International Journal of u- and e- Service, Science and Technology
Vol.8, No.1 (2015), pp.143-154
http://dx.doi.org/10.14257/ijunesst.2015.8.1.13

Research and Implementation of User Features Mining Model without Labels in Social Networks

Zhang Xu

Jiangsu Institute of Economic,Nanjing 211168,China
494848647@qq.com

Abstract

Under the tide of information technology revolution, social network services (SNS) become a typical application in Web2.0 era by its rich and interactive user participation, and has swept the world in a short time. More and more users begin to express themselves on Facebook, Micro-blog and other social networks, and user features are existing in the SNS in a more intuitive way. These information assign complete personality and image for each node in the SNS, which has enormous potential commercial value. For background, this paper studies the user features mining problem in social networks, and focuses on the user features without labels. Firstly, two models are established for user features mining without labels, in which clustering, classification, text mining and graph mining are used in the model. Then, the proposed models are implemented under two scenarios: user interest discovery and Micro-group structure discovery. Experiment results based on Sina Micro-blog show that the accuracy is above 80%, which can meet the demand of user features mining in social networks. Therefore, the techniques proposed in this paper are feasible.

Keywords: *social networks, data mining, clustering, classification, user features*

1. Background of Social Networks

Under the tide of information technology revolution, the Internet domain is transferring from Web1.0 era in which users mainly aim to obtain information to Web2.0 in which users are both acquisitions and sources of the Internet. Social networks have become the typical application in Web2.0 era by its rich and interactive user participation, and swept the world in a short time[1], such as foreign Facebook, Twitter, LinkedIn and domestic Renren, Micro-blog and other various vertical social networks. The appearance of these sites breaks the traditional barriers for friend interaction and the limitations of time and space, which greatly expends the user's circle of friends and their interactions. In turn, the appearance of SNS promotes the social morphology on the Internet and the boost of user behaviors to the real society, making virtual society and real society begin to cross. Therefore, it is SNS that starts the third revolution of the Internet domain.

To study SNS, we should firstly understand the basic concept of SNS. Broadly speaking, SNS is an Internet service aiming to help users build social networks. The definition of SNS in Wikipedia is as follows: SNS is a kind of social relationship networks which is built based on common hobbies or interests on the Internet[2]. In fact, there are many definitions of SNS, and they are constantly changing with the development of the networks. However, as the SNS is defined by connections and interactions between users, thus, it can be described by a series of nodes and edges from the perspective of graph theory. As shown in Figure 1, nodes represent the users or other entities in the network, while edges represent the connections or interactions between these entities. This model applies to all types of social network, but nodes and edges may differ with the specific network type and scenario demand. In addition to the rich graphical information constituted by the connections, SNS also contains a large amount of content information, which exists in nodes or edges in various multimedia forms, such as text, images, sound, video, and so on.

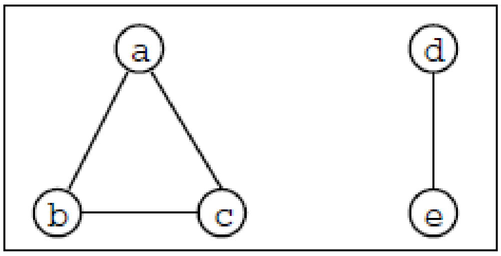

Figure 1. Graphic Representation of SNS

Under the booming waves of SNS, one and another information publishing and social networking platforms emerge in endlessly. Micro-blog is a kind of information dissemination and sharing platforms, which includes friends management, micro-blog publish, comment or forward others' content and other functions. Micro-blog has two basic functions: instant communication and social media. At present, there are many domestic micro-blogs, in which Sina Micro-blog is in the leading place due to its strong stickness and activity. Compared to Twitter, besides maintaining the communicating and social functions, Sina Micro-blog greatly strengthened the media and spread ability[3]. In early 2010, Sina Micro-blog launched its open platform service, providing massive micro-blog content, relationships and information dissemination channels for developers. It is an open platform for information subscription, sharing and exchanging based on Sina Micro-blog system. The Sina Micro-blog API is completely open, and any organization or individual can get data via the open platform interface just by simply registration for further research or application development. The data we need is acquired through the open platform API, supporting the algorithm simulation and model implementation proposed in this paper.

2. Introduction of Data Mining Technology

Massive information stored in social networks bring unprecedented opportunities and challenges for data analysis and mining, making data mining technology developed rapidly in the last two decades. Meanwhile, the results of data mining have been widely applied to many traditional subjects and produces design, greatly promoting the development of other areas.

2.1. Rise and Overview

We live in “data era”, and several terabytes or megabytes of data from business, society, science and engineering, medicine and other areas of our daily lives pours in the computer networks, the world wide web and various data storage devices. Among them, the community and social media have become increasingly important sources of data, which generate massive data, including digital images, video, blog networks, online communities and different kinds of social networks. The explosive growth, widely available and enormous amount make our times a real "data era". Therefore, we urgently need powerful and versatile tools to mining valuable information from these massive data and transfer them into organized knowledge. Knowledge discovery and data mining (KDD) is a young, vibrant, dynamic changed and fast-growing field, which has been and will continue to make contributions in the process from “data era” to “information era”[4].

2.2. Application of Data Mining in Social Networks

As mentioned in section 2.1, community and social media have become increasingly important data sources. They generate massive content data and diverse social networks, which contain a large amount of information related to user features. These data not only contains feedback data for SNS development and product design, but also shows a vital significance to society, behavioral science, marketing and other subjects. In order to filter out useful data, extract data model and draw effective conclusions, data mining technology has long been applied to the SNS field.

There are two most common methods in current SNS data mining: one is based on network structure and connections analysis [5]; the other is based on content information analysis[6]. Graph theory is the first technology applied to SNS analysis. At the end of 20th century, there appears complex network research in system science field, which is a further extension and supplement to graph theory, and shows great advantages in network structure, behavior and dynamic evolution analysis of social network [1]. At the same time, SNS content mining is also booming. As mentioned before, users contribute a lot of text and multimedia content information for SNS, which plays an important guiding role for our analysis. Therefore, there are a lot of researches on content mining techniques, and some effective methods and tools have been developed, such as ontology establishment, sentiment analysis, keyword-based search, semantic-based queries, etc. [7].

2.3. User Features Mining in Social Networks

SNS provides platforms for users to show themselves, so the users are not only just audiences of the network media, but also become active participants. These active users provide a large amount of user feature information for SNS. Different literatures have different definitions for user feature, in which literature [8] divides them into five aspects: basic information, user knowledge, user interest, browsing history and user preferences, which cover all intuitive features in SNS. In fact, user features information in SNS is far more than this, such as user's friends, mood changes and personality characteristics, however, these information cannot be obtained through direct observation, and they need more in-depth data mining techniques. The goal of this paper is to obtain user features in social networks by data mining methods, in which the combination of classification, clustering, graph mining and text mining is the main method in this process[9].

We call the feature we're going to mining "target feature", and the actual user features mining scenarios can be divided into two categories based on the results of target feature: user features mining with labels and user features mining without labels. User features mining with labels refer to that the labels of target feature have been known before mining, and our purpose is to classify user target features into different labels. Taking gender as target feature for example, the labels of target feature are "male" and "female", which is known before mining, and our purpose is to classify users into them, thus it belongs to user features mining with labels. User features mining without labels refer to that the labels of target feature is known before mining, and we need to not only determine the target feature labels, but also classify users into different labels. Taking geographical region as target feature for example (assume that users' regions are uncertain), we need to identify both the geographical labels and the label of each specific user, so it belongs to user features mining without labels. User features mining with labels is a typical classification problem, and is mainly solved by classification algorithms. User features mining without labels is a little complex, solutions are different according to different scenarios. Firstly, labels of the target feature are unknown at the beginning, making it more similar to clustering problem, thus it can be directly solved by clustering algorithms. Secondly, if target feature labels can be determined, it can be transferred to user features mining with labels and solved by classification algorithms. This paper studies user features mining without labels problem.

3. User Features Mining Model without Labels

3.1. Problem Description

Social networks are usually identified as graphs G(V,E), in order to describe the user features mining model without labels, we firstly introduce the basic variables as follows:

V: node set in social networks, and v_i represents for user i;

E: edge set in social networks, and e_{ij} represents for connection between user v_i and v_j;

W: weight set of the edges, and w_{ij} represents for the weight of e_{ij}, which shows the connection strength between user v_i and v_j;

F: target feature in specific scenario, such as user interest or preference mentioned before;

R: labels set of the target feature, and r_i represents for the label of feature F of user v_i.

Based on these variables, user features mining without labels can be described as follows: Given network G(V,E), determine the label r_i of target feature F of each user v_i, in which $r_i \in R$, but the number and types of labels in R are unknown at the beginning.

3.2. Clustering Model Establishment

Clustering is to divide data set into multiple clusters, which is suitable for the situation that the labels of all the data are unknown. Therefore, user features mining without labels is a typical clustering problem. Its purpose is to gather users with similar target features into a cluster, in order for that the users in one cluster are similar with each other, but very different from users in other clusters. Node clustering problem in social networks is developed from graph partitioning. Traditional graph partitioning measures nodes intimacy with distances defined in graph theory, however, users with similar target features are not directly reflected from nodes distances. Thus clustering model is established based on traditional clustering methods, adding a lot of SNS features, as shown in Figure 2.

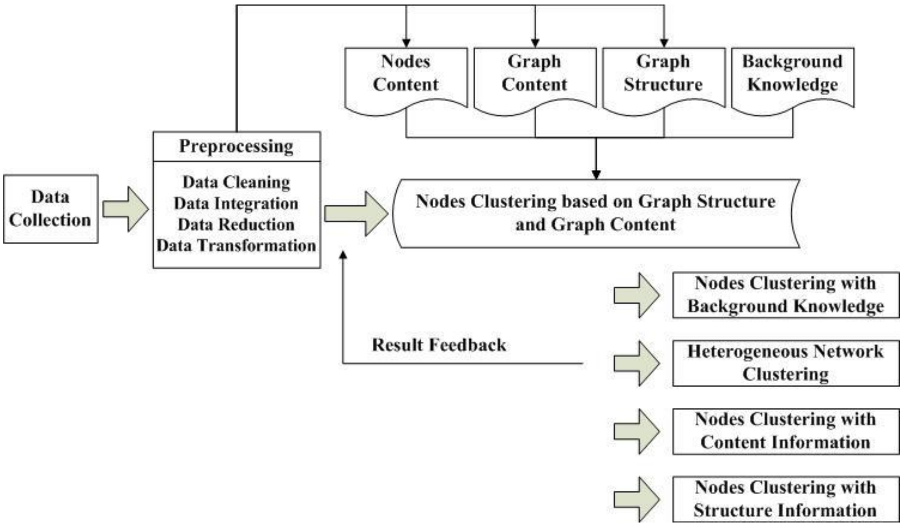
Nodes Content
Graph Content
Graph Structure
Background Knowledge
Preprocessing
Data Cleaning
Data Integration
Data Reduction
Data Transformation
Data Collection
Nodes Clustering based on Graph Structure and Graph Content
Nodes Clustering with Background Knowledge
Result Feedback
Heterogeneous Network Clustering
Nodes Clustering with Content Information
Nodes Clustering with Structure Information

Figure 2. Clustering Model of user Features Mining without Labels

User data collected from Sina Micro-blog firstly needs to be preprocessed, and then input into the nodes clustering module. In order to take full account of the user information in SNS, the clustering module combines analysis based on both content and structure. Node content, graph structure and graph content information are considered, and background knowledge may also be added as guidance to decide or adjust the clustering result. According to the needs of different scenarios, this model may be a fusion of different types of node clustering problems, such as node clustering with background knowledge, heterogeneous network clustering, node clustering with content and structure information, etc. There have been some mature studies and outcomes of these problems [10]. To further improve the algorithm accuracy, we require users to feedback on the clustering results, and the model will be trained and improved according to these feedbacks.

3.3. Classification Model Establishment

Since the difference between user features mining without labels and user features mining with labels is whether the target feature labels set is given or not, so if the labels set can be determined by background knowledge, artificial intelligence or some clustering algorithms, it can be transferred to user features mining with labels and solved by classification algorithm. The classification model is shown in Figure 3.

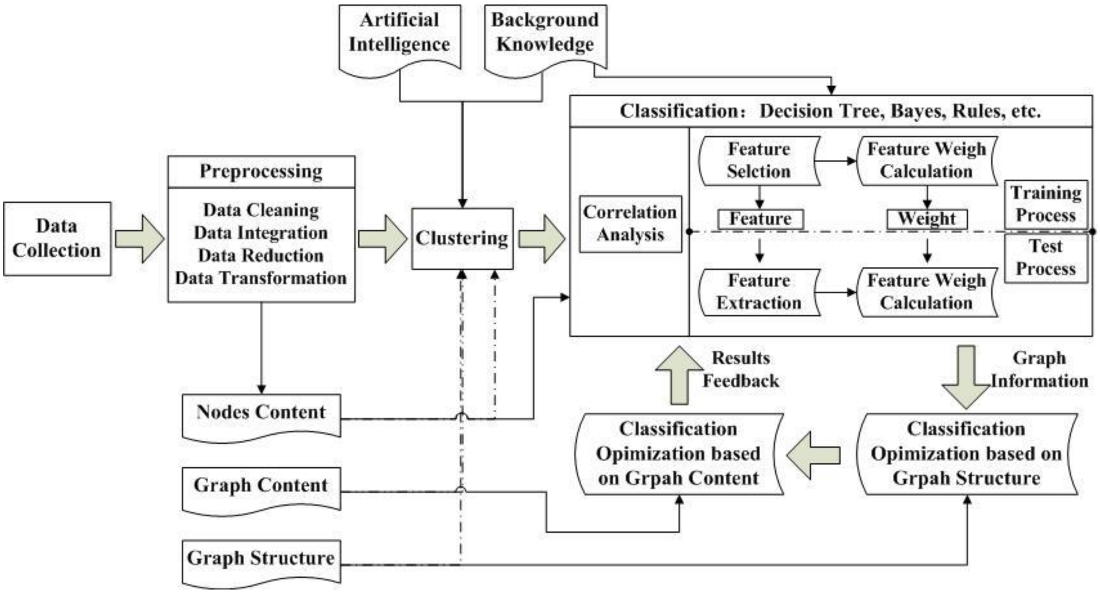

Figure 3. Classification Model of user Features Mining without Labels

In this model, target feature labels set R is determined by background knowledge, artificial intelligence or clustering algorithms after data preprocessing. Then the problem is transferred to user features mining with labels, and the subsequent modules are typical classification process. Preprocessed data is input into feature selection and weight calculation module. This module contains training and testing process, and background knowledge may also be needed. Then, the feature weights are input into classification module, which mainly considers nodes features. Taking the current classification results as initial labels, we make further iterations and update the labels based on structure information and graph content. Likewise, users are required to feedback on the results, and the model will be improved based on them.

4. Modeling User Interests Discovery

The behavior of Micro-blog users is usually closely related to their interest, hobbies or concerns, and mining user interest accurately is of a great significance for personalized recommendation services. The scope of user interest is very broad, and their labels are not known at first, thus it belongs to user features mining without labels. Further more, the community structure is an important feature of social networks. A large number of facts and research show that the community is "connected closely with internal nodes, and far away from other communities" in nature [11]. Many communities are formed based on user interest, so we discover the user interest in social networks with the clustering model.

We collected 2000 users as test data whose "tags" are valid, and their "tags" are used as the true value to measure the accuracy of our model.

4.1. Data Preprocessing and Feature Weight Selection

Through observation, we find that there show different keywords in different users' micro-blogs, which can reflects the user interest and concerns, and is very helpful for our problem. So we take these keywords as input feature and complete user interest discovery by clustering model.

We firstly extract the keywords appear frequently in micro-blogs through data preprocessing, and then convert it to a matrix: user (id, label, prob). "Id" represents for the user id, "label" represents all probable labels, and "prob" represents for whether the user owns the corresponding label. Each row of the matrix represents for a user, and each column represents for a label. If the user owns this label, the value is 1, otherwise the value is 0. User interest is discovered based on this matrix.

4.2. Connection Weight Calculation

Each user in SNS is equivalent to a node, and the connection or similarity between users is equivalent to the edge. Then, how to assign values for weighted graph becomes an important issue. This section presents the following three methods on it.

4.2.1. Cosine Distance: Cosine distance, also called as cosine similarity, measures the difference of two vectors with the cosine of the angle between them. Let the feature of X and Y is represented as X(x_1,x_2,…, x_n), Y(y_1,y_2,…,y_n), then the cosine distance between X and Y is calculated by the cosine of their angle θ, as shown in Equation 1.

$$\cos\theta = \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{\sqrt{x_1^2+x_2^2+\cdots+x_n^2}\times\sqrt{y_1^2+y_2^2+\cdots+y_n^2}} \qquad \text{(Equation 1)}$$

The closer to 1 the cosine of angle θ is, the more similar X and Y are; and the closer to -1 the cosine of angle θ is, the less similar X and Y are.

4.2.2. Euclidean Distance: Euclidean distance is a traditional edge assignment method[12], measuring the absolute distances between nodes. Let the feature of X and Y is represented as X(x_1,x_2,…, x_n), Y(y_1,y_2,…,y_n), then Euclidean distance between them is calculated as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{(Equation 2)}$$

Cosine distance distinguishes two features in direction, and is not sensitive to their absolute value, while Euclidean distance can better reflect the absolute difference of the values.

4.2.3. Jaccard Distance: Jaccard distance is proposed on the basis of Jaccard similarity coefficient. The ratio of the intersection elements to the union elements is called as Jaccard similarity coefficient, which measures the similarity of two sets, as shown in Equation 3[13].

$$J(X,Y) = \frac{|X\cap Y|}{|X\cup Y|} \qquad \text{(Equation 3)}$$

As shown in Equation 4, Jaccard distance measures the distance between two nodes by 1 minus Jaccard similarity coefficient. Jaccard distance effectively describes the proportion of different elements in two set to the all elements to measure the difference between two set [14].

$$J_\delta(X,Y) = 1 - J(X,Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} \qquad \text{(Equation 4)}$$

4.3. Clustering Algorithm

Clustering algorithms can be divided in different ways, such as based on criteria, cluster separation, similarity measure method, clustering space, etc. Partitioning method, hierarchical method and clustering based on density are common clustering methods. This section introduces two most commonly clustering methods.

4.3.1. K-means Clustering: The key of K-means algorithm is iteration and refinement. Given some observable objects (x_1,x_2,…,x_n) and each one is a multi-dimensional vector, the purpose of K-means algorithm is to divide them into k clusters, S={S_1,S_2,…,S_k}, to make the sum of squares in the cluster the smallest, as shown in Equation 5[15].

$$\operatorname{argmin} \sum_{i=}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \qquad \text{(Equation 5)}$$

in which u_i is the average node in cluster S_i. K-means algorithm contains two steps:

Step1: assignment, which aims to assign each observable node to the cluster to whose average node is the nearest. Let m_1,m_2,…,m_k is the initiate average nodes set. As shown in Equation 6, each x_p will be assigned to $S^{(t)}$, m_t is the nearest one to xp, after this step.

$$S_i^{(t)} = \left\{x_p : \left\| x_p - m_i^{(t)} \right\| \le \left\| x_p - m_j^{(t)} \right\| \, \forall\, 1 \le j \le k \right\} \qquad \text{(Equation 6)}$$

Step2: update, which aims to calculate the new average node of the updated cluster as shown in Equation 7. When the results of this step keep stable, the algorithm ends.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \qquad \text{(Equation 7)}$$

The main drawback of K-means algorithm is the significance of the initial average nodes, and different ones show a great impact on the clustering result, thus it is an unstable algorithm. In addition, K-means calculates the average of all the nodes in the cluster and take it as the new center node of the cluster. So, if there exists noise node in the system (refers to the isolated node distributing far from other nodes), the clustering result may deviated from the most nodes area, reducing the clustering accuracy.

4.2.2. K-medoids Clustering: To overcome the problem caused by noise nodes in K-means, K-medoids clustering is proposed. The biggest difference between K-means and K-medoids is the selection of the cluster center node. Instead of the average node, K-medoids selects an existing node in the cluster as its center node, effectively avoiding the influence of noise node [16].

Assuming the center node of a cluster is O(i), then we sequentially select a node as O(j) and calculate the cost of applying O(j) to replace O(i), which is measured with the distances between all the nodes to its cluster center node in the whole graph. After considering the impact of taking O(j) to replace O(i), the node whose cost is the least, namely improved the clustering best, will be selected as the new center node.

In addition, K-Means algorithm needs to know the node coordinates, and it cannot be a symbol or Boolean value, or the type of center node will change after calculation. K-medoids is not limited, and it can be conducted with only distances between nodes.

4.4. Model Implementation and Results Analysis

In order to compare the pros and cons of different methods above, 6 solutions are obtained after combining different feature weight calculation and clustering algorithm.

- Solution1: cosine distance for similarity calculation, K-means as clustering method;
- Solution2: cosine distance for similarity calculation, K-medoids as clustering method;
- Solution3: Euclidean distance for similarity calculation, K-means as clustering method;
- Solution4: Euclidean distance for similarity calculation, K-medoids as clustering method;
- Solution5: Jaccard distance for similarity calculation, K-means as clustering method;
- Solution6: Jaccard distance for similarity calculation, K-medoids as clustering method.

After repeated verification, we choose K=10 as the clustering parameter, and experience each solution 10 times. The mean and standard deviation of the results are shown in Table 1. Short distances in clusters, long distances between clusters and small ratios of them show an effective clustering result. Meanwhile, mean and standard deviation reflect the stability of the algorithms. As Table 1 shows, the clustering result of the solution6 is the most effective, and its mean and standard deviation are the smallest, thus is also the most stable method. In sum, Jaccard distance is feasible in measuring user features in social networks such as Micro-blog, and it shows an obvious advantage compared with traditional cosine distance or Euclidean distance. The clustering result is ideal by K-medoids combined with Jaccard distance.

Table 1. Experiment Result of User Interest Discovery

Solution	Ratio of the distance in clusters to the distance between clusters					Mean	Standard deviation
1	0.1578	0.2256	0.1965	0.1822	0.2098	0.24119	0.085036
	0.4368	0.3092	0.2503	0.1598	0.2839		
2	0.2802	0.5582	2.3925	0.9542	0.4129	0.92460	0.925636
	0.5274	0.4906	0.6213	2.8523	0.8564		
3	0.1245	0.1427	0.1202	0.1304	0.1189	0.13019	0.014173
	0.1264	0.1163	0.1190	0.1598	0.1437		
4	0.1197	0.1205	0.1089	0.1045	0.1280	0.11826	0.009109
	0.1070	0.1164	0.1306	0.1274	0.1196		
5	0.1152	0.1097	0.1206	0.1257	0.1232	0.11600	0.007391
	0.1184	0.1030	0.1064	0.1184	0.1194		
6	0.1084	0.1090	0.1124	0.1183	0.1046	0.11245	0.006857
	0.1219	0.1163	0.1082	0.1033	0.1221		

5. Modeling Micro-group Structure Discovery

Micro-groups, short for Micro-blog groups, are formed spontaneously due to the same hobbies or tags, which is similar with clubs to convenient users to gather and communicate. This section will study the social structure of the Micro-group, and compare it with the structure of real society offline to analyze the SNS features.

5.1. Labels Set Construction

At the beginning, we have no idea about the labels in Micro-group structure. So we refer to some background knowledge, analyze several typical Micro-groups, and divide the Micro-group structure into 3 categories, individuals, organizations and information sources, and 6 sub-categories as shown in Figure 4. Micro-blog media refers to some public accounts for information sharing or advertising. It is special information sharing way in SNS, and is also the only one not corresponding to the real society.

Now, Micro-group structure discovery has become user features mining with labels, and the labels shown in Figure 4 are the target labels for the classification process.

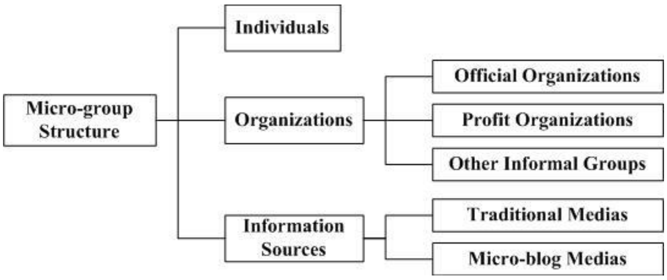

Figure 4. Labels of Micro-group Structure

5.2. Feature Selection and Extraction

After the classification labels are determined, we analysis the correlation between user features in Micro-blog with the classification result, and choose “username (UN)” and “user description (UD)” as input features of the classification algorithm. These two features are the most basic ones in Micro-blog, and can reflect user identity effectively. Most organizations and information sources take their real identity as usernames, such as “worldwide children safety organization” and “Chinese old newspaper”. Keywords analysis on these usernames can directly provides the classification result. However, the usernames of most users are casual, and their identities cannot get only by analysis on usernames. So another parameter, user description, is introduced to further guide the user identity classification.

UN in Micro-blog is limited in 4-30 characters, and UD less than 70 characters. They are both short text, which are not suitable for semantic algorithms, but their keywords features are obvious. Thus we choose keywords matching and keywords fuzzy matching based on weights as the method for Micro-group structure discovery [17].

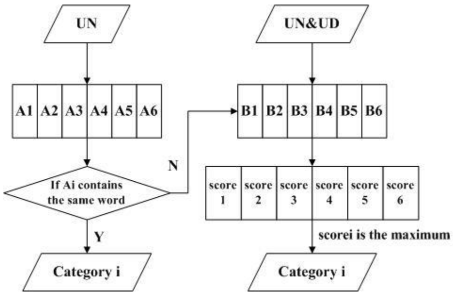

Figure 5. Flowchart of the Micro-group Discovery Algorithm

5.3. Micro-group Structure Discovery Algorithm

First of all, we construct the basic keywords vocabularies A and B to help classify users in Micro-groups into the correct category. The words in A and B are respectively divided into 6 categories, corresponding to 6 target labels. Words in A1-A6 are closely associated with the corresponding category, aiming for keywords matching with UN. If UN contains a word in Ai, then it will be directly classified into category i. Words in B1-

B6 are related to each category, and each word is assigned with a weight, represents for the correlation between the word and the corresponding category [18]. Therefore, users cannot be classified by vocabulary A will be input into vocabulary B. If UN or UD contain a word in Bi, its weight will be added to the weight of the corresponding category. At last, the user will be classified to the category whose weight is the maximum. Figure 5 shows the flowchart of this algorithm.

Pseudo code of this algorithm is shown as follows:

```
Read user-X's UN as input;
for i from 1 to 6;
if (UN contains a word in Ai)
   Classify user-X into category i;
else{
      Read user-X's UN&&UD;
   for i from 1 to 6, int score[i]=0;
   if(UN&&UD contain words in Bi)
       score[i]=score[i]+weight (in category i);
Classify user-X into category i when score[i] is the maximum.
```

5.4. Model Implementation and Results Analysis

In this section, we simulate above classification algorithm with "Adorable Pet Micro-group" as example. During April 1th, 2014-April 30th, 2014, we take "fanciful fish" as the entrance account and collect the one-hop and two-hop, a total of 50,000 users, by Sina Micro-blog open platform API. After removing the duplicate users, 41,660 users are left for our model implementation and simulation, and Table 2 shows the results.

Table 2. Experiment Result of Micro-group Structure Discovery

Categories		Number of users	Proportion of users
Individuals		16,529	39.68%
Organizations	Official organizations	3,764	9.04%
	Profit organizations	5,295	12.71%
	Other informal groups	972	2.33%
Information Sources	Traditional medias	5,966	14.32%
	Micro-blog medias	2,323	5.58%
Others		6,811	16.35%

Table 2 shows the structure of Micro-groups in Sina Micro-blogs, which illustrates the following characteristics:

Firstly, the categories of online and offline social structure are basically the same. Except for that micro-blog media is a specific promoting way in SNS, other categories can be directly mapped from online social networks to real society.

Secondly, in the Micro-group structure, individual category accounts for 39.68%, far higher than other categories, making individuals the most primary members in Micro-group.

In addition, it needs to be pointed out that "others" in Table 2 refers to the users cannot be classified due to lack of information. It accounts for 16.35%, ranking the second in all categories, showing many users are reluctant to reveal much information about themselves.

In order to analyze the accuracy of the algorithm, we select 1,000 users as test data from 41,660 users. We determine their true values by artificial classification to reveal the algorithm accuracy by comparing the experiment results with them. It shows that the algorithm results of 827 users are totally the same with the artificial results, thus the accuracy is higher than 80%, ensuring the effectiveness and reliability of our classification algorithm.

6. Conclusions

The unprecedented rise of SNS provides online platforms and channels for users to publish their information, so large amount of user information are exploded in SNS nowadays. This paper proposes user features mining without labels based on data mining techniques. Two basic models are firstly proposed: the clustering model and the classification model. Then these two models are implemented for user interest discovery scenario and Micro-group structure discovery scenario. User interest discovery is to cluster users based on their micro-blog content keywords; Micro-group structure discovery is to transfer the problem to user features mining with labels by determining the labels with background knowledge. Lastly, the experiment results based on Sina Micro-blog shows that these two models can both realize the expected results, thus the models and techniques proposed in this paper is feasible.

Due to space limitations, this paper only realizes the core modules of the model. Results feedback and algorithm adjustment modules are not reflected in this paper, so it is worth us to continue discussion in future studies.

Acknowledgement

These should be brief and placed at the end of the text before the references.

References

[1] C. H. Zhang, C. B. Yu, X. N. Zhu and Y. Gao, "Communication Network Technology", Posts & Telecom Press, **(2012)**; Beijing.
[2] Y. Gao, "The Topological Analysis and Social Algorithm Research Based on the Social Networks", Beijing University of Posts and Telecommunications, **(2012)**.
[3] http://www.weibo.com.
[4] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques, Morgan Kaufmann, **(2006)**.
[5] C. A. C. Lampe, N. Ellison and C. Steinfield, "A Familiar Face(book): Profile Elements as Signals in an Online Social Network", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, **(2007)**, pp. 435-444.
[6] J. Stan, V. H. Do and P. Maret, "Semantic User Interaction Profiles for Better People Recommendation", Advances in Social Networks Analysis and Mining (ASONAM), International Conference on IEEE, **(2007)**, pp. 434-437.
[7] C. C. Aggarwal, "An Introduction to Social Network Data Analytics", Springer US, **(2011)**.
[8] M. Mezghani, C. A. Zayani, I. Amous and F. Gargouri, "A User Profile Modelling Using Social Annotations: A Survey", Proceedings of the 21st International Conference Companion on World WideWeb, **(2012)**, pp. 969-976.
[9] C. C. Aggarwal and H. Wang, "Text Mining in Social Networks", Social Network Data Analytics, Springer US, **(2011)**, pp. 353-378.
[10] S. Parthasarathy, Y. Ruan and V. Satuluri, "Community Discovery in Social Networks: Applications", Methods and Emerging Trends, Social Network Data Analytics, Springer US, **(2011)**, pp. 79-113.
[11] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 22, no. 8, **(2000)**, pp. 888-905.
[12] P. E. Danielsson, "Euclidean Distance Mapping", Computer Graphics and Image Processing, vol. 14, no. 3, **(1980)**, pp. 227-248.
[13] R. Toldo and A. Fusiello, "Robust Multiple Structures Estimation with J-linkage", Computer Vision-ECCV, **(2008)**, pp. 537-547.
[14] S. H. Cha, "Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions", vol. 1, no. 2, **(2007)**, p. 1.
[15] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm", Applied Statistics, **(1979)**, pp. 100-108.
[16] X. Jin and J. Han, "K-Medoids Clustering, Encyclopedia of Machine Learning", Springer US , **(2010)**, pp. 564-565.
[17] A. Kanaegami, K. Koike, H. Ohgashi and H. Taki, "Text Search System for Locating on the Basis of Keyword Matching and Keyword Relationship Matching", vol. 297, no. 39, **(1994)**.

[18] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren and W. Lou, "Fuzzy Keyword Search over Encrypted Data in Cloud Computing", Proceedings of Infocom IEEE, **(2010)**, pp. 1-5.

Author

Zhang Xu received his Master's degree in Software Engineering from Nanjing University of Aeronautics and Astronautics, and is currently a Ph.D. student there. His research interest is mainly in the area of data mining concepts and techniques, and Software Engineering .He has published several research papers in scholarly journals and international conferences in the above research areas.