Advances in Computer and Communication, 2024, 5(2), 152-157
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Crowdsourcing User Community Division and Experimental Analysis Based on Attribute Heterogeneity

Tingting Zhang[1,*], Ke Chen[1], Yuanxing Zhao[2], Huangtao Zhao[1]

[1]School of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China.
[2]Jinken College of Technology, Nanjing, Jiangsu, China.

How to cite this paper: Tingting Zhang, Ke Chen, Yuanxing Zhao, Huangtao Zhao. (2024) Crowdsourcing User Community Division and Experimental Analysis Based on Attribute Heterogeneity. *Advances in Computer and Communication*, 5(2), 152-157.
DOI: 10.26855/acc.2024.04.010

Received: March 28, 2024
Accepted: April 26, 2024
Published: May 24, 2024

***Corresponding author:** Tingting Zhang, School of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China.

Abstract

Under the crowdsourcing environment, dividing the user community is not only beneficial for understanding the platform's user structure and managing resources within the platform but also enhances user group recommendations. In this paper, a comprehensive similarity calculation method based on heterogeneous attributes is proposed, building on existing research and considering the heterogeneity of user attributes and inter-user structure. Based on the similarity between nodes, a specific threshold is selected for community division using a systematic clustering approach. The community structure evaluation function is designed to assess the quality of community division, and the algorithm's accuracy is demonstrated through experimental analysis. The research results show that the method in this paper can not only effectively mine the actual structure within the user community and achieve the rational subgrouping of crowdsourcing users but also help enhance the accuracy of the platform's task recommendations, thereby improving the platform's operational efficiency and visibility.

Keywords

Crowdsourcing, attribute characterization, heterogeneity, community division

1. Introduction

In a crowdsourcing environment, dividing user communities is advantageous for understanding the platform's user structure, managing internal resources, and recommending user groups [1]. For instance, segmenting users on the website based on similar interests and preferences enables the recommendation of relevant user groups for large-scale complex tasks, thereby enhancing the quality of crowdsourced task completion. In current research on community segmentation, most are based on users' structural characteristics and attribute features [2]. These features can include information such as users' interests, behavior patterns, geographical location, age, gender, educational background, and more. Utilizing these features for community segmentation helps in understanding the similarities and differences between user groups, thus aiding in better management and provision of relevant services [3].

Users' attribute characteristics are usually heterogeneous due to the fact that they may have multiple belonging categories [4, 5]. In a crowdsourcing environment, during the process of task searching and selection, users may consider their own professional knowledge background to choose tasks that match their capabilities. Therefore, they typically possess domain-specific characteristics related to their affiliations. Furthermore, due to the dynamic and diverse nature of user interests, individuals may possess other interest preferences in addition to their professional expertise. For instance, an IT professional might have a certain level of interest in painting. Consequently, users typically exhibit corresponding preference characteristics apart from their domain background. Indeed, a user may

have multiple affiliations, leading to heterogeneous distribution characteristics in their attributes.

This paper conducts community segmentation based on the heterogeneous attribute characteristics possessed by users, differing from traditional studies that assume independence between user structure and attributes. We suggest that the relationships among user attributes can create a network structure among users, and user structure network simultaneously contains the relationships among user attributes. The research considers both the attribute characteristics of users in crowdsourcing environments and the structural relationships among users. It constructs corresponding structural relationship networks based on users' domain-specific characteristics and preference features separately. Building upon this, the research calculates the similarity between users and proposes a community partitioning algorithm based on the heterogeneity of attribute features. The study validates the effectiveness of the algorithm through experimental analysis.

2. User network relationship description

This paper conducts community division based on the heterogeneous characteristics of user attributes. The resulting crowdsourced user community network is a community structure composed of nodes and edges. The nodes contain rich heterogeneous information, encompassing both the domain-specific characteristics and preference features of the user nodes [6]. Edges represent the weight between two user nodes, describing the closeness of their connectivity relationship. Due to the heterogeneous nature of attributes, two users might have different partitioning results based on distinct partitioning criteria. That is to say, while being in the same domain, users might have different interest preferences; and although they have similar interest preferences, they might belong to different domain scopes. Therefore, based on the diversity of relationships among users, this article analyzes two types of network structures. It constructs networks of user-domain relationships and user-preference relationships respectively.

For the user-domain relationship networks, the edges between users indicate whether two users belong to the same domain scope, while the attributes of user nodes represent some interest preferences that users possess in the real world when completing tasks. For the user-preference relationship networks, the edges between users indicate whether two users have similar interest preferences, while the attributes of user nodes represent whether users possess corresponding domain-specific characteristics.

In the user-domain relationship networks, the weight of edges between users indicates the degree of connectivity in the network, which is based on the domain relationships between users. Assuming users U_i and U_j are connected due to the domain relationship D, the value describing the relationship between U_i and U_j should be a non-zero value.

Based on the user-domain feature relationship, we construct a network of domain feature structural relationships among users. Let Ga be an undirected network comprising m nodes, where Ui ($1 \le i \le m$) represents the i-th node in Ga. The set of attribute features preference for its nodes is $U^F = (U_1^F, U_2^F, ..., U_3^F)$. DSij denotes the relationship between the edges of node Ui and node Uj, which represents the domain relationship between two users. The structural network of the user-domain relationship Ga = (U, UF, DS) is illustrated in Fig. 1 as shown below.

Similarly, a structural relationship network based on user preference features among users can be constructed. The weight of edges between users represents the closeness of connection based on preferences, such as users' interests in network programming, English, music, photography, and more.

The structural network of user preference relationships is denoted as Gb = (U, UD, FS), illustrated in Fig. 2 below.

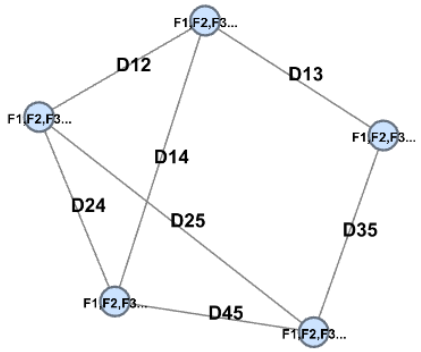

Figure 1. User-Domain Feature Relationship Network Ga.

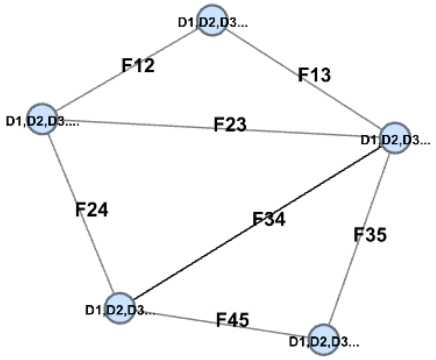

Figure 2. User Preference Feature Relationship Network Gb.

3. Community division based on heterogeneous characteristics

3.1 User Similarity Analysis Based on Domain Structure

Based on the domain features, the structural relationship among users represents whether two users possess similar knowledge backgrounds for a particular issue. In this article, this structural relationship is delineated based on the domains they both belong to. DS_{ij} denotes the strength of the structural relationship between users U_i and U_j. A higher value of DS_{ij} indicates a greater similarity in domain structures between the two users, represented as:

$$DS_{ij} = \begin{cases} N, & N_{(U_i \cap U_j)} \geq 1 \\ 0, & else \end{cases}$$

N represents the number of domains that both users U_i and U_j belong to. When the count of common domains is greater than or equal to 1, then DS = N. otherwise, DS = 0. Similarly, user attribute relationships which are based on the preferences represent the similarity in the ability features between two users. In this paper, the Jaccard similarity is used to quantify the similarity between user attribute features. DN_{ij} represents the similarity of preference features between users U_i and U_j. A higher value of DN_{ij} indicates a greater similarity between the preference features of the two users, expressed as:

$$DN_{ij} = \frac{\left| U_i^F \cap U_j^F \right|}{\left| U_i^F \cup U_j^F \right|}$$

3.2 User Similarity Analysis Based on Preference Structure

Preference-based relationships between user structures indicate whether two users have the same interests for a particular problem. In this paper, the relationship between user preference structures is inscribed based on the preference categories that users have in common. FS_{ij} denotes the degree of strength of the structural relationship between users U_i and U_j. The larger the value of FS_{ij}, the greater the similarity between the two user preference structures, denoted as:

$$FS_{ij} = \begin{cases} N, & N_{(U_i \cap U_j)} \geq 1 \\ 0, & else \end{cases}$$

N represents the number of preferences that both users Ui and Uj share. FS = N when the count of shared preferences is greater than or equal to 1; otherwise, FS = 0. Similarly, the attribute relationship network between user nodes based on domains represents the similarity in knowledge backgrounds between two users. Let FN_{ij} represent the similarity of domain attribute features between users U_i and U_j. A higher value of FN_{ij} indicates a greater similarity between the domain attribute features of the two users.

$$FN(x) = \frac{\left| U_i^D \cap U_j^D \right|}{\left| U_i^D \cup U_j^D \right|}$$

3.3 User Similarity Analysis

Based on the similarity analysis in this paper, the similarity between two users in the user-domain structural network should be characterized by both the structural similarity DS_{ij} among users and the node attribute similarity DN_{ij}. Similarly, in the user-preference structural network, the similarity between two users should also be delineated by both the structural similarity FS_{ij} among users and the node attribute similarity FN_{ij}.

Considering the heterogeneous nature of user attributes, the comprehensive similarity W_{ij} between users based on domain and preference structures is formulated as:

$$W_{ij} = ln^{\alpha \cdot \left(\frac{DS_{ij}}{\left| DS_{ij} \right|} \cdot \frac{DN_{ij}}{\left| DN_{ij} \right|} \right) + \beta \cdot \left(\left(\frac{FS_{ij}}{\left| FS_{ij} \right|} \cdot \frac{FN_{ij}}{\left| FN_{ij} \right|} \right) \right)}$$

DS_{ij} represents the similarity in user domain structure, DN_{ij} represents the similarity in user preference attribute features, FS_{ij} represents the similarity in user preference structure, and FN_{ij} represents the similarity in user domain attribute features. α and β are tuning parameters.

3.4 Community Partitioning Algorithm Based on User Similarity

Based on the above analysis, the goal is to partition the crowdsourced users into several communities, aiming to create communities where the similarity between nodes within each community is maximized. A relatively good community division results in a high similarity of nodes inside the community and a low similarity of nodes outside the community. Modularity is defined as the ratio of the total number of edges inside the community to the total number of edges in the network minus an expectation value. This expectation is the magnitude of the ratio of the total number of edges within a community to the total number of edges in the network formed by the same community assignment when the network is set as a randomized network. In this paper, based on Newman's idea of modularity division, the algorithm for calculating user similarity is integrated into the calculation method of modularity, so the modularity Q is:

$$Q=\frac{1}{2m}\sum_{U_i,U_j}\left[W_{(U_i,U_j)}-\frac{k_{U_i}k_{U_j}}{2m}\right]\cdot\sigma(C_{U_i},C_{U_j})$$

Where m represents the total number of edges in the network, $W_{(U_i,U_j)}$ represents the similarity between two nodes. k_{U_i} represents the degree of node U_i. The σ function's value is defined as follows: if U_i and U_j are in the same community, it equals 1, otherwise it is 0.

4. Experimental Analysis

This paper analyzes the validity of the model based on the experimental data provided by the SQUARE website, which provides a comparative benchmark dataset for crowdsourcing research. It contains 900 English documents, 149 participants, as well as information on participants' domain scope and preference characteristics.

Using the community division method in this paper, the modularity of the community division result is calculated based on the crowdsourcing user domain characteristics and preference features dataset, and the values of the similarity modularity function under different numbers of communities are given, as shown in Fig. 3. As can be seen from Fig. 3, the similarity modularity function obtained from the calculation shows a trend of rising and then falling, and reaches the maximum value (at the vertex) when the number of communities C=5, at which time the corresponding community structure is the final output of the algorithm's community segmentation results, and its visualization results are shown in Fig. 4.

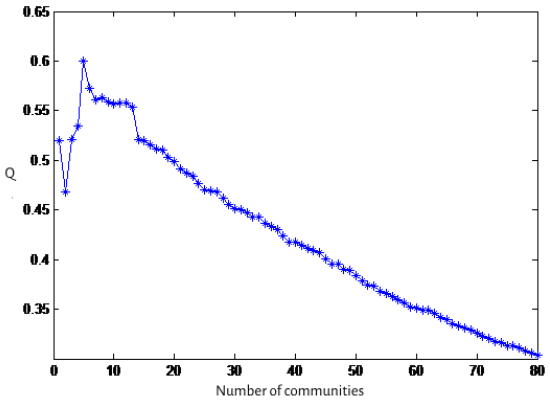

Figure 3. Modularity changes based on the number of communities.

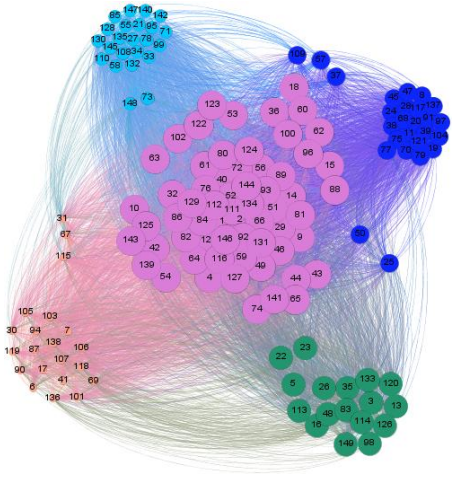

Figure 4. User community division results.

The analysis of Figure 4 shows that of the five communities divided, community C1 (the lower left corner) consists

of 21 users; community C2 (the upper left corner) consists of 23 users; community C3 (the upper right corner) consists of 26 users, of which user 57 community C4 (the lower right corner) consists of 17 users, and community C5 (the middle) is the largest community divided, and it consists of 62 users.

Analysis shows that in community C1, user 31, user 67, and user 115 are more remote from the community and close to C5, indicating that these users may meet the characteristics of community C5 at the same time; in community C2, user 148 and user 73 are more remote from the community, and these users may meet the characteristics of community C5 at the same time; in community C3, user 109, user 57, and user 37 are more remote from the community, these users may meet the characteristics of both C2 or C5 communities, while user 50 and user 25 may meet the characteristics of both C4 or C5 communities; in community C4, user 22 and user 23 may meet the characteristics of both C5 communities.

Analyzing the community as a whole, communities C1, C2, C3, and C4 are smaller in size, and the knowledge background and interest preferences of users in these communities may be more concentrated, which is more suitable for completing tasks with certain professional characteristics. Community C5 has a larger scale and a more dispersed community density, and the knowledge background of users in these communities may be more extensive, which is more suitable for completing popular tasks, and real-life users mostly have this kind of characteristic.

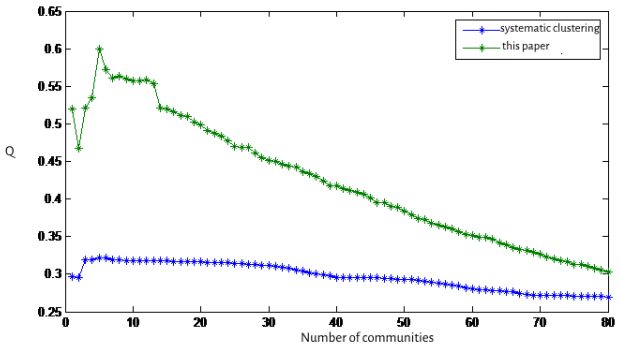

Figure 5. Comparative analysis with the example.

In order to further verify the effectiveness of the method in this paper, we compare it with the community structure obtained by the classical systematic clustering method. The Q function of modularity is used to measure the quality of the community, and the experimental results show the effectiveness of the method.

5. Conclusion

This paper takes users in the crowdsourcing environment as the research object and proposes a community division method based on attribute heterogeneity features by comprehensively considering users' heterogeneous attribute features and structural features. In order to improve the accuracy of the analysis, firstly, the domain similarity between users is determined based on their domain structure network and preference features, and the preference similarity between users is determined based on their preference structure and domain features; then, the comprehensive similarity between any two users in the network structure is calculated based on the domain similarity and preference similarity between users; Finally, based on the comprehensive similarity between users, the evaluation function of community structure is constructed, and the systematic clustering method is used for community division. This method can effectively mine the real structure inside the user community and can obtain both decentralized large communities and centralized small communities. In addition, the community division of crowdsourcing users can realize reasonable grouping of users, which helps crowdsourcing platforms to design tasks according to the distribution characteristics of users, thus improving the accuracy of the platform's task recommendation.

Acknowledgements

The authors are grateful to the reviewers for their valuable suggestions on how to improve the paper. The author acknowledges the support of Project No. 20KJB630012 funded by the University Science Research Project and 2020SJA0344 funded by Philosophy and Social Science Research of Jiangsu Province.

References

[1] Yang, M., Ooi, Y. M., & Han, C. (2022). Lead users as idea supplier in online community platform: How to choose the right ideas to implement? International Journal of Production Economics, 244.

[2] Gu, Z., Bapna, R., Chan, J., & Gupta, A. (2022). Measuring the Impact of Crowdsourcing Features on Mobile App User Engagement and Retention: A Randomized Field Experiment. Management Science: Journal of the Institute of Management Sciences, (2), 68.

[3] Fuger, S., Schimpf, R., Füller, J., & Hutter, K. (2017). User roles and team structures in a crowdsourcing community for international development—a social network perspective. Information Technology for Development, 1-25.

[4] Kong, Z, Sun, Q., Kou X.Y., Wang, L.F. (2022). Research on the Importance of Network Nodes Based on Attribute Information and Structural Characteristics. Journal of Northeastern University (Natural Science), 43(5), 625-631.

[5] Yan, J. K., Leidner, D., & Benbya, H. (2023). The Differential Innovativeness Outcomes of User and Employee Participation in an Online User Innovation Community. Social Science Electronic Publishing, 35(03), 1-44.

[6] Guo, X., Omar, M. H., & Zaini, K. M. (2020). Multiattribute Access Selection Algorithm Supporting Service Characteristics and User Preferences in Heterogeneous Wireless Networks. Wireless Communications and Mobile Computing, 2020, 1-27.