Postprint: Multi-dimensional Attribute Weighted Analysis for Weibo User Clustering

Authors: Zhang Haitao, Tang Shiman, Wei Mingzhu, Li Zezhong

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] Accurately grasping user interest tendencies in social networks, classifying users, and forming highly aggregated user groups is of great significance for research on social network information ecology and information recommendation.

[Method/Process] By constructing a hierarchical model for user attribute description based on multiple dimensions, crawling user sample data from Sina Weibo according to the model's data requirements, quantifying second-order variables under the multi-dimensional attributes of relevant user background information, user post information, and user behavior information, constructing user vector expressions, comparing user classification effects between single-dimensional and multi-dimensional scenarios, further performing weighted analysis by assigning different weight values to attributes, conducting variance analysis after obtaining optimal clustering results, and improving the model.

[Results/Conclusion] User clustering effects based on weighted multi-dimensional attributes are significantly higher than those under single-dimensional and multi-dimensional non-weighted conditions, and the user post content dimension has the greatest effectiveness in improving user clustering results.

Full Text

Preamble

Vol. 62 No. 24 December 2018
Research on Microblog User Clustering Based on Multi-Dimensional Attribute Weighted Analysis
Zhang Haitao[1],[2], Tang Shiman[1], Wei Mingzhu[1], Li Zezhong[1]
[1] School of Management, Jilin University, Changchun 130022
[2] Information Resource Research Center, Jilin University, Changchun 130022

Abstract: [Purpose/Significance] Accurately grasping the interest tendencies of social network users, classifying users, and forming highly aggregated user groups is of great significance for studying social network information ecology and information recommendation. [Method/Process] By constructing a hierarchical model for describing user attributes based on multiple dimensions, sample user data were crawled from Sina Weibo according to the model's data requirements. Second-order variables under multi-dimensional attributes related to user background information, user blog information, and user behavior information were quantified to construct user vector expressions. The classification effects under single-dimensional and multi-dimensional conditions were compared, and further weighted analysis was conducted by assigning different weights to attributes. After achieving optimal clustering results, variance analysis was performed to improve the model. [Result/Conclusion] User clustering based on multi-dimensional attribute weighting is significantly more effective than user clustering under single-dimensional and multi-dimensional non-weighted conditions, and the user blog content dimension contributes most to improving the validity of user clustering effectiveness.

Classification Number: G250
Keywords: Microblog, Multi-dimensional, User Clustering, Weighted Analysis
DOI: 10.13266/j.issn.0252-3116.2018.24.016

Introduction

In recent years, social media riding the wave of rapidly developing internet technology has become increasingly pervasive, breaking traditional social patterns. Social networks based on internet technology are more complex than traditional social networks and offer more research opportunities. In-depth research not only promotes related social network studies but also provides guidance for the further development of social media platforms. According to data from the 41st "Statistical Report on Internet Development in China" released by the China Internet Network Information Center (CNNIC), as of December 2017, China's internet user population reached 772 million, with an internet penetration rate of 55.8%; mobile internet users reached 753 million, accounting for 97.5% of the total, with mobile internet permeating every aspect of people's lives. As a social

media platform, Weibo's user adoption rate continued to grow in 2017, reaching 40.9%. Sina Weibo's Q3 2017 financial report shows that as of September 2017, monthly active users totaled 376 million, a 27% year-over-year increase, with mobile users accounting for 92%; daily active users reached 165 million, a 25% year-over-year increase. Clearly, Weibo dominates the social media landscape with significant influence.

On Weibo, users can pre-assign tags to themselves, fill in personal information such as education and birth date, and provide background information. They can also create original posts, repost content, and generate behavioral traces through activities like following other users, reposting, commenting, and liking. This information and information behavior reflect users' interest tendencies. Mastering these interest tendencies enables the creation of corresponding information recommendation models to improve user benefits from Weibo and allows the platform to conduct more targeted information push and marketing. Mining user interests and clustering users with similar interests can not only reduce the complexity of social network research but also better guide the development of personalized information recommendation services.

2 Literature Review and Research Approach

2.1 Literature Review

Social media originated abroad, with platforms like Twitter and Instagram emerging before Sina Weibo, making foreign research on social networks earlier than domestic research. In recent years, as social media user numbers have surged, domestic scholars have conducted increasingly numerous and in-depth studies on social media information dissemination mechanisms, user classification, and community detection. User classification has become a research hotspot in both computer science and library and information science. Computer science focuses on clustering algorithm research and improvement, while library and information science emphasizes personalized information recommendation through user clustering to improve information utilization efficiency. The library and information field has conducted considerable research on user clustering from single dimensions such as content, user behavior, and user background information, but few studies have approached this from a multi-dimensional perspective.

Some studies have examined social media user classification from a behavioral perspective, primarily investigating which behavioral features most effectively contribute to user clustering and how to optimize clustering algorithms. For example, M.C. Alarcón-del-Amo et al. [1] classified social networking site users into four categories—"introvert," "expert communicator," "versatile," and "novel"—based on usage frequency, experience, and interaction patterns, and summarized the behavioral characteristics of these four user types. Zhang Lin et al. [2] conducted clustering analysis of Weibo users based on four characteristic variables reflecting user information behavior: follower count, following count, number

of posts, and favorite count, and analyzed the features and influence of each category.

Some scholars have approached user classification from the content dimension, establishing information associations between users and content. For instance, J. Hannon et al. [3] calculated content-based correlations between users, analyzed and mined interest similarities, clustered users with high similarity, and provided personalized information recommendations for different categories. D.M. Blei et al. [4] studied the LDA model, extracting topics from user-published content to model document classes and form thematic documents. Each document's topics were given probabilistically and estimated via likelihood to obtain similarity between user documents, achieving association between documents and users, and then conducted topic clustering or text clustering based on topic distribution to cluster users. L.J. Hong et al. [5] constructed a new content-based user clustering model using a method that merged two topics with the LDA model. M. Efron [6] designed a method to analyze microblog content from multiple perspectives, modeling user information. Content-based user clustering focuses on textual similarity in what users express or are interested in, using this textual similarity as user similarity, with the main concern being how to model more accurate mapping between content and users.

Beyond these two dimensions, some scholars have analyzed the impact of user background information on user interests and behavioral performance. For example, Xu Zhiming et al. [7] considered user background information in user similarity measurement, with empirical results showing that user background information has significant influence on user similarity measurement. However, research on user clustering based on user behavior or information needs often overlooks the impact of background information on user clustering analysis.

2.2 Research Approach Design

In recent years, most research on community user clustering analysis has been based on single dimensions of user-published content or user behavior [8]. However, literature review combined with real-world analysis reveals that factors influencing user interests are often multifaceted, related to user background information, user blog content, and user information behavior. The importance of these three dimensions in expressing user interests should be considered simultaneously in the user clustering process [9]. The research approach of this paper is shown in Figure 1 [Figure 1: see original paper], attempting to consider user background information, user blog information, and user behavior information from three dimensions, quantify second-order variables under these multi-dimensional attributes through data collection and processing, obtain single-dimensional optimal user clustering, multi-dimensional optimal user clustering, and weighted optimal user clustering, and conduct comparative analysis of clustering effects. The specific clustering approach is: first, subdivide the attributes of these three dimensions to obtain second-order variables affecting them, thereby constructing a multi-dimensional user attribute description

model; second, obtain user data for relevant dimensions according to this model, process the data to obtain user vectors, measure user similarity using vector similarity, set thresholds, and classify users with interest-based similarity greater than the set value into one category, where this similarity is related to user background information, user behavior, and user blog content; third, since the three dimensions may not contribute equally to user interest expression, conduct weighted analysis according to their impact strength on the user description model to explore which weighting conditions yield the best clustering results; finally, use variance analysis to determine which factors among the second-order variables have minimal impact on user description, appropriately remove them, and revise the user description model to achieve better clustering results. How to fully utilize user background information, user blog information, and user information behavior for scientific analysis, how to quantify this information, how to allocate weights to achieve optimal clustering results, and how to achieve precise user positioning for interest mining, information recommendation, and precise operation are the key considerations of this paper.

By referring to relevant literature [10-11], this paper proposes a user description model based on three dimensions: user background information, blog content information, and user behavior information. The multi-dimensional user attribute description model is shown in Figure 2 [Figure 2: see original paper].

3 Data Acquisition and Processing

3.1 Microblog Data Crawling

Clustering analysis of Weibo users should first determine which characteristic variables can effectively reflect differences between users as the basis for classification, and how to quantify these features [12]. Weibo users' original characteristic attributes include follower count, following count, mutual follower count, personal description, favorite count, verification status, gender, registration time, repost count, topic count, URL count, date of first post, platform count, etc. Many related studies lack goal orientation in data acquisition, retaining data of little significance for subsequent research, resulting in significant data redundancy. This paper, based on the multi-dimensional user attribute model presented earlier (see Figure 2), subdivides the characteristic variables of underlying attributes. Drawing on previous research, the data requirements for second-order variables were obtained, as shown in Table 1 :

Table 1 Microblog User Second-Order Attribute Data Requirements

Dimension	Second-Order Variable	Data Processing Method
User Background Information	U1 User Gender (Boolean: 1 for male, 0 for female)	Boolean encoding

Dimension	Second-Order Variable	Data Processing Method
	U2 Weibo Verification (Boolean: 1 for verified, 0 for unverified)	Boolean encoding
	U3 Region (Boolean: 1 for developed city, 0 for underdeveloped city)	Boolean encoding
	U4 Education Information (Boolean: 1 has education info, 0 no education info)	Boolean encoding
	U5 Occupation Information (Boolean: 1 has occupation info, 0 no occupation info)	Boolean encoding
Blog Content Information	I1 Tags (Keyword extraction)	Keyword extraction
	I2 Introduction (Keyword extraction)	Keyword extraction
	I3 All Blog Content (Word segmentation and frequency statistics)	Word frequency analysis
	I4 Liked Blog Content (Word segmentation and frequency statistics)	Word frequency analysis
User Behavior Information	A1 Number of Posts (Boolean: 1 above average, 0 below average)	Boolean encoding
	A2 Following Count (Boolean: 1 above average, 0 below average)	Boolean encoding
	A3 Follower Count (Boolean: 1 above average, 0 below average)	Boolean encoding
	A4 Repost Count (Boolean: 1 above average, 0 below average)	Boolean encoding
	A5 Like Count (Boolean: 1 above average, 0 below average)	Boolean encoding

Note: The content in parentheses describes data processing methods, detailed in the data processing section below.

The relevant data used in this study were collected using the Octoparse data collection software. The Weibo homepage categorizes user interests into 20 categories: fashion, travel, humor, emotions, science, anime, food, sports, movies,

TV series, horoscope, music, fitness, military, digital, history, photography, cute pets, games, and beauty. To form experimental control groups, these 20 categories were retained. Based on these 20 categories, the "Find People" portal was used to input corresponding categories, and 120 users' relevant data were crawled for each category according to Table 1 (selected in ranking order). After screening, the final user sample was obtained.

3.2 Data Processing

3.2.1 Data Cleaning The collected data of 120 Weibo users in each of the 20 categories underwent data cleaning, primarily to avoid adverse effects from data sparsity and cold-start problems on subsequent data analysis and user clustering effectiveness. During data cleaning, it was necessary to delete user samples with NULL values in data items such as introduction, profile, and tags, as well as corresponding items with no practical meaning in introduction or profile content (e.g., "Work contact: XXXXXX," "Heart like still water"). Additionally, considering that institution-certified Weibo users represent organizations rather than individuals in their various activities on the platform, with strong specificity but weak research viability, institution-certified Weibo users were removed. Moreover, since the subsequent step involved crawling each user's most recent 20 posts for text analysis, users with fewer than 20 posts were also removed. From the remaining samples, a sample size of 20$\times$50 was randomly selected as the final user sample.

3.2.2 Text Information Quantification Based on Cloud Model User classification based on Weibo content information typically involves constructing a user-text correspondence model, such as LDA or three-layer Bayesian models, calculating structured text similarity to measure user similarity for clustering. However, considering the incompatibility between text description models and multi-dimensional user attribute description models, this study adopts the cloud model to quantitatively represent qualitative user blog information [13].

User nodes are treated as cloud droplets in the cloud model, and the original 20 categories are treated as 20 rating items for user scoring. User Weibo content information is processed using the bag-of-words model, ignoring text grammar and word order, treating it merely as a collection of words where each word's occurrence is independent. The IK-Analyzer segmentation tool is used for word segmentation and frequency statistics on microblog text. Projects are scored based on word frequency to obtain a user rating table (sample shown in Figure 3 [Figure 3: see original paper]). The username column contains the target user's Weibo account name. In the user rating table, the set of rated items for the target user, S1, is counted. S1 is the set of all items, and the user's unrated items S2 = S - S1, where S is the set of all items. Based on the user rating table, a user-item matrix can be obtained, from which similarity between user i and user j, Sim(i,j), is calculated. This is the data processing method for the single dimension of user blog content information. When considering the

simultaneous effect of three dimensions—user background information, user blog information, and user behavior—this user-item matrix has strong portability. For user clustering, it is only necessary to add the second-order variables of user behavior information and user background information as new items to the project rating table, simplifying the operation and directly constructing the user-item matrix to calculate user similarity through the same operation for subsequent clustering.

3.2.3 Data Standardization The benefit of data standardization is that it can improve precision and is particularly effective for user similarity measurement algorithms based on distance calculation. Standardization allows each feature variable to contribute equally to the results. In the multi-dimensional description model, due to the different nature of each dimension, they typically have different units and orders of magnitude. When the levels between dimensions differ greatly, direct analysis using raw indicators will highlight the role of indicators with higher values in comprehensive analysis while relatively weakening the role of indicators with lower values. Therefore, to ensure result reliability, raw indicator data must be standardized. This study first considers the binarization of data based on user behavior and user background information, with specific processing methods given in Table 1 .

4 Data Analysis and Discussion

4.1 Single-Dimension User Clustering Visualization Analysis

This clustering used Tableau 10.5 software to analyze user behavior data, user profile data, and blog content data separately, then conducted non-weighted analysis on these three dimensions simultaneously to obtain different clustering results. A longitudinal comparative analysis of clustering effects under these four conditions was performed, and a horizontal comparative analysis was conducted against the original 20 categories. For convenience in horizontal comparison, the original category (keyword) was selected as the horizontal axis and the clustered group as the vertical axis during clustering visualization [14].

Tableau 10.5 uses the k-means algorithm for clustering. For a given number of clusters k, the algorithm divides data into k classes. Each class has a center (centroid), which is the average of all points in that class. Centers are found through a K-means iterative process that minimizes the distance between points in a class and the class center. Tableau uses the Lloyd algorithm combined with squared Euclidean distance to calculate k-means clustering for each k. Combined with a splitting process to determine initial centers for each k > 1, the generated clusters are deterministic, with results depending only on the clustering. The Calinski-Harabasz criterion was used to evaluate clustering quality and determine the optimal number of clusters. The Calinski-Harabasz criterion is defined as in Equation (1):

$$\frac{(N-k)}{(k-1)} \times \frac{SSB}{SSW}$$

where SSB is the total between-class variance, SSW is the total within-class variance, k is the number of clusters, and N is the number of observations. A larger ratio value indicates higher class cohesion (small within-cluster variance) and greater class dispersion/separation (large between-cluster variance). When determining the optimal number of clusters, the number corresponding to the first local maximum Calinski-Harabasz index is selected.

First, clustering analysis was performed on user behavior data. When the number of classes was 5, optimal clustering results were achieved, with visualization results shown in Figure 4 [Figure 4: see original paper].

In single-dimension user behavior clustering, users with the keyword "emotions" were separately classified into one category, with inter-class attributes showing high following count, high post count, high like count, low follower count, and low repost count. Users with keywords "TV series," "movies," "anime," "horoscope," "games," and "music" were grouped into one category, with inter-class attributes showing low following count, low post count, medium like count, medium follower count, and medium repost count. Users with keywords "anime," "travel," and "fashion" were grouped into one category, with inter-class attributes showing medium following count, relatively low post count, low like count, high follower count, and medium repost count. Users with keywords "fitness," "military," "science," "history," "cute pets," and "sports" were grouped into one category, with inter-class attributes showing medium following count, medium post count, high like count, medium follower count, and high repost count. Users with keywords "beauty," "food," "photography," and "digital" were grouped into one category, with inter-class attributes showing relatively high following count, medium post count, relatively low like count, medium follower count, and medium repost count.

Under this dimension, the original 20 user categories were highly aggregated into 5 categories with high inter-class distinction, but the impact of dimensional attributes on total user attributes was relatively small. Clustering users from only this single dimension is clearly highly inaccurate.

The single-dimension clustering results based on user background information are shown in Figure 5 [Figure 5: see original paper], with optimal clustering achieved when the number of classes was 4. Users with keywords "military," "food," "emotions," and "music" were grouped into one category, with inter-class attributes showing unclear regional distribution, high proportion of verified users, balanced gender distribution, and high proportion of users providing education and occupation information. Users with keywords "TV series," "movies," "anime," "history," "travel," and "horoscope" were grouped into one category, with inter-class attributes showing balanced regional distribution, high proportion of verified users, balanced gender distribution, medium proportion of

users providing education information but low proportion providing occupation information. Users with keywords “funny,” “fitness,” “science,” “cute pets,” “photography,” “digital,” and “games” were grouped into one category, with inter-class attributes showing regional distribution 偏向 non-developed cities, medium proportion of verified users, high male proportion, and low proportions of users providing education and occupation information. Users with keywords “beauty,” “fashion,” and “sports” were grouped into one category, with inter-class attributes showing high proportion of developed city users, low proportion of verified users, high female proportion, low proportion of users providing education information, and relatively low proportion providing occupation information.

The clustering results clearly show that user classification based on single-dimension user background information has some reference value, with higher reliability than clustering based on single-dimension user behavior. For example, females are more interested in beauty and fashion, males are more interested in fitness, science, photography, digital, and games, and users from developed cities have stronger privacy protection awareness and are less willing to provide education and occupation information—all reflected in the clustering results under this dimension. Inter-class distance is relatively large, but the cohesion of user clustering based on single-dimension user background information is insufficient. Using only this dimension for user classification lacks guidance for Weibo content aggregation.

The single-dimension clustering results based on user blog information are shown in Figure 6 [Figure 6: see original paper], with optimal clustering achieved when the number of classes was 2. Users with keywords “funny,” “emotions,” and “horoscope” were grouped into one category, with inter-class attributes showing concentrated word frequency frequently appearing in the four sub-attributes I1 (tags), I2 (introduction), I3 (all blog content), and I4 (liked blog content), while also showing high co-occurrence frequency with the three keywords “funny,” “emotions,” and “horoscope.” The remaining users, excluding those with keywords “funny,” “emotions,” and “horoscope,” were grouped into another category, with inter-class attributes showing dispersed word frequency and low co-occurrence frequency with the four sub-attributes, as well as low co-occurrence frequency with the three keywords “funny,” “emotions,” and “horoscope.”

Under the single-dimension user blog information clustering, the optimal solution under non-weighted conditions clearly suffers from significant classification errors due to double counting. The goal of clustering is to group users with high blog content similarity, which is reflected in the tendencies shown by users longitudinally on first-order variables. However, without weighted analysis, when users have values across multiple second-order variables, the operations performed on each item in the user rating table are simply additive. For user vectors, this operation method distorts vector similarity measurement. In practice, this manifests as: if a user’s tags, introduction, posts, and liked posts repeatedly contain words related to the keyword “funny” (e.g., “joke,” “spoof,”

"prank") but rarely contain words related to other keywords, the current rating table operation naturally assigns high weight to this tendency. When a user shows tendencies spanning two or more keywords across tags, introduction, posts, and liked posts, the automatically assigned weight will definitely be lower than in the former case. However, the former only indicates that the user's blog content tends toward the keyword "funny," meaning such users may only be interested in "funny" information, while users in the latter case may be interested in both "funny" information and other types. This does not mean the latter users' interest level in "funny" information is lower than the former.

The data sources used for user clustering based on user background information and user behavior information are highly standardized binary matrices. Therefore, user blog information single-dimension data also requires standardization to ensure the accuracy of subsequent comparative clustering analysis and the usability of weighted analysis data.

The standardization of user blog information used the Z-score normalization method built into SPSS software. Z-score standardizes original values X to X' based on the mean and standard deviation of the raw data. After data standardization, the optimal clustering results for users based on standardized blog information are shown in Figure 7 [Figure 7: see original paper].

For user clustering based on standardized blog information, optimal results were achieved when the number of classes was 6. Figure 7 shows that within a certain error range, user interest tendencies expressed in blog information often manifest in multiple aspects. Users with keywords "funny," "beauty," "emotions," and "horoscope" were grouped into one category, with inter-class attributes showing high similarity between introduction and liked content but low similarity between tags and posted content. Users with keywords "anime," "science," "cute pets," "digital," "music," and "games" were grouped into one category, with inter-class attributes showing high similarity in user tags and liked content but low similarity in introduction and posted content. Users with keywords "TV series" and "movies" were grouped into one category, with inter-class attributes showing high similarity in introduction, tags, posted content, and liked content. Users with keywords "travel," "food," "photography," and "fashion" were grouped into one category, with inter-class attributes showing high similarity in user tags, posted content, and liked content but low similarity in introduction. Users with keywords "military" and "history" were grouped into one category, with inter-class attributes showing low introduction similarity, medium posted content similarity, and high similarity in tags and liked content. Users with keywords "fitness" and "sports" were grouped into one category, with inter-class attributes showing high similarity in posted content and liked content, medium similarity in tags, and low similarity in introduction.

Standardized data significantly improved the effectiveness of the single-dimension user clustering model based on user blog information. The clustering results have high reference value. For example, users with keywords "movies" and "TV series" were grouped into one category with high inter-class attribute

similarity, reflecting that users interested in TV series often also show interest in movie information.

4.2 Multi-Dimensional Attribute Weighted User Clustering Visualization Analysis

4.2.1 Multi-Dimensional Non-Weighted User Clustering Visualization Analysis After data standardization, a total user rating table based on the cloud model was obtained. This table was imported into Tableau software for user clustering analysis, yielding optimal user clustering results shown in Figure 8 [Figure 8: see original paper].

When the number of classes was 4, optimal user clustering was achieved. The clustering summary diagnosis yielded a total between-group sum of squares of 9.1821 and a total within-group sum of squares of 8.2197. The between-group sum of squares metric quantifies the distance between classes as the sum of squared distances between each class center and the dataset center, with class centers measured using weighted averages based on data points assigned to the class. A larger between-group sum of squares indicates better separation between classes. The within-group sum of squares metric quantifies class cohesion as the sum of squared distances between each class center and individual marks within the class. A smaller within-group sum of squares indicates higher class cohesion. Analysis shows that under multi-dimensional conditions, the clustering effect is still not optimal in the optimal case, with small between-group distances and large within-group distances, and unclear category boundaries. Although multi-dimensional user clustering more comprehensively considers the impact of each dimension on user similarity measurement, it disperses the intensity of user similarity performance across dimensions, resulting in suboptimal clustering effects.

4.2.2 Multi-Dimensional Weighted User Clustering Visualization Analysis Considering that the three dimensions may contribute unequally to the user clustering process, different weights were assigned to each dimension in the model. Since there are three dimensions and prior research on weight allocation is lacking, weight allocation experiments were conducted simultaneously with weighted analysis. This paper used the common linear weighting method to conduct weighting experiments on the three dimensions, with weight allocations as follows:

Weighting 1: User Vector = (User * 0.25, Content * 0.5, Action * 0.75)
Weighting 2: User Vector = (User * 0.25, Content * 0.75, Action * 0.5)
Weighting 3: User Vector = (User * 0.5, Content * 0.25, Action * 0.75)
Weighting 4: User Vector = (User * 0.5, Content * 0.75, Action * 0.25)
Weighting 5: User Vector = (User * 0.75, Content * 0.5, Action * 0.25)
Weighting 6: User Vector = (User * 0.75, Content * 0.25, Action * 0.5)

The diagnostic data for optimal user clustering under different weightings are

shown in Table 2 :

Table 2 Optimal User Clustering Effect Diagnosis Under Multi-Dimensional Weighting

Diagnostic Metric	Weighting 1	Weighting 2	Weighting 3	Weighting 4	Weighting 5	Weighting 6
Between-Group Sum of Squares	5.6279	10.799	9.8483	1.6438	8.156	6.4661
Within-Group Sum of Squares	15.06	5.6028	7.5535	18.5	10.132	10.936

Comparing the clustering effects of optimal user clustering under the six weighting methods in Table 2, Weighting 4 yielded the best user clustering results. That is, when the weight of user blog content is 1/2, user background information weight is 1/3, and user behavior information weight is 1/4, with the number of clusters being 14, optimal clustering effects were achieved. The clustering visualization is shown in Figure 9 [Figure 9: see original paper].

Compared with user clustering effects based on single dimensions of user blog content, user background information, and user behavior information, the weighted user clustering has more explicit inter-class characteristic attributes. For example, in optimal user clustering based on single-dimension user blog content, users with keywords “funny,” “emotions,” “beauty,” and “horoscope” were grouped into one category. However, under multi-dimensional weighted conditions, users with keywords “funny” and “emotions” were grouped together, while users with keywords “beauty” and “horoscope” were separately divided into two individual categories. This demonstrates that similarity differences based on user background information and user behavior information dimensions affect clustering results. Similarly, differences in clustering results compared to single-dimension user clustering are caused by influences from the other two dimensions. This influence increases user differences, making clustering results more accurately reflect user interest tendencies.

Compared vertically with multi-dimensional non-weighted user clustering effects, weighted user clustering better reflects actual conditions. For example, under non-weighted conditions, users with keywords “TV series,” “movies,” “anime,” “military,” “history,” “emotions,” and “anime” were grouped into one category. Under weighted conditions, classification changed significantly. When treating multiple dimensions equally, classification alienation may occur—users with high blog similarity may be dispersed into multiple categories due to low similarity in user behavior and user background information dimensions. In reality, user similarity based on blog content, user background information, and user behavior contributes differently to overall user similarity measurement. Only

by recognizing this difference can we obtain a more realistic user similarity measurement model and clustering results that better match actual conditions. Weighted analysis confirms that optimal clustering effects are achieved when the single dimension of user blog content has the greatest weight, indicating that the user blog content dimension has the strongest influence on user similarity measurement.

Meanwhile, compared with Weibo's original classification standards, some users originally belonging to different categories were merged into one category. For example, users with keywords "funny" and "emotions" were grouped together, indicating that in user classification, we must 重视 the concept of overlapping communities, recognize intersections between user interests, and consider the importance of non-linear user classification. In subsequent information recommendation, we should comprehensively consider the overlap of user classifications and appropriately push information from other categories to users with overlapping interests.

4.3 Model Optimization

4.3.1 Variance Analysis Analysis of Variance (ANOVA) is a collection of statistical models and related procedures used to analyze differences between and within classified observations. Variance is calculated for each variable to generate an ANOVA table that can determine which variables are most effective for clustering.

Tableau's relevant ANOVA statistics include F-statistics, P-values, model mean square, and error sum of squares. The F-statistic for one-way or single-factor ANOVA is the variance fraction explained by the variable, representing the ratio of between-group variance to total variance. A larger F-statistic indicates greater distinction between classes for the corresponding variable. The P-value refers to the probability that values in the F-distribution of all possible F-statistics exceed the actual F-statistic for the variable. If the P-value is below the specified significance level, the null hypothesis (that individual elements of the variable are random samples from a single population) can be rejected. This F-distribution has degrees of freedom (K-1, N-K), where K is the number of classes and N is the number of items in the established classes. A lower P-value indicates greater distinction between expected values of the variable's elements across classes.

Model mean square is the ratio of between-group sum of squares to model degrees of freedom. Between-group sum of squares measures differences between cluster means. If cluster means are very close to each other (and thus close to the overall mean), the value will be small. Model degrees of freedom is k-1, where k is the number of clusters. Error sum of squares is the ratio of within-group average sum to error degrees of freedom. Within-group sum of squares measures differences between observations within each cluster. Error degrees of freedom is N-k, where N is the total number of observations (rows) in estab-

lished clusters and k is the number of clusters. Error sum of squares can be regarded as overall mean square error, assuming each cluster center represents the cluster's "true value."

Under optimal weighted conditions, the model's ANOVA results are shown in Table 3 :

Table 3 ANOVA for Optimal Weighted Clustering

Variable	F-Statistic	P-Value	Model Mean Square	Error Sum of Squares
All Blog Content	12.922	0.001323	1.295	0.1003
Liked Blog Content	9.164	0.002703	0.9266	0.1011
Occupation Information	7.115	0.004951	0.7185	0.1010
Weibo Verification	1.018	0.005113	0.1028	0.1010
Repost Count	0.576	0.01661	0.0582	0.1010
Number of Posts	2.575	0.04358	0.2596	0.1009
Follower Count	0.822	0.08282	0.0829	0.1009
Following Count	5.999	0.09724	0.6053	0.1009
Introduction	7.931	0.1025	0.8002	0.1009
Tags	0.2883	0.2552	0.0291	0.1010
Region	0.8246	0.2636	0.0832	0.1010
Gender	0.9467	0.4759	0.0955	0.1009
Education Information	0.241	0.6662	0.0243	0.1010
Like Count	0.1873	0.7885	0.0189	0.1010

Table 3 shows that among the 14 second-order variables, the variable contributing most to the model is blog content, with 8 variables having P-values less than 0.1: blog content, tags, and liked blog content from the user blog information dimension; gender, verification, region, number of posts, and following count from user background and behavior dimensions. This means these 8 second-order variables are generalizable for user description in the original model. The practical significance lies in identifying the influence strength of second-order variables on user vector expression—larger F-values and smaller P-values indicate more reliable variables for user clustering, meaning the variable better distinguishes different user classes. For example, the blog content variable has an F-value of 12.922 and a P-value of 0.001323, indicating 99.8677% confidence that using this variable to distinguish users is correct. In statistics, 0.01 is generally used as the P-value threshold; P-values below 0.01 indicate the variable is fully effective for the model. For second-order variables in Table 4 with P-values greater than 0.01, removal can be considered to improve model effectiveness.

4.3.2 Improvement of Multi-Dimensional User Attribute Description Model Under optimal weighted conditions, ANOVA was performed on optimal user clustering. Considering that the P-value for the second-order variable "introduction" is very close to 0.01, using P=0.011 as the threshold, variables

with P-values less than 0.11 were retained, while those with P-values greater than 0.11 were removed. Considering optimal weighting, the improved multi-dimensional weighted user attribute description model is shown in Figure 10 [Figure 10: see original paper].

Through clustering analysis, weighted analysis, and variance analysis of user sample data, this study examined user clustering effects based on multi-dimensional attributes and weighting, obtaining a multi-dimensional weighted user attribute description model. This model is significant for guiding Weibo user classification and subsequent information recommendation research, while also providing new ideas for community user clustering. This paper also has some limitations, such as: Do clustering effects differ under various clustering methods based on the weighted approach? How to construct an information recommendation mechanism based on this model? These require further research.

References

[1] Alarcón-del-Amo MC, Lorenzo-Romero C, Gómez-Borja M. Classifying and profiling social networking site users: A latent segmentation approach[J]. Cyberpsychology, behavior, and social networking, 2011, 14(9): 547-553.
[2] Zhang Lin, Xie Zhonghong. Research on Microblog User Types and Influence Based on Clustering[J]. Information Science, 2016, 34(8): 57-61.
[3] Hannon J, Bennett M, Smyth B. Recommending Twitter users to follow using content and collaborative filtering approaches[C]//Proceedings of the 4th ACM conference on recommender systems. New York: ACM, 2010: 199-206.
[4] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.
[5] Hong LJ, Davison BD. Empirical study of topic modeling in Twitter[C]//Proceedings of the first workshop on social media analytics. New York: ACM Press, 2010: 80-88.
[6] Efron M. Information search and retrieval in microblogs[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6): 996-1008.
[7] Xu Zhiming, Li Dong, Liu Ting, et al. Similarity measurement of microblog users and its application[J]. Chinese Journal of Computers, 2014, 37(1): 207-218.
[8] Huang Jing. Research on user behavior characteristics and applications in consumer virtual communities[J]. Library and Information Service, 2011, 55(3): 97-100, 51.
[9] Cui Jindong, Sun Yaoyao, Wang Xin, et al. Research on microblog information recommendation method based on Folksonomy and ontology fusion[J]. Information Science, 2015, 33(10): 27-31.
[10] Xue Yunxia. Research on microblog user attribute identification methods[D]. Suzhou: Soochow University, 2015.
[11] Gu Xiaoxue, Zhang Chengzhi. Research on tag clustering combining

annotation content and user attributes[J]. New Technology of Library and Information Service, 2015(10): 30-39.
[12] Peng Xixian, Zhu Qinghua, Liu Xuan. Research on characteristics analysis and classification of microblog users—Taking "Sina Weibo" as an example[J]. Information Science, 2015, 33(1): 69-75.
[13] Zhang Guoying, Sha Yun, Liu Xuhong, et al. High-dimensional cloud model and its application in multi-attribute evaluation[J]. Journal of Beijing Institute of Technology, 2004(12): 1065-1069.
[14] Li Xiaohui. Research on personalized recommendation algorithm based on Jaccard item category similarity[D]. Changsha: Central South University, 2010.

Author Contributions

Zhang Haitao: Responsible for article structure framework design and guidance;
Tang Shiman: Responsible for paper writing and revision;
Wei Mingzhu: Responsible for data crawling related work;
Li Zezhong: Responsible for variance analysis.

English Abstract

Research on the Clustering of Microblog Users Based on Multi-Dimensional Attribute Weighting Analysis

Zhang Haitao[1],[2], Tang Shiman[1], Wei Mingzhu[1], Li Zezhong[1]
[1] The Management College of Jilin University, Changchun 130022
[2] The Information Resource Research Center of Jilin University, Changchun 130022

Abstract: [Purpose/significance] It is of great significance for the study of social network information ecology and information recommendation to accurately grasp the interest tendency of social network users and classify users into highly aggregated user groups. [Method/process] In this paper, by constructing the user attributes describe hierarchical model based on multi-dimensional, according to the model data requirements fetching user sample data from Sina microblog, quantify the second-order variable based on the multi-dimensional property of the users' background information, users' blog information and user behavior information to construct user vector expression, comparing the classification results based on single dimension and the multi-dimensional, given different weights to attribute for weighted analysis, when achieve the optimal clustering results, based it do variance analysis to improve the model. [Result/conclusion] User clustering effect based on the multi-dimensional attribute weighting is significantly better than the user clustering effect based on the single-dimensional and under the condition of the multi-dimensional unweighted, and users microblog content dimension for improving the validity of user clustering effect is the largest.

Keywords: microblogs, multi-dimensional, user-cluster, weighted-analysis

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