Self-assigned role identification method focuses on when the application background and the value set of the role is unknown, how to identify patterns of user activities, and abstract them as user roles based on the analysis of user data. Self-assigned role identification methods mainly including three kinds of methods, namely mathematical based method, social network analysis method and machine learning based method.

Mathematical based method has been widely used in user role identification [17-18]. McCallum [17] proposes three kinds of Bayesian hierarchical model namely ART, RART1 and RART2 to identify user roles in the mail data set. Peleshchyshyn [18] propose a set of user activity indicators such as "content author placement" and "influence action", which are basis of user role identification, and then develop special marks in mathematical expressions to represent these indicators.

Social network analysis method can also have been used in self-assigned user role identification [19-21]. S. L. Toral identifies the role of "regular contributors" and "brokers" based on average emergence and centrality in the social network [19]. J. Fuller identifies the role of "creative users", "experts", "efficient contributors" and "passive reviewers" according to external centrality, degree centrality and creative contribution in the social network [20]. While Q. Ma identifies roles in social network based on the analysis of network typologies such as degree centrality and so on [21].

Machine learning based methods, especially the clustering methods have been used to identify self-assigned user roles, because clustering methods are suitable for abstracting the characteristics of users from data without labels. Hacker [22] using K-means to identify nine user roles in corporate social network such as "power users", "focused information sharers", "team members" and so on.

Some researchers also try to combine more than one kind of method to achieve better results in self-assigned user role discovery [23-24]. Guo [23] combines machine learning method with social network analysis to identify user roles in professional virtual communities. W. Liu [24] combines Dirichlet mixture model of Hawkes Processes with clustering method for user roles discovery.

Although method have been proposed to identify user roles, current research concerns with the identification of roles, while the effectiveness of the clustering algorithm is in lack of analysis, and as far as we know, tools specifically designed to identify user roles from social network data are missing. Therefore, a tool is implemented to discover and optimize user role from social network data in this paper.

3 URDO&RG

3.1 Framework

As the tool aims to provide one-stop user role analysis and requirement generation services, the main function of the tool consists of four parties, which is shown in Figure 1.

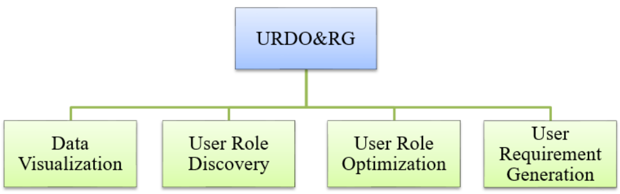


Figure 1. Framework of the tool

The first part is data visualization, aims to display data features with colorful visual diagrams. The second part is user role discovery, aims to discover user roles from user data using clustering algorithms. The third part is user role optimization, which aims to optimize the user role discovery algorithms. The fourth part is user requirements generation, aims to describe user requirements in nature language using fuzzy logic.

The overall layout of the tool is shown in Figure 2. The function bar and function buttons are located in the upper left of the panel. The display part of the panel is divided into 3 areas, which are visualization area on the left, file management area in the upper right and process information area in the lower right. The visualization area is used to display the data analysis results and role generation results visually. The file management area is used to display the file structure in the project. The process information area is used to display the data and execution progress during algorithm running.

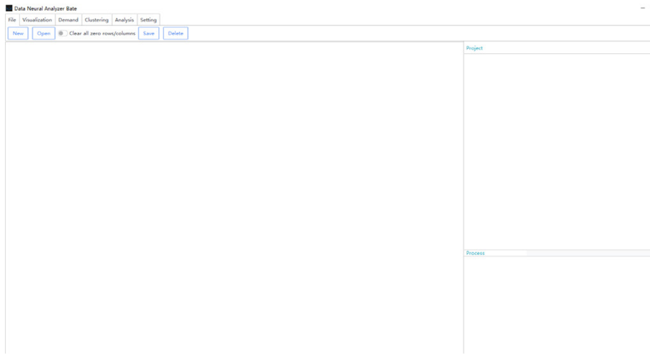


Figure 2. Layout of the tool

3.2 Data Visualization

The first step is data loading, the function buttons are shown in Figure 3. In which "New" is used to open a data table, "Open" is used to open an existing project, "Save" and "Delete" are used to save or delete the output result shown in the panel.

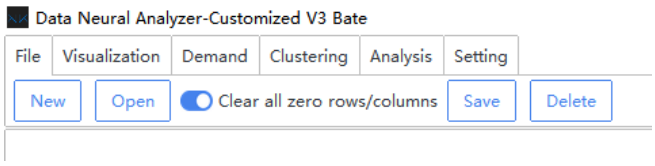


Figure 3. Buttons for data loading

In order to satisfy the multi-angle visual analysis of data by managers, this tool provides a variety of perspective functions, including heat map, correlation matrix, scatter matrix, histograms, density plots and boxplots. All of these functions exhibit the data from different perspectives and can help the managers to analyze data easily. It is worth noting that before using these functions, the dimension to display the data must be selected as a scenario for analysis. Figure 4 is an illustration of the buttons for data visualization.

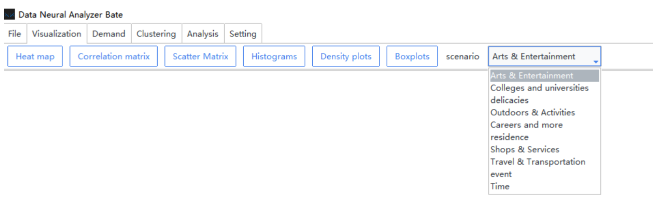


Figure 4. Buttons for data visualization

3.3 User Role Discovery

Five clustering algorithms are used to identify user roles, including the K-means algorithm, DBSCAN algorithm, Bi-Kmeans algorithm, OPTICS algorithm and Agglomerate algorithm. After the clustering process is completed, the output of the algorithms including the clustering results and various indicators. Multiple pictures of the clustering results can be switched to view. All of these algorithms were implemented in Python on a personal computer.

3.4 User Role Optimization

In order to improve the effectiveness of user role discovery using clustering algorithm, two optimization algorithms named GBK-means and RDK-means is designed to optimize user role discovery.

3.4.1 GBK-means

The definitions used in GBK-means are shown as follows.

Definition 1: Data separation degree (SEP)

SEP denotes the separation degree of the data, which is calculated as formula 1. A small degree of separation means relatively centralized data, while large degree of separation means relatively dispersed data. In which MAX_{si} and MIN_{si} denote the maximum and minimum value of the i-th dimension data in data set S.

$$SEP_s = \frac{\sum_{i=1}^{n} \frac{STP_{si}}{MAX_{si} - MIN_{si}}}{n} \tag{1}$$

Definition 2: Number of grid divisions (M)

M is used to divide each dimension of data set S into several parts, which is calculated as formula 2. In which k is cluster number.

$$M = \left[\frac{\sqrt{k}}{SEP_s} + 1 \right] \tag{2}$$

Definition 3: Dense grid (gd)

Dense grid denotes whether data number contained in the grid is greater than or equal to the density threshold β or not. Where β is calculated as formula 3.

$$\beta = \frac{R}{M^n (1 - B)} \tag{3}$$

R is the elements number in the data set, B is the blank grid ratio, which is calculated in formula 4. Grids without data are called blank grids, labeled as G_B.

$$B = \frac{\lVert G_B \rVert}{M^n} \tag{4}$$

Based on the above definition, the process of GBK-means is shown in Algorithm 1.

Algorithm 1. GBK-means algorithm

Input: cluster number k, data set S with R elements and n dimensions

Output: k initial clustering centers

1: divide each dimension in S into M equal parts
2: divide the data set into M^n grids
3: identify the dense grid for each dimension, denoted as C_i
4: identify the data in C_i, denoted as D
5: initial the state for each of the dense grid C_i to unvisited
6: for each c in the dense grid set C:
7:     if the state of c is unvisited:
8:         denote the state of c to visited, assign a new cluster tag CT, create list L, add c to L
9:     else:
10:         continue
11: endfor
12: gets L's header, check adjacent unvisited grids, change it to visited, if it is dense grid, denote its tag, and add to list L
13: connect to dense grid with the same tag, forming dense grid area
14: using k-means to obtain the primary cluster centers
15: calculate the average value of primary clusters as initial centers
16: if initial center numbers j>k, DBSCAN algorithm is used to get k clustering centers
17: if j<=k, K-means++ is used
18: print the resulting cluster centers

The result of running GBK-means is shown in Figure 5.

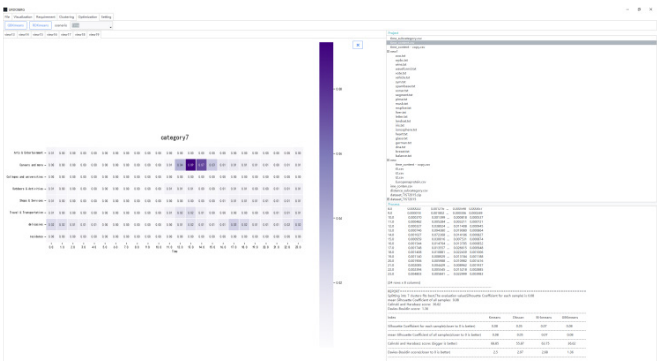


Figure 5. Result of running GBK-means

3.4.2 RDK-means

The process of RDK-means is shown in Algorithm 2.

Algorithm 2. RDK-means algorithm

Input: data set S with R elements and n dimensions, k clustering centroid

Output: k initial clustering centers

1: calculate the variance of all sampling points
2: add the minimum variance sample point to the centroid set
3: take the average distance of the minimum variance sampling points as the radius, and delete the sampling points in the circle
4: determine whether the initial centroid number is less than k. If yes, go to step 3; else to next step
5: save k initial centroid
6: calculate the distance between the centroid and save it in the distance list
7: take the minimum distance from the distance list, denoted as d_min
8: each centroid is plotted as a circle of radius γ
9: check whether the intersection of circles exists. if yes, go to step 10, else go to step 11
10: the points in the intersection area are divided into clusters at the center of the nearest intersection circle
11: the points are divided into clusters with the centers of the circles
12: delete the points

13: delete d_min from the distance list
14: check whether the distance list is empty; if yes, go to the next step otherwise, return to step 7
15: update the center point and take the mean of all sample points as new center point
16: check whether the center point still changes. If yes, return to step 6
17: print the final centroid

The result of running RDK-means is shown in Figure 6.

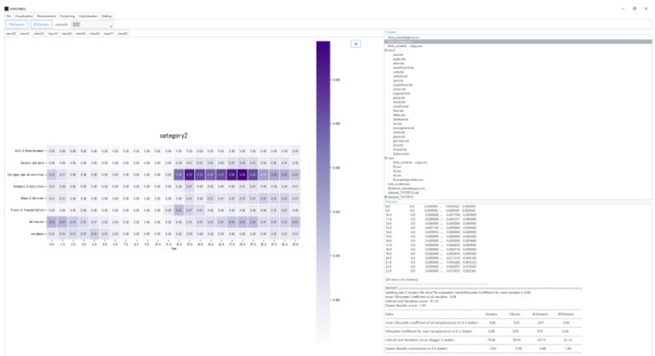


Figure 6. Result of running RDK-means

3.5 User Requirement Generation

In this part, the fuzzy logic is used to generate user requirements according to user data. Note that it is necessary to select a scenario to start the process, after all the requirements are generated, the program will automatically pop up the corresponding data histogram, so that managers can understand user requirements more intuitively. The information used in user requirements including the number of users, the context elements of user's activity, analyst's observe dimension on data, the fuzzy descriptions generated based on the calculation of fuzzy logic. The calculation rules are summarized in formula 5, 6 and 7.

$$M\_d(\text{low}) = \frac{\frac{avg}{2} - u\_c}{\frac{avg}{2} - min},\ u\_c < \frac{avg}{2} \tag{5}$$

$$M\_d(med) = \frac{u\_c - \frac{avg}{2}}{avg + \frac{\max - avg}{2} - \frac{avg}{2}}, \quad \frac{\text{avg}}{2} \le u\_c < avg + \frac{\max - avg}{2} \tag{6}$$

$$M\_d(high) = \frac{u\_c - \left(avg + \frac{\max - avg}{2}\right)}{\max - \left(avg + \frac{\max - avg}{2}\right)}, \quad u\_c \ge avg + \frac{\max - avg}{2} \tag{7}$$

In these formulas, low, middle and high are used to describe the degree to which a requirement is established, avg, max and min denote the average times, maximum times and minimum times of user activity.

4 Experiments

Based on publicly available data sets, experimental methods are used to verify the usefulness of the tool and the effectiveness of algorithm.

4.1 Data Set and Preprocessing

The data set used in this paper contains 1083 users, 38333 points of interest and 227428 check-ins [25]. Each data item of a check-in activity including user number, POI (point of interest)'s identifier, category number and category name, the position of the activity, and the time of the activity.

As the category name of POI includes hundreds of categories, and too many categories drown out the key characteristics of user in user features analysis. Therefore, The POI root category is introduced, which is generated based on the hierarchy tree in Foursquare website.

The time factor has an important effect on user characteristics. In this study, time related feature is calculated in hours. Time measured in hours is discrete, which may result in inaccurate constructing user features. Such as when the time is 08:59:00, hour of the time is 8, however, this check-in activity has greater impact on user's feature at 9 o'clock. Therefore, a temporal smoothing strategy is proposed in the construction of user's features. The strategy smooths an activity to adjacent hours, which is shown in formula 8.

$$\begin{cases} Cou(h) = Cou(h) + \left(1 - \frac{\text{m}}{60}\right) \\ Cou((h+1)\%24) = Cou((h+1)\%24 + \left(1 - \frac{60*1-m}{60}\right) \end{cases} \tag{8}$$

In which h represents hours, and m represents minutes, as second have limited effect on hours, it is not considered in this study. After using the strategy, if an activity happened in 16:12:00, the impact of this activity is: Cou(16)=Cou(16)+ 0.8, Cou(17)=Cou(17)+ 0.2.

The effectiveness of the smoothing strategy has been evaluated based on the data set. As feature matrix is established for each user separately, the data for each user is identified firstly, then, 80 percent of the data are used to construct user feature matrix, remaining are used to verify the effectiveness of the prediction using user feature matrix.

As analysts are concerned with whether the service provided to the user is suitable, accuracy of the prediction is more important. The top K accuracy rate is used, which is shown in formula 9.

$$Accuracy@K = \frac{\left|\{u,l,t,a\} \mid a \in P_{u,l,t}(K), (u,l,t,a \in TS)\right|}{|TS|} \tag{9}$$

In the formula, u denotes user, l denotes location, t denotes time and TS denotes the test set. In the experiment, user features constructed with and without the smoothing strategy are both constructed, they are used separately to predict user activities, and the accuracy rates are calculated using formula 9. The result of Top 1 accuracy and Top 2 accuracy are shown in Figure 7.

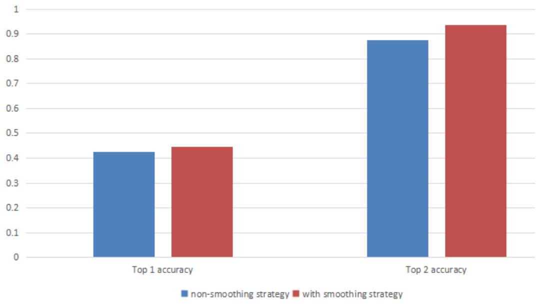


Figure 7. Result of with or not with smoothing strategy

The user feature matrix is constructed based on user data, which is the basis of user analysis, such as user time-root category feature matrix is shown in Figure 8.

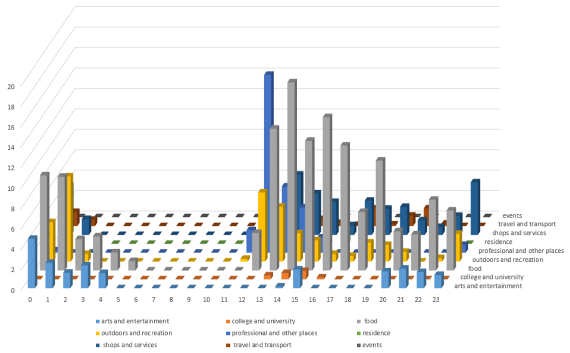


Figure 8. User time-root category feature matrix

As shown in the figure, the horizontal axis denotes time, the longitudinal axis denotes root categories, and the vertical axis denotes the total times of user's activity in corresponding time and root category.

4.2 Data Visualization

The graphical display helps to understand the data. The demonstration of the data in variety perspective functions, namely heat map, correlation matrix, scatter matrix, histograms, density plots and boxplots are shown in Figure 10.

4.3 User Role Discovery

Clustering algorithm can divide users into user sets with similar characteristics, that is, user roles. In keeping with the previous chapter, the result of DBSCAN clustering algorithm is shown in Figure 9. As shown in the figure, 7 categories from view0 to view6 in visualization area is the visual illustration of the clustering results, such as view0 shown in the figure, users tend to be active in about 6:00 to 9:00 in outdoors & activities. This means that users of these avatars like to participate in outdoor activities in the morning to maintain a healthy body. The data shown in the process information area describe the number of rows and columns of the user feature matrix, the evaluation metrics of the algorithm, the number of categories of the clustering, the number and percentage of users for each category.

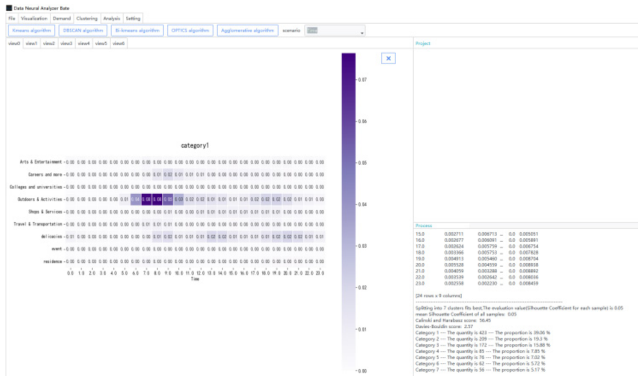


Figure 9. Clustering results for DBSCAN algorithm

4.4 User Role Optimization

The evaluation metrics and effect of the algorithm and the running result are detailed in this section.

- Evaluation metrics

Silhouette Coefficient and Davies-Bouldin score are used as evaluation metrics in this study.

Silhouette Coefficient is an import evaluation metric to evaluate clustering effect [26]. When the distance between the points in the same cluster is relatively close, and the distance between one cluster and its nearest cluster is farther, the closer the contour coefficient value is to 1, that is, the better the clustering effect is. Similarly, all sample points in a cluster are relatively far apart, and the cluster where the sample is located is closer to its nearest cluster, the worse the clustering effect is.

Silhouette Coefficient is suitable to evaluate the effectiveness of user role discovery algorithm. because the larger the value is, the higher the similarity of users in the same role group is, and the greater the difference between the characteristics of this role group and those of other groups is. Which means the better the user role discovery algorithm is.

Davies-Bouldin score is used to evaluate the similarity degree between each cluster and its most similar cluster, and then the average value of all these similarity degrees is used to evaluate the quality of the whole clustering result. If the similarity between clusters is higher, the value of Davies-Bouldin score is larger, which means the distance between one user group to others is smaller, then the user role discovery algorithm is worse.

The experiment result of silhouette coefficient and Davies-Bouldin score is shown in Figure 11 and Figure 12.

As shown in these figures, the Silhouette Coefficient score of RD-Kmeans and GB-Kmeans is Higher than DBSCAN, Bi-Kmeans and OPTICS algorithm, and not less than K-means and Agglomerate algorithm. The Davies-Bouldin score of RD-Kmeans is lower than the five clustering algorithms, namely DBSCAN, Bi-Kmeans, OPTICS, K-means and Agglomerate algorithm, The Davies-Bouldin score of GB-Kmeans is well below all of the five clustering algorithms and is below than RD-Kmeans algorithm. The results show that the two optimization algorithms RD-Kmeans and GB-Kmeans are effective in user role discovery optimization.

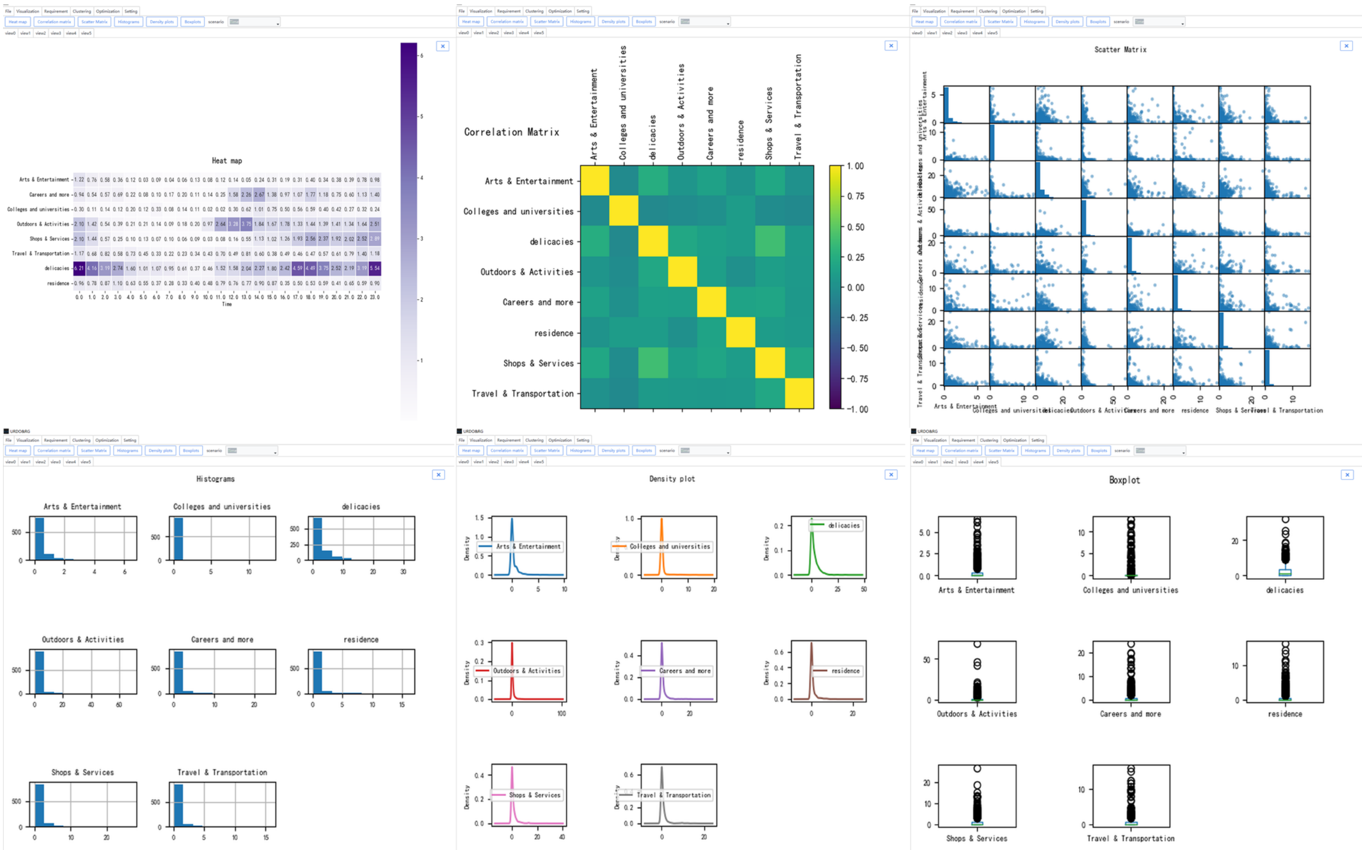


Figure 10. Comprehensive demonstration of data visualization

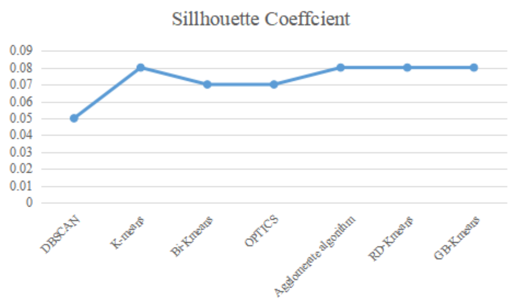


Figure 11. The experiment result of silhouette coefficient

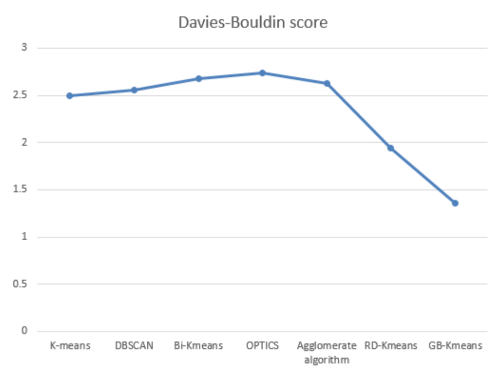


Figure 12. The experiment result of Davies-Bouldin score

4.5 User Requirement Generation

The user's check-in is an exact number, and fuzzy logic can map the exact number to a fuzzy description in natural language. In this paper, fuzzy logic is used to mapping accurate data into three levels, namely high, middle and low, and combined with the categories of user sign-in locations, user requirements are generated in natural language.

The maximum values, minimum values and average values are needed in using the fuzzy logic method. Based on the user feature matrix, these values are calculated, such as the maximum values and average values are shown in Figure 13 and Figure 14.

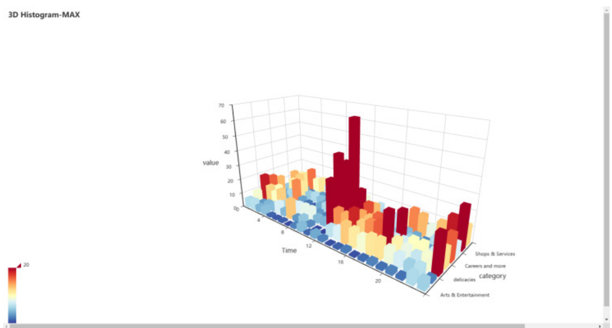


Figure 13. The maximum value of the data

When these values are obtained, the membership degree of low, medium and high are calculated according to formulas 5, 6, and 7. And then, a natural language-based template is used to generate user requirements. Note that there are many items of requirements based on the data, only the representative requirements, namely the probability is larger than a threshold value will be reserved. The illustration of these requirements is shown in Figure 15.

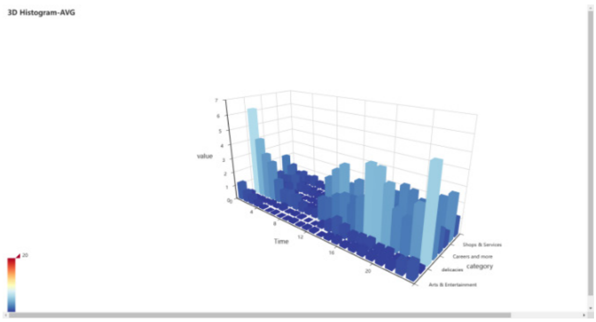


Figure 14. The average of the data

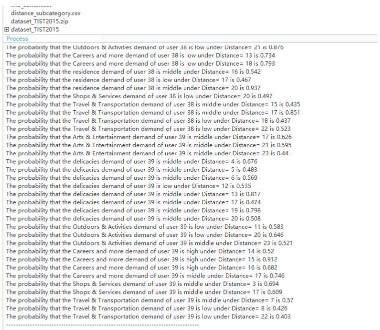


Figure 15. The requirements generation results

5 Conclusion and Future Work

Millions of mobile netizens generate a volume of check-in data which contains rich user information. User role is an abstraction of a group of users with similar characteristics. And it is an assist to understand the social attributes of users and provide them with intelligent services. In this study, a tool is designed and implemented to realize the process from user check-in data preprocessing and display, user role discovery, user role optimization and user requirement generation, which can provide convenience and help for users' analysis.

Future work consists of the following parties, the first one is to verify the effectiveness of the optimization algorithm on more data. The second part is to verify the efficiency of the tool on a large amount of data. And the third one is combining other types of data such as social network data or comments data into the tool to extend the functionality of the tool.

Acknowledgement

This research was funded by funded by University Key Scientific Research Project of Henan Province, grant number No. 22A520052. And High-level Talent Scientific Research Start-up Fund under Grant No. ZKNUC2021023.

References

[1] Y. Zheng, Location-based social networks: users, in: Y. Zheng, X. Zhou (Eds.), *Computing with Spatial Trajectories*, Springer, New York, 2011, pp. 243-276.

[2] J.-V. Hacker, F. Bodendorf, P. Lorenz, A framework to identify knowledge actor roles in enterprise social networks, *Journal of Knowledge Management*, Vol. 21, No. 4, pp. 817-838, July, 2017.

[3] B.-J. Biddle, *Role theory: Expectations, identities, and behaviors*, Academic Press, 2013.

[4] A. Arora, A. Taneja, Research issues, innovation and associated approaches for recommendation on social networks, *International Journal of Performability Engineering*, Vol. 17, No. 12, pp. 1027-1036, December, 2021.

[5] H. Yang, Y. Zhang, J. Zhang, D. Chen, G. Sun, Event-based Community Detection in Micro-Blog Networks, *International Journal of Performability Engineering*, Vol. 17, No.1, pp. 60-73, January, 2021.

[6] Y. Liu-Thompkins, E. Maslowska, Y. Ren, H. Kim, Creating, metavoicing, and propagating: a road map for understanding user roles in computational advertising, *Journal of Advertising*, Vol. 49, No. 4, pp. 1-17, August, 2020.

[7] M. Morzy, On Mining and Social Role Discovery in Internet Forums, *International Workshop on Social Informatics*, Warsaw, Poland, 2009, pp. 74-79.

[8] Oxford University, *Digital Engagement in GLAM*, Gardens, Libraries & Museums, 1060891, February, 2021. https://www.glam.ox.ac.uk/digital-engagement-glam

[9] W.-W. Zhang, J. Zhu, Review of User Roles in Professional Virtual Communities, *Journal of Modern Information*, Vol. 40, No. 7, pp. 167-177, August, 2020.

[10] P.-K. Roy, J.-P. Singh, A. Nag, Finding Active Expert Users for Question Routing in Community Question Answering Sites, *Proceedings of the Machine Learning and Data Mining in Pattern Recognition*, New York, NY, USA, 2018, pp. 440-451.

[11] T. Zhao, H. Huang, X. Fu, Identifying Topical Opinion Leaders in Social Community Question Answering, *Proceedings of the Database Systems for Advanced Applications*, Gold Coast, QLD, Australia, 2018, pp. 372-387.

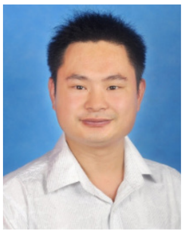
[12] S. Chatterjee, R. Chaudhuri, D. Vrontis, R. Piccolo, Enterprise social network for knowledge sharing in MNCs: Examining the role of knowledge contributors and knowledge seekers for cross-country collaboration, *Journal of International Management*, Vol. 27, No. 1, pp. 1-14, March, 2021.

[13] Y. Hu, Identification of Opinion Leaders in the Network Academic Community, *Journal of Library and Information Science*, Vol. 2, No. 11, pp. 72-77, 2017.

[14] I. Kayes, X. Qian, J. Skvoretz, Adriana Iamnitchi, How

Influential Are You: Detecting Influential Bloggers in a Blogging Community, *4th International Conference on Social Informatics (SocInfo 2012)*, 2012, Lausanne, Switzerland, pp. 29-42.
[15] K. Song, D. Wang, S. Feng, D. Wang, G. Yu, Detecting Positive Opinion Leader Group from Forum, *Proceedings of the Web-Age Information Management*, Harbin, China, 2012, pp. 95-101.
[16] X. Tang, C.-C. Yang, Identifying influential users in an online healthcare social network, *Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics*, British Columbia, Canada, 2010, pp. 43-48.
[17] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on enron and academic email, *Journal of Artificial Intelligence Research*, Vol. 30, pp. 249-272, October, 2007.
[18] A. Peleshchyshyn, V. Vus, O. Markovets, Solomiia Albota, Identifying Specific Roles of Users of Social Networks and Their Influence Methods, *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies*, Lviv, Ukraine, 2018, pp. 39-42.
[19] S.-L. Toral, M.-R. Martínez-Torres, F. Barrero, Analysis of Virtual Communities Supporting OSS Projects Using Social Network Analysis, *Information and Software Technology*, Vol. 52, No. 3, pp. 296-303, March, 2010.
[20] J. Fuller, K. Hutter, J. Hautz, Kurt Matzler, User Roles and Contributions in Innovation-Contest Communities, *Journal of Management Information Systems*, Vol. 31, No. 1, pp. 273-307, Summer, 2014.
[21] Q. Ma, X. Luo, M. Zhao, User Behavior Network Based User Role Mining of Web Event, *2018 14th International Conference on Semantics, Knowledge and Grids*, Guangzhou, China, 2018, pp. 211-217.
[22] J. Hacker, K. Riemer, Identification of user roles in enterprise social networks: method development and application, *Business & Information Systems Engineering*, Vol. 63, No. 4, pp. 367-387, August, 2021.
[23] W. Guo, Q. Zheng, W. An, W. Peng, User roles and contributions during the new product development process in collaborative innovation communities, *Applied Ergonomics*, Vol. 63, pp. 106-114, September, 2017.
[24] W. Liu, Q. Zhang, W. Qiao, Dirichlet Mixture Model of Hawkes Processes Based Patent User Role Discovery Model, *International Joint Conference on Neural Networks*, Padova, Italy, 2022, pp. 1-8.
[25] D. Yang, D. Zhang, V.-W. Zheng, Z. Yu, Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 45, No. 1, pp. 129-142, January, 2015.
[26] P.-J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, Vol. 20, pp. 53-65, November, 1987.

Biographies



Yuanbang Li received the M.E. degree from Zhengzhou University of China in 2010 and the PhD from Wuhan University in 2021. Currently, he is worked in the School of Computer Science and Technology at Zhoukou Normal University. His research interests include requirement engineering and social network.

Kai Wang is now a bachelor candidate in Zhoukou Normal University, He participated in the 3rd, 4th and 5th “Kingduns Information Security Cup”, the 7th “Strong Network Cup” National Cyber Security Challenge and won multiple awards. His research interests include machine learning and cybersecurity.

Ping Liu received the master. degree from PLA Information Engineering University in 2008. Currently, She is worked in The 988th Hospital of the Joint Logistics Support Force of the People’s Liberation Army. Her research interests include information retrieval code as research direction.

Zhenzhu Fu received the Bachelor degree in Wuhan University of Technology, and she is pursuing a Master’s degree at Zhongnan University of Economics and Law. She is now working in Zhoukou Normal University as a lecture. Her research interests include data mining and data analysis.

Chi Xu received the PhD from Sun Yat-Sen University in 2020. Currently, he is worked in the School of Computer Science and Technology at Zhoukou Normal University. His research interests include Geological big data and computer numerical simulation.