International Journal of Grid Distribution Computing
Vol.7, no.3 (2014), pp.41-48
http://dx.doi.org/10.14257/ijgdc.2014.7.3.05

A New Clustering Model Based on Word2vec Mining on Sina Weibo Users' Tags

Bai Xue[1], Chen Fu[1] and Zhan Shaobin[2]

[1]*Department of Computer Science and Technology*
Beijing Foreign Studies University
Beijing, China
[2]*Shenzhen institute of information & technology*
Shenzhen, China

Abstract

Clustering of Weibo users is one of the most important topics in data mining on social network. Clustering can help dig out the relations among people or between people and resources. A lot of work relating to clustering has been done on analyzing personal relationship, whereas we focus our clustering model on preferences and interests. In this article, we propose a new clustering model focusing on users' tags people choose to describe themselves. First, we will study the characteristics of Sina Weibo tags of users, which are the foundation of the new clustering model. Second, we will use the word2vec tool to cluster Weibo users based on their tags and verify the accuracy of the results.

Keywords: *Weibo Tags, nodes cluster, word2vec, Sina Micro Community*

1. Introduction

Social networks have gained popularity recently with the advent of websites such as MySpace, Facebook, Sina Weibo, *etc.*, these networks are a rich source of data as users populate their sites with personal information. A fundamental problem related to these networks is the discovery of clusters or communities. Intuitively, a cluster is a collection of individuals with dense friendship patterns internally and sparse friendships externally. A lot of work has been done on this area [1] to dig out personal relationship on social network.

While people intimately relating to each other doesn't mean they have similar interests or personalities. In this article, we define the community as a group of people who share same hobbies and have a lot in common no matter whether they have direct links or not. Users' tags are any keywords chosen by users to meet their own preferences or describe their traits and characteristics. Weibo users' tags can help users organize, share and discover information resources more effectively, and deliver users' interests of the general public. So we assume that users who have same or similar tags have the same or similar preferences and interests, which is the fundamental sociological assumption of this article.

We try to measure the semantic distance between tag words and cluster tags that have relatively short semantic distances. Then people with tags in the same class should be in the same cluster. It is easy to understand that people labeling themselves with' IT' are more likely to have similar interests and preferences with those who label themselves '互联网' than those labeling '文科男'.

ISSN: 2005-4262 IJGDC

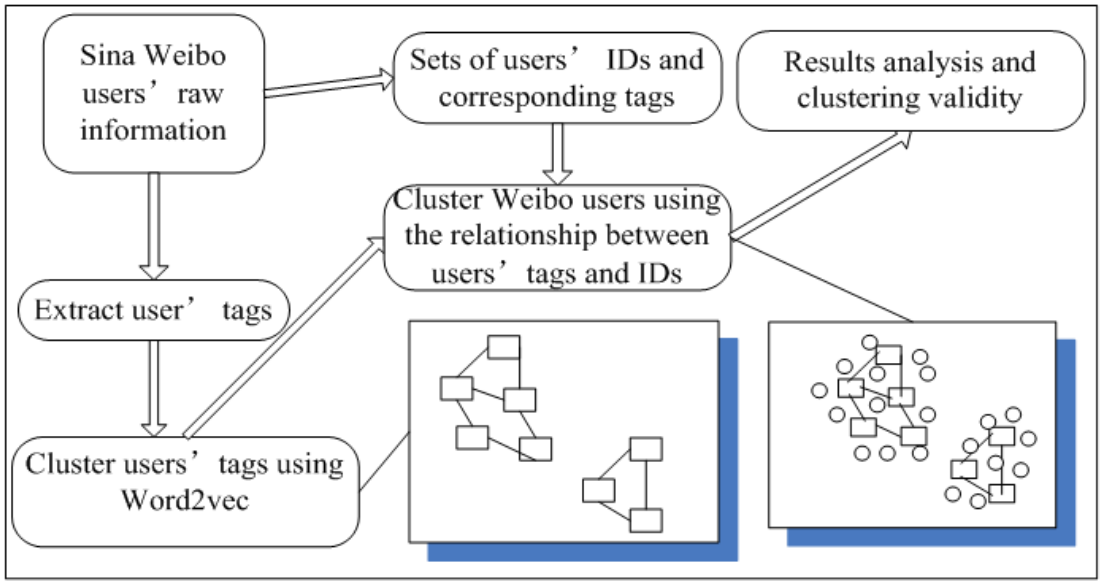

Figure 1. Flow Diagram of the New Clustering Model

People with same or similar interests and preferences joining together are a popular phenomenon nowadays, particularly on Internet. Because virtual social networks have broken the limit of distance in real world, which makes it easier to share resources. Our model presents a way to dig out the community on Sina Weibo and help users find other users sharing the same properties.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents statistical and empirical analysis of Weibo tags. Section 4 elaborates on clustering using word2vec tool and we also analyze the clustering results. In Section 5, we do a Clustering Validity to prove the effectiveness and accuracy of the clustering model. Finally, we provide some concluding remarks and suggestions for future work in Section 6.

2. Related Work

This section discusses related work on two key aspects of our model. Firstly, Subsection 2.1 presents tag clustering methods. Then, Subsection 2.2 introduces the word2vec tool.

2.1. Tag Clustering

Echarte, *et al.*, [2] discuss the problem of syntactic variations in folksonomies. They propose the utilization of pattern matching techniques to identify syntactic variations of tags. They evaluate the performance of the Levenshtein and Hamming distances using the 10,000 most popular tags and 1,577,198 annotations from CiteULike. Results show that the Levenshtein measure provides the best overall performance. However, both techniques do not perform well with tags shorter than 4 characters.

Specia and Motta [3] propose a method for building semantically related clusters of tags using a non-hierarchical clustering technique based on the co-occurrence of tags. They also explore the relationships between pairs of within-cluster tags. The authors perform a

statistical analysis of the tag space in order to identify clusters of possibly related tags. Clustering is based on the cosine similarity among tags given by their co-occurrence vectors.

2.2. Word2vec

The word2vec tool provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research. It takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The resulting word vector file can be used as features in many natural language processing and machine learning applications.

There are two main learning algorithms in word2vec: continuous bag-of-words and continuous skip-gram. The switch -cbow allows the user to pick one of these learning algorithms. Both algorithms learn the representation of a word that is useful for prediction of other words in the sentence.

A simple way to investigate the learned representations is to find the closest words for a user-specified word. The distance tool serves that purpose. For example, if you enter 'france', distance will display the most similar words and their distances to 'France'. The word vectors can be also used for deriving word classes from huge data sets. This is achieved by performing K-means clustering on top of the word vectors. The output is a vocabulary file with words and their corresponding class IDs.

3. Study of Weibo Users' Tags

In Weibo social networks, users can label themselves with tags which are any keywords chosen by users to meet their own preferences or describe their traits and characteristics. Weibo users' tags can help organize, share and discover information resources more effectively, and deliver users' interests of the general public. Users tend to use tags to present their personalities, studies of field, things they pay attention to. Thus, users' tags contain important information worthy of mining. There are 3 characteristics of Sina Weibo user's tags: suitable, long tail and ambiguity.

3.1. Suitable

From around 5 million users' information we get from Sina Weibo, we find out that users with tags are far more active than those who don't. About 20 percent users have written their own tags. Among those who have tags, about 27 percent of users have more than 100 weibos. Whereas among users who don't have tags, about 8 percent of them have more than 100 weibos. So using tags for clustering can effectively eliminate the effect of zombies and help improve the accuracy of clustering result. Furthermore, users with tags are more tend to take their Weibo seriously, so we can make sure that tags are real reflection of users' interests.

3.2. Long Tail

The default pattern of word2vec tool drops the words appearing less than 5 times. Through analyzing tags of 5 million users, we find that about 6.5% tags have appeared more than 5 times, whereas these tags cover about 84% of total users with tags. Furthermore, users' tags have high repetitive rates, only 10% tags left after deducting the repetition. So the tags users choose have high concentration, a small part have very high usage rates while large part have very small usage rates, as showed Figure 2.

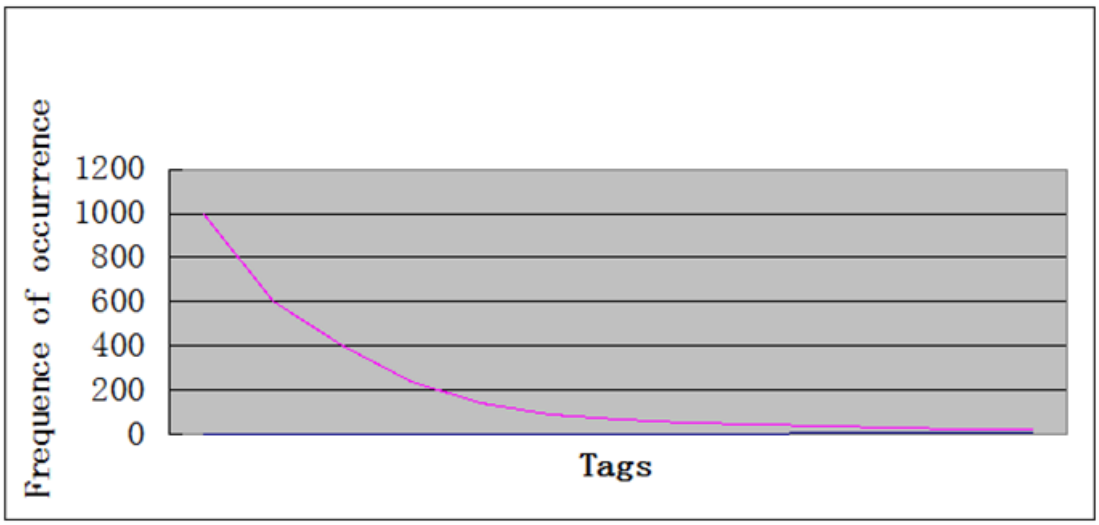

Figure 2. Distribution of Tags' Frequencies of Occurrence

3.3. Ambiguity

Because users tags have no definite patterns and rules, people can label themselves with words freely chosen according to their preferences. With regard to this situation, same words may have totally different meanings, for example writing 'apple', users can mean the fruit or the company. Also, seemingly different words may have very similar meaning, sometimes 'justin bieber' can be written as 'B 宝'. This problem can be reduced by the word2vec training process, we must include as many language contexts as possible and expand the scale of the training input text. With the development of 'Internet language', the ambiguity of language is becoming a common problem, which can be reduced but not be eliminated. The input training material and the training process are important and the key to the problem.

4. Clustering Using Word2vec Tool

4.1. Similarity Measures

We empirically analyze a real social network dataset of about 5 million users crawled from Sina Weibo. We use the word2vec tool to measure the semantic distance between tag words. The training materials are users' tag words. After training word2vec, we get distributed representation of tag words saved in vectors.bin. The semantic similarity is measured by Cosine value between two vectors.

For two n-dimension vectors a (x11,x12,…,x1n) and b(x21,x22,…,x2n), the Cosine value is calculated as follow, the bigger the value, the nearer the two vectors are in semantic sense.

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^{2}} \sqrt{\sum_{k=1}^{n} x_{2k}^{2}}} \quad (1)$$

Here are some output result calculating semantic distance between Chinese words.

Word: 互联网 Position in vocabulary: 97

Word Cosine distance

数码	0.795761
营销	0.793491
网络	0.771275
IT	0.754762
IT 互联网	0.726590
电子商务	0.701084
自由职业	0.685272
手机	0.663465
媒体	0.651999
新闻	0.649853

Word: IT Position in vocabulary: 106

Word	Cosine distance
数码	0.789488
互联网	0.754761
电子商务	0.752686
网络	0.740814
营销	0.729009
手机	0.648838
IT 互联网	0.646523
广告	0.624902
微博	0.622221
新闻	0.615561

4.2. Tag Clustering and Result Analysis

Using word2vec is a good way to cluster the words. We cluster Weibo users' tag words into 400 classes. Here is part of the result: (Words with same number are in the same class)

夜店 4
性感 4
穿衣 4
辣妹 4
爱音乐的 90 后 5
80 后 5
自由 5
旅游 5
旅行 5
听歌 5

Then we cluster Weibo users according to their tag classes. For example, if a user's tags have the word '80 后', the user should be put into the cluster 5.

The result shows that there are 10 clusters which have more than 10 thousand users, 18 clusters which have more than 1 thousand users, 246 clusters having more than 100 users, and 126 clusters having less than 100 users. The distribution of number of users in each cluster is

presented in Figure 3. The distribution also shows long-tail property, from which we know that a large number of users gather around some hot and regular topics while most clusters have unique and special topic with their own specific, professional and relatively small and stable number of members, which is Sina Micro communities. The top 10 clusters' key words are：'IT 技术', ' 精彩生活', ' 新闻资讯', ' 吃', ' 星座', ' 大学生活', ' 文学', ' 睡觉', ' 宅', ' 地点', which make sense because these words are what we usually hear about in real world, and we can learn more about people's common interests online and offline.

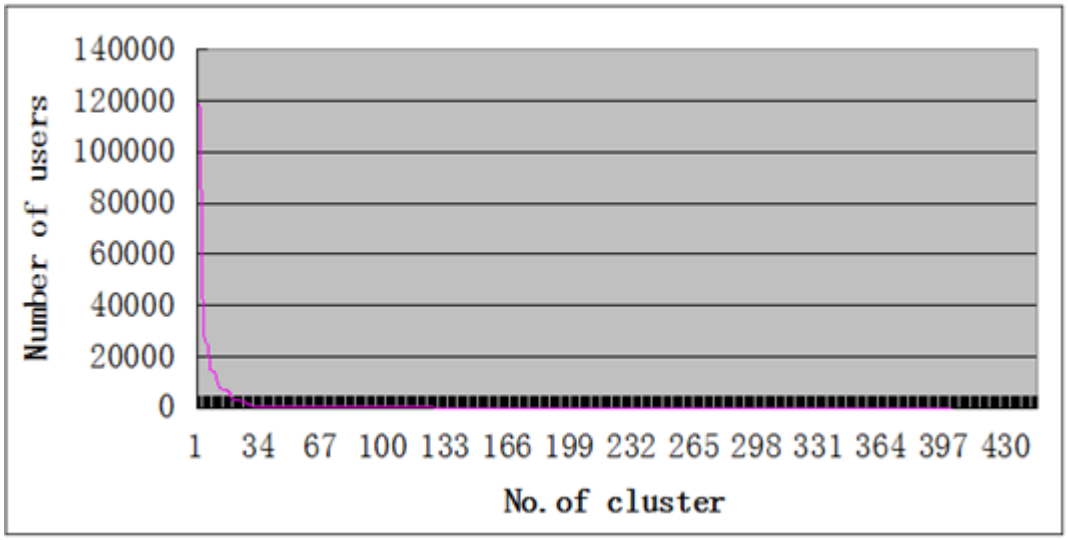

Figure 3. Distribution of Number of Users in Each Cleuster

We use the bubble chart to present the relationship between intra-cluster density, overlap rate and the size of the cluster (The size of bubble shows the size of cluster). (Figure 4) We can see that with the size of cluster increasing, the density of cluster tends to increase and the overlap rate tends to decrease. This shows that users are more concentrated in small clusters or Micro Communities and are relatively dispersed among large communities which have more everyday topics. With knowledge and information exploding on the Internet, generalization and specialization are two important properties of social network communities. The Figure 4 shows this trend well with generalization represented by large clusters and specialization represented by small clusters.

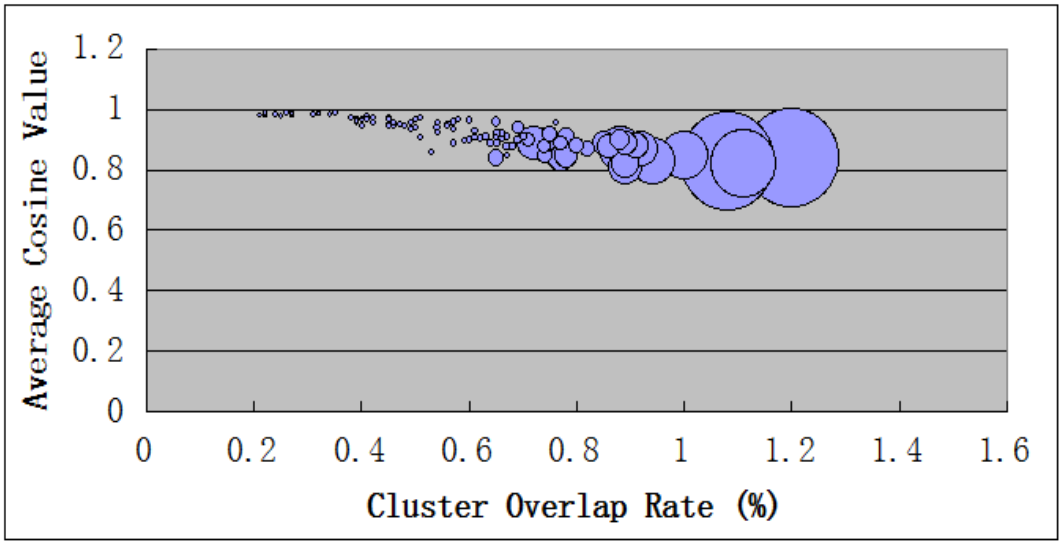

Figure 4. Relationship between Intra-Cluster Density, Overlap Rate and the Size of the Cluster

5. Clustering Validity

Because of the huge amount of data, we can't predict a cluster or two before clustering. So we will use the Internal Criteria (the properties of clustering result itself) to evaluate the clustering results.

We find that the 400 clusters have covered more than 85% users, and the overlap among clusters is only 0.5%. The high coverage and low overlap means that this cluster model is relatively accurate and acceptable. The average distance among tags in the same cluster is 0.93, while the average distance among tags in different clusters is 0.23. Inside the cluster is dense and outside the cluster is sparse, which is an important property of good clustering.

We can also use Entropy [4] to evaluate the result of clustering:

$$Entropy(X) = \sum_{t=1}^{n} - p_t \log_2 p_t \qquad (2)$$

n: number of clusters in set sample X

pt: probability of items in i cluster appearing in X

The bigger the entropy, the more the set sample is dispersed. When the probability of items in n clusters appearing in X is the same, the Entropy will get the biggest value log2 (n).

The Entropy of out clustering result is 4.63, while the Entropy of the evenly distributed clustering result is 8.96. The value shows that the clustering result is concentrated and also dispersed. This result adheres to our common sense and also to Weibo users' tags long-tail distributed property: people are inclined to center on a small number of topics, and large part of tag clusters have small number of people adhering to and also have small inter-cluster density and overlap rate, so they are more dispersed.

6. Conclusion and Future Work

We present a new clustering model based on Sina Weibo users' tags. Users'tags are any keywords chosen by users to meet their own preferences or describe their traits and characteristics. So tags can be used to find the user's properties. We use word2vec tool to measure the semantic distance between users' tag words and cluster users according to their tag classes.

We use 5 million users' information on Sina Weibo to implement our model. The clustering result has high coverage rate and low overlap rate. Also the intra-cluster density is high and inter-cluster density is low. Furthermore, the Entropy value of clustering adheres to tags' long-tail distribution property. So the result of clustering is effective and acceptable and can be used in friend's recommendation. Tags' cleansing is still important for further study. We are also looking at the problems of tag spamming and inherently ambiguous tags.

References

[1] N. Mishra, R. Schreiber, I. Stanton, *et al.*, "Clustering social networks Algorithms and Models for the Web-Graph", Springer Berlin Heidelberg, **(2007),** pp. 56-67.
[2] F. Echarte, J. J. Astrain, A. Cordoba and J. Villadangos, "Pattern Matching Techniques to Identify Syntactic Variations of Tags in Folksonomies", In: Lytras, M. D., Carroll, J.M., Damiani, E., Tennyson, R.D. (eds.) 1st World Summit on The Knowledge Society (WSKS 2008). Lecture Notes in Computer Science, Springer, vol. 5288, **(2008)**, pp. 557-564.
[3] L. Specia and E. Motta, "Integrating Folksonomies with the Semantic Web", In: Franconi, E., Kifer, M., May, W. (eds.) 4th European Semantic Web Conference (ESWC 2007). Lecture Notes in Computer Science, Springer, vol. 4519, **(2007)** pp. 503-517.
[4] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Cluster validity methods: part I", ACM Sigmod Record, vol. 31, no. 2, **(2002),** pp. 40-45.
[5] G. Begelman, P. Keller, F. Smadja, "Automated tag clustering: Improving search and exploration in the tag space", Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, **(2006),** pp. 15-33.
[6] J. Gemmell, A. Shepitsen, B. Mobasher, *et al.*, "Personalizing navigation in folksonomies using hierarchical tag clustering", Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, **(2008),** pp. 196-205.
[7] K. Kasahara, K. Matsuzawa, T. Ishikawa, *et al.*, "Viewpoint-based measurement of semantic similarity between words", Learning from Data, Springer New York, **(1996),** pp. 433-442.

