Postprint: Research on Group Profiling Construction Methods Based on Network Structure and Text Content

Authors: QIU Yunfei, Zhang Weizhu

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] In user profiling research based on social networks, to address the problems that traditional user modeling struggles to handle complex network relationships, group construction is mostly content-based, and groups exhibit low similarity or poor tightness, we propose a group profiling construction method based on network structure and text content. [Method/Process] First, we adopt the convolutional neural network method to fuse features from both network structure and text content, representing network users as spatial vectors; second, we combine modularity calculation methods on the basis of the k-means algorithm to cluster the spatial vectors; then, we conduct comparative studies on crawled Chinese and English datasets respectively; finally, we select 1,000 important users from the Chinese dataset for case analysis. [Results/Conclusion] Experimental results show that the density value of this method increases by an average of 0.105 compared to content-based methods, the entropy value decreases by an average of 0.955 compared to structure-based (including structure and content) methods, and case analysis further demonstrates the feasibility of the proposed method.

Full Text

Preamble

Research on Group Profile Construction Method Based on Network Structure and Text Content

Qiu Yunfei, Zhang Weizhu
School of Software, Liaoning Technical University, Huludao 125105

Abstract

[**Purpose/Significance**] In user profiling research based on social networks, traditional user modeling struggles to handle complex network relationships, group construction relies primarily on content, and groups exhibit low similarity or poor cohesion. To address these issues, this paper proposes a group profile construction method that integrates network structure and text content.

[**Method/Process**] First, we employ a convolutional neural network approach to represent network users as spatial vectors by fusing features from both network structure and text content. Second, building upon the k-means algorithm and incorporating modularity calculation methods, we cluster these spatial vectors. We then conduct comparative studies on crawled Chinese and English datasets. Finally, we select 1,000 influential users from the Chinese dataset for case analysis.

[**Result/Conclusion**] Experimental results demonstrate that the proposed method increases density values by an average of 0.105 compared to content-based methods, while reducing entropy values by an average of 0.955 compared to structure-based methods (including those based on both structure and content). The case analysis further validates the feasibility of the proposed approach.

Keywords: social network; network relationship; text content; deep learning; clustering algorithm; user profile
Classification Number: TP391
DOI: 10.13266/j.issn.0252-3116.2019.22.003

Introduction

In the Internet era, people interact with various online platforms daily, where they can follow topics of interest, browse preferred content, add friends, or gain followers. Through these behaviors, users establish connections with others, much like in real life. These connections evolve with changes in relationships, ultimately forming networks. Such networks encompass not only social platforms like Weibo, Zhihu, and Twitter, but also citation networks (e.g., CNKI, IEEE Xplore) and propagation networks (e.g., Douban, Digg). Beyond directly observable data such as user attributes, published content, and follower counts, these networks contain substantial indirect relationship data, including follow relationships and citation links.

Scholars worldwide have conducted research on network representation learning, community detection, and user profiling based on network relationships. User profiling has emerged as a particularly active research area in recent years, delivering significant value for personalized services, recommendation systems, and precision marketing. Effectively leveraging social network data to accurately and comprehensively characterize user profiles has become a key research direction.

Early user profiling studies focused on predicting user attributes, interests, behaviors, and credibility based on published text, followed content, and online reviews. For instance, S. Alaoui et al. detected user purchase intentions from semantic analysis of user-generated text and combined this with demographic attributes for product recommendations. W. X. Zhao et al. identified purchase intentions from users' follow content and published text on microblogs, constructing social media user profiles and e-commerce user profiles, then mapping and associating these profiles to enable social-media-based product recommendations. Shan Xiaohong et al. built a user profile conceptual model using online review data from Ctrip hotels to characterize hotel user features. Yu Chuanming et al. performed deep representation learning on stock forum users' posts and combined behavioral features (follower count, following count, posting frequency, comment volume, forum age) to propose a behavior-content fusion model for identifying noise investors. Guo Guangming processed and analyzed multi-source heterogeneous data to construct user credibility profiles and predict user credibility scores. Fan Xiaoyu et al. integrated data from multiple sources including personal homepages, CNKI, and funding networks to propose a researcher profiling method fusing multi-source heterogeneous data, profiling researchers from three dimensions: basic attributes, research preferences, and research relationships.

As research progressed, scholars recognized that users on online platforms establish connections through adding friends, mutual following, and text citation. Consequently, researchers began leveraging network relationships for user modeling and label prediction. A. Mislove et al. constructed network topology from follow relationships in social networks for community detection, using clustering algorithms to predict unknown user attributes based on known attributes. Cao Jiuxin et al. utilized Weibo follow relationship topology with probabilistic cascade models and machine learning to predict user forwarding behavior. Liu Kan et al. fused user behavior, published content, and social relationships using random forest algorithms to build identification models for Weibo bot users. Xu Zhiming et al. treated Weibo social networks as weighted undirected graphs, determining user similarity based on edge weights. Although these studies incorporated network relationships, machine learning methods in complex networks have limitations, particularly when large training sets are unavailable, resulting in low modeling accuracy for traditional approaches.

To simplify complex relationship processing, scholars construct network groups for holistic user analysis. Lin Yanxia and Xie Xiangsheng defined Weibo topics based on social identity theory, classified Weibo content by topic, and implemented group classification using multidimensional scaling. Zhang Hongxin et al. constructed topic models from mobile terminal log data, clustering users based on correlation between log data and topics. Xiong Wei et al. used LDA topic models to categorize webpage content by topic and performed group profiling based on user behavioral information. While these methods avoid complex network relationship processing, they cluster solely based on user text content, achieving content similarity but lacking structural cohesion in groups.

Consequently, some researchers focus on group construction based on network structure to improve cohesion. V. D. Blondel et al. proposed community construction methods from social network graphs. J. Leskovec et al. constructed directed unweighted graphs to cluster social users into groups. While structure-based clustering improves group cohesion, these groups may not be similar in content or attributes. Therefore, K. Steinhaeuser and N. V. Chawla incorporated node attributes as edge weights in network graphs, proposing a random walk-based group construction method. Y. Zhou et al. defined a distance metric combining structural and attribute similarity, adding node attributes and edges to graphs for group construction. Z. Xu et al. proposed a Bayesian probability model-based graph clustering method that models both graph structure and attribute information without distance calculation. Chen Kehan et al. clustered groups by integrating graph summarization methods based on similarity of user Weibo content for user interest recommendation. Wu Shufang et al. measured user similarity by linearly harmonizing in-link and out-link label similarities based on user relationships. Although these studies strive for content similarity while considering structure, group cohesion and similarity are often incompatible—groups with high similarity exhibit low cohesion, and vice versa, failing to achieve ideal group construction.

2 Research Methods

2.1 Network User Representation

On various online platforms, users build network relationships through adding friends or following behaviors. Once interaction occurs, users' information may change, potentially exhibiting different characteristics based on interaction partners or engaging with different users based on followed content. For example, in a real-world academic network with three scholars A, B, and C—where A researches deep learning and data mining, B researches data mining and natural language processing, and C researches natural language processing and user profiling—scholar B may collaborate with A on data mining and with C on natural language processing. We can further infer that scholars A and C might collaborate through B, applying deep learning methods to user profiling research. Traditional methods struggle to accurately model such complex network relationships. To represent network users more accurately, this study employs deep learning methods for user modeling through deep training, considering not only explicit network relationships but also inferring implicit network relationships from contextual semantic information in text.

2.1.1 Method Overview In network representation methods, neural network-based approaches typically establish a loss function during model construction, optimizing it to find more suitable parameter values and build more accurate models. Compared to matrix-based methods, neural network approaches often offer faster computation, higher execution efficiency, and improved accuracy. Therefore, this study uses a neural network-based method

to represent users as spatial vectors. To account for both structural and content features, we divide the representation into structural representation vectors and content representation vectors.

Recent classic neural network-based network structure representation methods include DeepWalk, LINE, and Node2vec. Among these, only LINE generates context-dependent node representations. Since the "content representation vectors" generated from node content later are also context-dependent, adopting the LINE method facilitates merging the two representations. Additionally, LINE employs first-order and second-order proximity (first-order proximity refers to directly connected nodes, while second-order proximity refers to nodes connected through intermediate nodes), which better reflects real-world user relationships (users establish connections through direct following or by following the same user). Due to complex network structures and large information volumes, machine learning methods have low processing efficiency, whereas convolutional neural network methods operate extremely fast in GPU-configured environments. Moreover, when representing text content features, convolutional neural networks offer greater advantages than n-grams in characterizing high-dimensional features. Therefore, this study implements structural representation vectors using the LINE model proposed by J. Tang et al., content representation vectors using convolutional neural networks, and obtains final user representation vectors by summing the two representation vectors.

The convolutional neural network architecture consists of three layers. User text content and network relationships serve as the input layer, generating text matrices. Convolution and pooling operations are performed on these matrices as the hidden layer. The output layer uses a softmax function to normalize hidden layer results into unit vectors, which are then multiplied by text matrices to obtain content representation vectors. Figure 1 [Figure 1: see original paper] illustrates the process of generating content representation vectors through convolutional neural networks.

2.1.2 Method Implementation (1) Process Description. Define the network graph as $G = (V, E, T)$, where V represents network nodes, $E \subseteq V \times V$ denotes edges representing relationships between nodes, and T represents node text content. As shown in Figure 1 [Figure 1: see original paper], taking two nodes $u \in V$ and $v \in V$ on edge $e(u, v) \in E$ as an example, the specific process for obtaining content representation vectors using convolutional neural networks is as follows:

First, convert the follow content of both nodes into word sequences S_u and S_v. Through convolutional layer operations, generate two text matrices $P \in \mathbb{R}^{d \times m}$ and $Q \in \mathbb{R}^{d \times n}$, where m and n represent the lengths of S_u and S_v respectively, and d denotes spatial dimension.

Second, introduce an auxiliary matrix $A \in \mathbb{R}^{d \times d}$ to compute the correlation

matrix $F \in \mathbb{R}^{m \times n}$ as:

$$F = \tanh(P^T A Q)$$

Third, apply mean pooling to rows and columns of matrix F for row and column pooling respectively, obtaining pooled vectors for P and Q:

$$p = [p_1, \ldots, p_m]^T, \quad q = [q_1, \ldots, q_n]^T$$

Fourth, use the softmax normalization function to convert pooled vectors of P and Q into unit vectors a_p and a_q. The i-th element of a_p is expressed as:

$$a_{p_i} = \frac{\exp(p_i)}{\sum_{j \in [1,m]} \exp(p_j)}$$

where $p_i = \text{mean}(F_{i1}, \ldots, F_{in})$ represents the i-th element of p, i.e., the row-pooling result for each row of matrix F. The calculation of a_q follows the same process as a_p.

Fifth, the product of the unit vector and text matrix yields the content representation vector:

$$u_c = P a_p, \quad v_c = Q a_q$$

(2) Loss Function. In statistics and machine learning, loss functions typically measure error and loss magnitude. This study employs the loss function for user modeling evaluation, aiming to minimize the loss function and improve modeling accuracy. The loss function is defined as the negative of the conditional probability of predicting interaction object node v from node u. The conditional probability of u generating v is defined as:

$$\log(v|u) = \alpha \cdot \log(v_s|u_s) + \beta \cdot \log(v_s|u_c) + \beta \cdot \log(v_c|u_s) + \gamma \cdot \log(v_c|u_c)$$

where:

$$\log(v_s|u_s) = \frac{\exp(u_s \cdot v_s)}{\sum_{x \in V} \exp(u_s \cdot x_s)}$$

Here, u_s, v_s, and x_s represent the structural representation vectors of nodes u, v, and x respectively, while α, β, and γ denote parameters.

The overall model loss function is defined as:

$$\text{Loss} = -\sum_{e \in E} \log(v|u)$$

2.2 Network Group Construction

User representation vectors obtained from model training serve as input for clustering algorithms. We propose a group construction method based on network structure and text content (GCNSTC) that clusters users with close relationships and high similarity into groups. The k-means algorithm is employed for

clustering, with modularity used to evaluate clustering results through iterative updates until results stabilize.

Modularity, proposed by M. E. J. Newman, measures the strength of network community structure. To achieve better clustering performance, we define modularity from both network structure and text content perspectives, using the weighted sum of structural modularity and content modularity as the final evaluation metric.

(1) Structural Modularity. Calculated using the traditional Newman modularity formula:

$$M_s = \sum_{(x,y)\in C} \left[\frac{A_{xy}}{2m} - \frac{d_x \cdot d_y}{(2m)^2} \right] \delta(C_x, C_y)$$

where x and y represent nodes; C_x and C_y represent groups; A_{xy} equals 1 when nodes x and y are connected and 0 otherwise; d_x denotes the degree of node x; m represents the total number of edges in the network, with $2m$ being the total degree of the network; $\delta(C_x, C_y)$ equals 1 when x and y belong to the same group and 0 otherwise.

(2) Content Modularity. Node text content is converted into measurable vector representations, and cosine similarity is used to calculate content modularity:

$$M_c = \sum_{(x,y)\in C} \frac{\cos(x_i, y_i)}{2m} \delta(C_x, C_y)$$

where:

$$\cos(x_i, y_i) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum x_i^2} \cdot \sqrt{\sum y_i^2}}$$

and d represents the dimension of text representation vectors.

(3) Final Modularity Metric. Structural modularity and content modularity are weighted and summed to obtain the final modularity evaluation metric M:

$$M = \omega M_s + (1 - \omega) M_c$$

where ω is a weighting factor with $0 < \omega < 1$.

Based on structural and content modularity, the algorithm for clustering network users is presented as Algorithm 1:

Algorithm 1: Clustering Algorithm Based on Structural-Content Modularity
Input: User representation vectors, user follow content, initial group count k
Output: Group count K, set of groups

1. Call k-means clustering algorithm to obtain k groups
2. **Repeat**

3. **for** each node i **do**
4. **for** each group j **do**
5. **if** node i is not in group j **then**
6. Remove node i from its current group and add it to group j
7. Calculate modularity increment after adding node i
8. **endif**
9. **endfor**
10. Select group j with maximum modularity increment and move node i to group j; otherwise, node i remains
11. **endfor**
12. **until** groups no longer change

3 Data Acquisition and Preprocessing

This study employs both Chinese and English datasets for comparative experiments to demonstrate the feasibility and general applicability of the proposed method across languages. Since the research aims to fuse network structure and text content features for network user modeling, the collected data includes both text data and network structure data.

3.1 Data

3.1.1 Zhihu Dataset. We used the web crawler tool "ache" to crawl Zhihu users' followed topic content and follow relationships between users. The ache tool returns relevant search pages based on specified search topics or attribute content. When configuring ache, we set the crawl content to "follow topic description content," the number of users to crawl to 10,000, and the number of followed topics per user to $ $3. For the crawled text, we first used Python regular expressions to remove HTML tag content, then applied jieba segmentation for word segmentation to convert text representation into word sequence representation. Since Chinese text processing often encounters encoding issues, we used GBK encoding for reading data and segmentation processing, then stored the data in UTF-8 encoding after segmentation. Segmented text typically contains invalid words such as "这" (this), "这个" (this one), "了" (particle), "什么" (what), and "呢" (particle). We removed these invalid words using a Chinese stopword list to obtain the final Chinese text dataset. For network structure data, each user has a unique ID. When a follow relationship exists between users, both users' ID information is stored in the dataset; otherwise, no storage is required.

We used the LDA topic model to classify followed content by topic. Table 1 shows the Zhihu text dataset (partial list). The Zhihu structure dataset contains 10,000 nodes and 43,894 edges.

3.1.2 Cora Dataset. Cora is a crawled citation network, but the network data is relatively large. This study filtered articles related to "machine learning" as text content. We preprocessed the obtained text using the Python library

pyenchant for spell checking and correction, removing misspelled words like "liike" and "lke," then performed English text segmentation. For English text segmentation, we used Python' s nltk SnowballStemmer class for stemming and WordNetLemmatizer class for lemmatization. Additionally, since English letters have case distinctions (e.g., "Hello" and "hello" represent the same meaning but are often treated as different words due to case differences), we converted all uppercase letters to lowercase using Python' s API. Finally, English text typically contains invalid words like "of," "to," "an," and "a," which we removed using an English stopword list to obtain the required English text dataset. For network structure data, each article has an ID, and citation relationships were identified and stored based on the filtered article IDs.

From the Cora citation network, we selected 2,277 machine learning-related papers, which were divided into 7 categories based on research content. Table 2 shows the Cora text dataset (partial list). The Cora structure dataset contains 2,277 nodes and 5,214 edges.

3.2 Comparison Methods

This study employs three group clustering algorithms as baselines for comparative experiments from three perspectives: text content-based, network structure-based, and structure-content-based approaches.

(1) K-means Algorithm: The content-based clustering algorithm uses K-means as the baseline, converting text content into vector form and clustering based on text distance calculation.

(2) Louvain Algorithm: Pan Li et al. noted that community detection algorithms focus only on structural density of clustering results without considering node attribute information. Therefore, the structure-based method adopts the classic Louvain algorithm from community detection, which constructs weighted networks and uses node relationships to build communities by calculating modularity gain when adding nodes to neighbor communities.

(3) SA-Cluster Algorithm: The structure-content-based method employs the SA-Cluster algorithm proposed by Y. Zhou et al., which adds node attribute information to networks to build augmented networks, then defines structural and attribute similarity and uses random walk algorithms to calculate distances between network nodes for group construction.

3.3 Evaluation Metrics

We evaluate the above group construction methods using density and entropy metrics. Density primarily reflects the closeness of relationships among group members, with higher density values indicating tighter relationships. The density calculation formula is:

$$\text{Density} = \frac{2\sum_i m_i}{\sum_i n_i(n_i - 1)}$$

where k represents the number of groups, m represents the total number of edges in the network, and m_i represents the number of edges within group i.

Entropy, originally a thermodynamic measure of system disorder, here reflects similarity among group members. If a social network node joining a group increases entropy, the node' s introduction brings additional information, indicating greater difference from other group members. Therefore, lower entropy values correspond to lower disorder and higher similarity among group members. The entropy calculation formula is:

$$\text{Entropy} = -\sum_i \frac{n_i}{n} \sum_j p_{ij} \log(p_{ij})$$

where n represents the total number of nodes in the network, n_i represents the number of nodes in group i, and p_{ij} represents the percentage of nodes in group i with category j.

3.4 Experimental Results and Analysis

We compare the proposed method with the three group construction methods described above, conducting comparative experiments on both Zhihu and Cora datasets (with $\omega = 0.5$) and evaluating them using density and entropy metrics. Table 3 shows the experimental results on the Zhihu dataset, while Table 4 shows the results on the Cora dataset.

On the Zhihu dataset, the GCNSTC method' s density is higher than K-means but lower than Louvain and SA-Cluster, indicating that GCNSTC outperforms K-means in group cohesion but is slightly inferior to Louvain and SA-Cluster. Through entropy comparison, K-means achieves an entropy value of 0, indicating highest similarity among group members, while GCNSTC' s entropy is smaller than both Louvain and SA-Cluster, demonstrating better similarity among group members.

In the Cora dataset, GCNSTC' s density value is 0.57, again higher than K-means but lower than Louvain and SA-Cluster, indicating that GCNSTC produces groups with relatively ideal cohesion. K-means still achieves the lowest entropy among the four methods, showing best group similarity, but GCNSTC' s entropy differs from Louvain and SA-Cluster by a factor of 1/2, significantly outperforming them in group similarity.

3.5 Discussion

Group Similarity: The experimental results show that K-means achieves the lowest entropy values across both datasets, indicating highest group similarity. However, GCNSTC' s entropy is lower than Louvain and SA-Cluster by an average of 0.955, demonstrating superior group similarity compared to structure-based methods (including those based on both structure and content).

Group Cohesion: Although GCNSTC improves upon K-means in density, it is inferior to the other two methods. This occurs because our structural representation vector processing uses the LINE model, a shallow neural network model. When merged with content representation vectors from deep models, the deep model' s more precise results relatively weaken the structural representation vector' s contribution to the overall user representation vector, making the user representation vector more content-oriented. Consequently, GCNSTC' s structural cohesion only outperforms content-based K-means clustering.

Based on experimental results, we rank the four methods by group cohesion and similarity. For cohesion: SA-Cluster > Louvain > GCNSTC > K-means, with GCNSTC ranking third, only outperforming K-means. For similarity: K-means > GCNSTC > Louvain > SA-Cluster, with GCNSTC achieving second-best performance, optimal compared to structure-based methods. This demonstrates that the proposed method improves both group cohesion and similarity, with more pronounced improvement in similarity.

To further illustrate this conclusion, Table 5 compares density and entropy changes between GCNSTC and the other three methods. Although GCNSTC improves density relative to K-means, its overall performance considering entropy is inferior to K-means, which clusters purely based on text similarity using Euclidean distance. However, for SA-Cluster and Louvain, GCNSTC shows significant entropy improvement, making it overall superior to both methods. In summary, the proposed method is effective, particularly in entropy reduction, producing groups with higher similarity though slightly lower cohesion.

4 Case Study

We analyzed the Chinese dataset by counting each user' s follower numbers in descending order. Among the top 50 users, all IDs were within 1000; among the top 100, only two IDs exceeded 1000; and among the top 200, approximately 89% of user IDs were within 1000. Users with IDs below 1000 and their followers constitute about 80% of the entire network. We designate these 1000 users as influential users in the network.

We selected these 1000 users for analysis from both content similarity and structural cohesion perspectives. Each user follows 1-3 topics. Using our clustering method, users were grouped into 10 clusters. We defined major topics as those accounting for over 10% of a group' s total topic follows. Table 6 summarizes each group' s member count, total topic follows, and major topics.

From the major topic perspective, six groups have major topic proportions exceeding 50%, indicating that users'followed topics generally align with the group' s overall content, showing high content similarity. Although Group 3' s major topics account for only 43.9% of total follows, it has only two major topics, making group users relatively similar overall. Groups 2, 8, and 10 show less ideal clustering with lower major topic proportions, indicating that group members

follow more topics with each topic having relatively low follow counts (below 10%), suggesting greater differences among members.

From the average major topic perspective, Group 5 shows the best clustering effect with "entrepreneurship" as its sole major topic, accounting for 61.5% of follows, indicating that this user group primarily focuses on entrepreneurship content. Although Group 7' s major topic proportion reaches 75.4%, its average major topic proportion is only 18.9%, indicating that the group overall follows many main topics with low similarity.

For these 10 groups, we counted each user's followed users and sorted them in descending order. Since some groups have few members while some followed users have multiple followers, Table 7 summarizes followed users with follow counts >10 among the top 3. The results show that Groups 2, 8, and 10 have over half their members following the same users, forming subgroups centered around followed users that constitute 50% of the entire group, making these groups relatively structurally compact. Groups 3 and 9 have lower cohesion than the first three groups, containing some unlisted small subgroups that weaken overall cohesion. The remaining five groups have less than 50% of members following the same user, with their largest subgroups similar to Groups 3 and 9, but containing some relatively larger subgroups that significantly impact overall structure. For example, Group 4 has two listed subgroups of similar size, effectively dividing Group 4 into two subgroups. While subgroups are internally cohesive, the overall integrity is poor. Additionally, these five groups contain some unlisted small groups, making the overall structure relatively loose.

This case analysis demonstrates that Groups 2, 8, and 10 have poorer content similarity but relatively compact structure, while other groups show high content similarity but relatively loose structure, further illustrating that group construction based on network structure and text content struggles to simultaneously optimize structural cohesion and content similarity. The proposed method achieves greater improvement in entropy, tending toward content similarity. Therefore, in the analysis of 1000 influential users, 7 out of 10 constructed groups show high content similarity. Although structure is relatively loose, groups can still be divided into moderately cohesive subgroups. Additionally, some users follow multiple users with overlapping follow counts, reducing overall structural cohesion.

5 Group Profile Analysis and Research Significance

Using Group 6 as an example, we further analyze the significance of group profile research.

(1) Holistic Analysis for User Analysis, Product Recommendation, and Industry Trend Prediction. Statistics show Group 6 primarily focuses on four areas: Internet, movies, travel, and lifestyle. The interest in movies, travel, and lifestyle indicates that users value quality of life and entertainment. The combination of Internet and movies suggests some users may work in In-

ternet film or be Internet entrepreneurs interested in business or investment. Some users follow psychology knowledge to manage work stress and psychological pressure. We can recommend new movie releases, tourist attractions, or travel guides to this group. The highest follow count is for the Internet, followed by movies and travel. In the "Internet Plus" era, we can predict development trends in film and tourism industries. For example, in Internet Plus film, besides online reviews and ticket purchasing, users can watch film works anytime on mobile terminals. In Internet Plus tourism, we can create communities on social networks to share travel notes, stimulate travel interest, design travel guide software to help people quickly make decisions about destinations, accommodation, and transportation, and facilitate ticket/hotel booking through e-commerce platforms like Ctrip.

(2) Content-Based Analysis for Group Message Push, Group Recommendation, and Friend Recommendation. Table 8 shows partial results of users following each topic in Group 6. Users in the same topic group share identical follow content, enabling group message push or recommendation. For new movie releases or highly-rated films, we can push recommendations to the movie group or push film industry news. When a new album requires promotion, we can recommend it to the music group, with refined recommendations based on music genres. Additionally, collaborative filtering enables friend recommendation and personalized services. For instance, if User 21 wants to learn psychology, we can recommend User 554 to User 21 because both follow movies and Internet topics, suggesting similar psychology interests. If User 662 wants to start an Internet business, we can recommend User 662' s designs to User 129 because both follow Internet and economics topics, indicating they may work in Internet-related fields. Since User 129 also follows e-commerce, they can publish User 662' s products on e-commerce platforms to facilitate online economic activities.

(3) Structure-Based Analysis for User Analysis and Friend Recommendation. Statistics show approximately 40% of Group 6 users follow User 2. Table 9 shows User 2' s partial followed users and their topics. Users following User 2 likely focus on entertainment and quality of life, primarily watching movies, traveling, and listening to music. Professionally, they may work in Internet film or be Internet entrepreneurs interested in business or investment. Using User 2 as an intermediary enables friend recommendation. For example, if User 92 wants music information, we can recommend Users 15, 404, or 944 to User 92 because User 2 has follow relationships with all four users. We can connect User 92 with three users through User 2, prioritizing Users 15 and 404 due to more shared preferences. If User 404 wants to choose a car for self-driving tours, we can recommend User 426 because both follow travel topics and User 426 follows car topics, likely having knowledge about suitable cars for self-driving tours. Since both users follow User 2, they can become friends through this connection.

Conclusion

Based on complex relationships among users in real-world networks, this study employs deep learning methods to represent network users as spatial vectors, modeling and clustering network users from both network structure and text content perspectives. We use modularity to measure clustering strength and iteratively improve clustering performance. To validate feasibility, we conduct comparative experiments on Chinese and English datasets against three different clustering algorithms. Experimental results demonstrate the method' s general applicability across languages, with density values increasing by an average of 0.105 compared to content-based methods and entropy values decreasing by an average of 0.955 compared to structure-based methods (including structure-content methods), simultaneously improving group cohesion and similarity.

We discuss experimental results, explaining why neither density nor entropy reaches optimal values. By comprehensively analyzing changes in density and entropy, we demonstrate more pronounced effects on entropy reduction. The case analysis shows approximately 7 out of 10 groups exhibit high content similarity. Analysis of 1000 influential users illustrates that group construction based on network structure and text content struggles to simultaneously optimize structural cohesion and content similarity, though the proposed method achieves greater entropy improvement. This research holds significant importance for product recommendation, industry prediction, personalized services, message push, user analysis, and friend recommendation.

Current group profiling research on social networks faces challenges including single text features for group construction, generalized group profile characterization, and insufficient research on group structure. Future research will incorporate multi-feature data to analyze and predict other user attribute labels, with focused investigation on group structure.

References

[1] He Juan. Research on personalized book recommendation application based on combination of user personal and group profiles[J]. Information Studies: Theory & Application, 2019, 42(1): 129-133, 160.

[2] Zhao W X, Wang J, He Y, et al. Mining product adoption information from online reviews for improving product recommendation[J]. ACM transactions on knowledge discovery from data, 2016, 10(3): 1-23.

[3] Liu Hai, Lu Hui, Ruan Jinhua, et al. Research on precision marketing segmentation model based on "user profile" mining[J]. Journal of Silk, 2015, 52(12): 37-42, 47.

[4] Alaoui S, Ajhoun R, Idrissi Y E B, et al. Semantic approach for the building of user profile for recommender system[C]//Global summit on computer & information technology. Sousse: IEEE, 2016: 114-119.

[5] Zhao W X, Guo Y, He Y, et al. We know what you want to buy: a demographic-based system for product recommendation on microblogs[C]//ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 1935-1944.

[6] Zhao W X, Li S, He Y, et al. Exploring demographic information in social media for product recommendation[J]. Knowledge and information systems, 2016, 49(1): 61-89.

[7] Shan Xiaohong, Zhang Xiaoyue, Liu Xiaoyan. Research on user profile based on online reviews: a case study of Ctrip hotels[J]. Information Studies: Theory & Application, 2018, 41(4): 99-104, 149.

[8] Yu Chuanming, Tian Xin, Guo Yajing, et al. User profile research based on behavior-content fusion model[J]. Library and Information Service, 2018, 62(13): 54-63.

[9] Guo Guangming. Research on user credit profile method based on social big data[D]. Hefei: University of Science and Technology of China, 2017.

[10] Fan Xiaoyu, Dou Yongxiang, Zhao Pengwei, et al. Research on researcher profile construction method fusing multi-source data[J]. Library and Information Service, 2018, 62(15): 31-40.

[11] Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: inferring user profiles in online social networks[C]//ACM international conference on web search and data mining. New York: ACM, 2010: 251-260.

[12] Cao Jiuxin, Wu Jianglin, Shi Wei, et al. Analysis and prediction of information dissemination in Sina Weibo network[J]. Chinese Journal of Computers, 2014, 37(4): 779-790.

[13] Liu Kan, Yuan Yunyun, Liu Ping. Research on Weibo bot user identification based on random forest classification[J]. Journal of Computer Research and Development, 2014, 37(1): 207-218.

[14] Xu Zhiming, Li Dong, Liu Ting, et al. Similarity measurement and application of microblog users[J]. Journal of Computer Research and Development, 2014, 37(1): 207-218.

[15] Lin Yanxia, Xie Xiangsheng. Microblog group user profiling based on social identity theory[J]. Information Studies: Theory & Application, 2018, 41(3): 142-148.

[16] Zhang Hongxin, Sheng Fengfan, Xu Peiyuan, et al. Crowd feature visualization based on mobile terminal log data[J]. Journal of Software, 2016, 27(5): 1174-1187.

[17] Xiong Wei, Hang Bo, Li Bing, et al. A service redirection method integrating user profile and content[J]. Journal of Peking University (Natural Science Edition), 2015, 51(2): 289-300.

[18] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10): 10008-10019.

[19] Leskovec J, Lang K J, Mahoney M W. Empirical comparison of algorithms for network community detection[C]//ACM international conference on World Wide Web. Raleigh: ACM, 2010: 631-640.

[20] Steinhaeuser K, Chawla N V. Identifying and evaluating community structure in complex networks[J]. Pattern recognition letters, 2010, 31(5): 413-421.

[21] Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities[J]. Proceedings of the VLDB endowment, 2009, 2(1): 718-729.

[22] Xu Z, Ke Y, Wang Y, et al. A model-based approach to attributed graph clustering[C]//ACM SIGMOD international conference on management of data. Scottsdale: ACM, 2012: 505-516.

[23] Chen Kehan, Han Panpan, Wu Jian. Heterogeneous social network recommendation algorithm based on user clustering[J]. Chinese Journal of Computers, 2013, 36(2): 349-359.

[24] Wu Shufang, Xu Jianmin, Wu Xiaobo. Similarity measurement of microblog users fusing user tags and relationships[J]. Journal of Intelligence, 2014, 33(12): 170-173, 126.

[25] Tang J, Qu M, Wang M, et al. LINE: large-scale information network embedding[C]//International conference on World Wide Web. Florence: WWW, 2015: 1067-1077.

[26] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E statistical nonlinear soft matter physics, 2003, 69(6): 066133.

[27] McCallum A K, Nigam K, Rennie J, et al. Automating the construction of internet portals with machine learning[J]. Information retrieval journal, 2000, 3(2): 127-163.

[28] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.

[29] Pan Li, Wu Peng, Huang Danhua. Research progress on online social network group discovery[J]. Journal of Electronics & Information Technology, 2017, 39(9): 2097-2109.

Author Contributions

Qiu Yunfei: Proposed research ideas and guided paper revision.
Zhang Weizhu: Designed research scheme, conducted experiments, and wrote the paper.

Note: Figure translations are in progress. See original paper for figures.