

## User Clustering Topic Recommendation Algorithm Based on Two Phase in the Social Network

Li Pei, Pan Xiaoying and Chen Hao

*School of Computer Science & Technology, Xi'an University of Posts and  
Telecommunications, Xi'an, 710121, China*  
*lipei@xupt.edu.cn*

### **Abstract**

*To deal with the issues like existing common data sparseness in weibo social network and the phenomena of cold start, this paper puts forward a two-stage clustering based on the recommendation algorithm GCCR. The algorithm firstly selects users' focused nodes which have higher number, so as to extract a dense subset of sparse data, and by using the method of graph paper, similar concerned interested core clustering is formed to this dense subset. Then, it is extracted that weibo content features of seed clustering and the whole data set other users. Then the entire user group is clustered based on content similarity. Finally the clustering results are used in subject recommendation. Through clustering the two phases of dense data subset and the whole data set, the clustering effect of extreme sparse data sets are improved. At the same time, because of fuzziness of graph clustering, this thesis retains a certain diversity in the process of user interest clustering, so as to avoid convergence too fast when cold start. This method is verified through the real social network data, and the experimental results show that this algorithm can effectively solve the problems such as data sparseness and cold start phenomenon.*

**Keywords:** Collaborative Filtering, Clustering, Data set, Fuzzy Degree

### **1. Introduction**

Social network shows the trend of rapid development with the popularity of Internet users, not only the number of users with explosive growth, but also its service form is also changing rapidly. In recent years, a large number of new social networking services constantly emerging, of which weak relationship between social network service of represented by sina weibo at home both and Facebook at abroad is becoming a major form of social network [1-3]. Different from the traditional social network, due to the unidirectional of the weak relationship, based on the nodes of social network of weak relationship (that is, a one-way relationship) presents obvious heterogeneity characteristics, including a large number of users in a natural man as major body node (e.g., "zhangsan") and in the media, institutions and various sources as the main theme node (e.g., "the weather of Beijing", "south weekend", "popular video", etc.). Among them, the user nodes, usually as a message subscriber, one-way attention to a large number of topics node, the one-way subscription relationship, often based on the user tendency of interest for different types of theme; at the same time the user node often form a two-way relationship with other users, this is usually based on the user's real social relations [4-6].

In this paper, the problem is in heterogeneous social networks of weibo classes, recommend topics node to the user (*i.e.* subscription recommendation), and deal with social networking data sparseness and cold start scenario which are common exist. To this, this paper puts forward a kind of the theme recommended method GCCR based on the user clustering of two stages. First, select the user focus in the higher number of nodes, so as to extract a dense subset of sparse data, using the method of graph paper, dense subset formed concerned interested in similar core clustering. Then, extract Weibo content features of seeds clustering and a data set focus other users, based on content similarity clustering to the entire user group, finally the clustering results used in theme recommended.

Through two-stage clustering process of the dense data subset and the whole data set, improve the clustering effect of extreme sparse data sets. At the same time, because of fuzziness of graph clustering, can retain certain diversity in the process of user interest clustering, so as to avoid convergence too fast when cold start. This paper conducted the development and innovative work mainly in the following aspects:

(1)Micro blogging social network based on weak relationship has obvious heterogeneous characteristic, according to the characteristics, the nodes can be divided into user (subscribers) and topic (publishers), user-oriented recommended the interest subjects become one of the main goals of information system in the social network, at the same time this kind of social network's phenomena of data sparseness and cold start has become the main problems in the recommendation system. This paper puts forward a kind of recommendation algorithm based on a two-stage clustering. The algorithm firstly select user focus on the higher number of nodes, so as to extract a dense subset of sparse data, the method of using graph paper, dense subset formed core clustering of concerned interested similarity. Then, extracting Weibo content features of seeds clustering and a data set focus other users, based on content similarity clustering to the entire user group, finally the clustering results used in theme recommended. Through clustering the dense data subset and the whole data set into two phases, improve the clustering effect of extreme sparse data sets. At the same time, because of kind fuzziness of graph clustering, can retain certain diversity in the process of user interest clustering, so as to avoid convergence too fast when cold start.

(2)In order to further verify the correctness and effectiveness of recommendation algorithm proposed in this paper based on a two-stage clustering, this method through the real social network data is verified, and analyzes various parameters' influence on the recommended effect. Compared with the traditional recommendation method based on the content, the algorithm can produce higher quality's recommended result across categories; it comes from the fuzzy clustering generated by the algorithm in this paper. And the method based on content in the case of lack of more subject categories, has extremely low recommend diversity, lead to rapid convergence recommendation results.

## 2. GCCR Framework

GCCR (Graph - the Content Clustering Recommendation) algorithm is designed to social networks based on users' weak relationship degree interest in the different themes, recommend theme content which might like by users. Through the analysis from users - theme preference matrix and its own published Content reflects the user preference information, and the comprehensive utilization, improve the recommended effect on sparse data sets. At the same time using the class ambiguity of the algorithm, to ensure the diversity of recommended under the cold start conditions. GCCR main steps including pretreatment, core Clustering, all user clustering, themes recommend phase, the main process is shown in Figure 1.

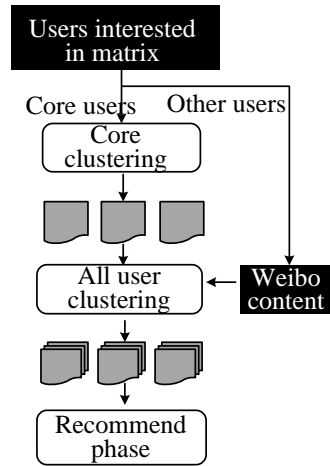


Figure 1. GCCR Main Processes

(1) Preprocessing stage. Select the collection of core users that interest vector non-zero value ratio is greater than the density threshold  $\lambda$ . According to the corresponding interest vector of core users extract dense sub-matrix of the original interest matrix.

(2) Core clustering stage. According to the subscription relationship form by core users' interest matrix to Figure the calculation diagram, using the iterative process to generate the core cluster ambiguity and independence constraint.

(3) All users clustering stage. Using the core clustering generated in the previous step, extract the core content feature vector of clustering, and extract weibo content feature vector published by the non-core users, according to similarity constant iteration of the content feature vector, non-core user will be added to the existing cluster set, until finish the clustering of all users.

(4) According to the result of clustering, and similarity of the clustering and class interest characteristics of each user, generated recommended vector within the class members, at the same time, according to different themes' interest differences between different clustering, forming across class recommendation vector, the comprehensive sequencing of the results as the final recommendation results.

In the implementation process recommendation algorithm, the training data set carries out (1) ~ (3) offline calculation generated clustering results and class interest vector, for any new users, only need to go on cluster calculation in (3) and the recommendation process. (4) Through the separation of offline and online processing operations as far as possible, it can make the algorithm achieve a higher recommend online calculation.

### 3. Two Stages' Users Clustering Topic Recommendation Algorithm

#### 3.1. Problem Modeling

(1) For  $N$  users,  $M$  theme, respectively for the user set  $y = \{y_1, y_2, \dots, y_n\}$  with theme set  $h = \{h_1, h_2, \dots, h_m\}$ . For each user  $UI$ , have corresponding interest vector  $r_i = (k_1, k_2, \dots, k_m)$ , all users' interest vector may constitute a interest matrix  $M$  of  $N \times M$ , for existing users subscribe to the relationship between the  $UI$  and theme  $h_j$ , corresponding element  $k_{ij} > 0$ , represent the user  $UI$ 's interest degree to the subject  $s_j$ , if there is no subscription, the corresponding  $k_{ij} = 0$ .

(2) Interest Figure  $G_m$  based on interest matrix  $m$  can be expressed as a directed graph  $D = (V, E)$ ,  $V$  for collection of users and theme nodes:

$$V = Y \cup h$$

E for collection of subscribing relationship:

$$E = \{e(y_i, h_j) | y_i \in y, h_j \in h, k_{ij} > 0\}$$

(3) For each user UI, define its interest density value  $des$  (UI) as the proportion of zero in the interest vector  $v_i$ , so the user  $y_i$  of  $des(y_i)$  is greater than the density threshold  $\lambda$  (usually 10%) is defined as the core. Then the core users set can be defined as

$$y' = \{y_i | y_i \in y, des(y_i) > \lambda\}$$

Interest matrix is constructed by the core user interest vector for dense sub-matrix  $m'$ , based on the dense sub-matrix can construct Figure Gm 'core interests.

### 3.2. Core Clustering

(1) through the dense interest matrix  $m'$  constructing core interest Figure Gm 'S (V, E) in core user set  $y'$  and theme collection S, a set of clustering set  $Clus$  on the user set  $y'$  can be expressed as the user clustering  $c_i$  collection, including:

$$y' = \bigcup_{i=1}^n c_i, c_i \neq \emptyset, \text{ and } i \neq j, c_i \cap c_j = \emptyset$$

For each theme  $s_j$ , we define the participation set of  $c_i$ :

$$E_{s_j}(c_i) = \{y | y \in c_i \text{ and } (y, s_j) \in R\}$$

So participation  $q_{ij}$  meet:

$$q_{ij} = \frac{|q_{s_j}(c_i)|}{|c_i|} > \sigma (\sigma > 0, \text{ is Intensity threshold})$$

$c_i$  and  $s_j$  are called "Clustering  $c_i$  strong focus on the theme of  $s_j$

(2) Define the user clustering  $c_i$ 's  $Amb_{ij}$  on the main topic of the  $s_j$ :

$$Amb_{ij} = \begin{cases} |c_i - q_{s_j}(c_i)|, & q_{ij} \geq \sigma \\ |q_{s_j}(c_i)|, & q_{ij} < \sigma \end{cases}$$

Which can be defined as  $c_i$  collection for subject S ambiguity:

$$Amb_{ij} = \sum_{s_j \in S} Amb_{ij}$$

The degree of  $Clus$  on user of set  $y$ 's global fuzzy is the theme set S

$$Amb = \log \left[ \frac{\sum_{c_i \in Clus} Amb_{ij}}{|Clus|} \right]$$

The exponential here is to ensure that changes linearly ambiguity with the clustering of global growth trend.

(3) User clustering  $c_i$ 's interest degree is defined as 3 on theme  $s_j$

$$ca_{ij} = \begin{cases} \frac{\sum_{y_k \in c_i} a_{kj}}{|c_i|}, & q_{ij} \geq \sigma \\ 0, & q_{ij} < \sigma \end{cases}$$

The  $c_i$  class interest vector on the theme set S is

$$cv_i = (ca_{i1}, ca_{i2}, \dots, ca_{im})$$

Each nonzero component corresponds to a strong focus on relationship. With different user clustering on the theme set  $S$  interest vector distance from each other; to measure the clustering results reflect user community interest difference degree. The interest distance between two clusters using cosine distance:

$$\text{diff}(c_i, c_j) = \frac{cv_i \cdot cv_j}{|cv_i| \cdot |cv_j|}$$

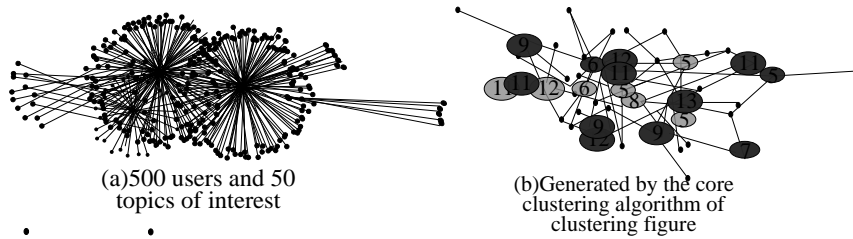
A set of clustering  $Clus$  on the user set  $y'$  for theme collection  $s's$  difference index

$$dvst = \frac{\sum_{c_i \in Clus, c_j \in Clus} \text{diff}(c_i, c_j)}{|Clus|}$$

### 3.3. All User Clustering

After getting user core clustering  $s_{imik}$ , we need to extract the core clustering and non-core content feature vector.

For the users  $y_i$ , published weibo for OriginTweet  $s_i$ , its first to the original weibo data preprocessing, such as removing the emoticons in weibo, remove the "@" someone's information, and so on, to get users to post the plain text weibo content Tweet  $s_i$ .



**Figure 2. Interest Figure Contrast Before and After Clustering**

Define the feature vector of user  $y_i$  is  $v_{y_i}$ ,  $v_{y_i} = (\text{Tweet } s_i)$ . Core clustering  $Clus_j$  characteristic vector for  $VClus_j$ , have  $VClus_j = (\text{Tweet } s_m)$ ,  $y_m \in Clus_j$ .

We adopt the improved edit distance algorithm to calculate the characteristic vector similarity, edit distance was originally used to measure the similarity between the strings, and using the single character as the basic computing unit, in order to make it more suitable for Chinese sentences with semantic similarity calculation, the algorithm used a single word in automatic segmentation of a sentence, as the basic editor unit. In addition algorithm considering the edit operation cost and sentence length influence on similarity, put forward the new block switching operation, and according to the semantic similarity between words gives different weight to different editing operations, under the premise of not using semantic disambiguation and syntactic analysis, semantic information both the sentence structure and vocabulary.

For the users  $y_i$ , we use the improved edit distance algorithm to calculate he and all his core clustering  $Clus_j$  similarity  $s_{imij}$ , if the maximum is  $s_{imik}$ , the user  $U_i$  join the clustering  $Clus_k$ . After all of its non-core user added to the corresponding clustering, can get to all users clustering  $GClus$ .

### 3.4. Recommend Stage

Get full user clustering  $GClus$ , can calculate theme set  $S$  class' interest vector in each user clustering  $c_i$ :

$$cv_i = (ca_{i1}, ca_{i2}, \dots, ca_{im})$$

All the class interest vector of clustering may constitute a kind of interest matrix  $m$ , for the zero value, using the Slope One algorithm to predict. Defined average interest deviation between theme  $s_i$  and  $s_j$ .

$$dev_{i,j} = \sum_{c_i \in G_{clus}} \frac{ca_{ki} - ca_{kj}}{|G_{clus}|}$$

So for any zero component, all can be predicted by the following formula which  $\bar{ca}_i$  is for the average value of each component of vector  $cv_i$ ,  $M-1$  is under the situation when  $I = j$ ,  $dev_j$ ,  $I$  value is zero

$$ca_j = ca_i + \frac{\sum_{i=1}^m dev_{i,j}}{M-1}$$

The zero filled with predictive value in original vector, get forecast interest vector  $CV'$ , sorting for each component interest value, for each user, except it is already the subject of attention, interest in the rest of the theme in accordance with the Top - K value is recommended. In practice, we usually take  $K$  value for the user has concerned topics or half that number.

In the case of online recommendation, first of all, need to recommend the user, you can extract the content characteristic vector and use the process of classification in all users clustering process, the user assigned to the appropriate clustering, using the predictive vector  $CV'$  of clustering to the recommendation. You can see, the whole process, in addition to user classification process need real-time calculation, the user clustering and interest value prediction can be directly used in advance after offline processing results.

The computational complexity of online recommendation is only related to the user clustering number, and user clustering number is very limited in the actual case, it also ensures the online recommendations efficiency of the algorithm. For the result of clustering and recommended, the need in the new user increased to a certain number and obvious impact on the interest distribution adjustment.

## 4. Experiment and Analysis

### 4.1. Data Set

Despite the research content of this article is based on an existing user - topics of interest numerical matrix, but we cannot directly get this interest in quantitative index on the real data set. Therefore we need to build the a measure of user interest index in the experiment, although this has nothing to do with the algorithm's description in this article, for the sake of performance experiment, the effect of this work is necessary.

Use "users expect review rate" to describe the interest degree of user in the topic, its significance in weibo system can be understood as a user on a particular topic comment content or forwarded by the probability of potential, the index turn after with users itself evaluation of rate regulation, approximate probability formula can use the following conditions:

$$a = \frac{q(r|R)}{q(r)} = \frac{q(r) \cdot q(R|r)}{q(R) \cdot q(r)} = \frac{q(R|r)}{q(R)}$$

Among them, the  $q(R)$  for probability from reading to the subject  $R$ ,  $q(R|r)$  forward for users, comments from the content of the theme of  $R$  probability, more than two probability can be used approximate the statistical results of the experimental data set. In the following discussion, use the interest value measurement as the basis.

Experimental data are grabbed through open platform of sina weibo and API. Because there are a mass of users information in the social network, simple random fetching nodes

can lead to the experimental data too sparse, also cannot reflect the weak relationship in the structure features of the social network. Therefore, we use the way of generating interest Figure in the network of sina weibo to simulate the formation process of online community based on weak relation step by step, with the opening of the seed users, thus obtains local samples of the heterogeneous social network features. Main processes are: (1) 5 ~ 10 nodes adjacent or close to the user as a seed. (2) For each iteration, with the method of depth priority, crawl users nodes adjacent with the current users; Or with the method of breadth first, grasp the current theme section of the user's attention points. (3) According to the average ratio of user nodes and theme nodes, to adjust the proportion of two kinds of grab in the process of iteration. (4) according to the crawled users set and theme set, obtain detailed attention, forward, review data, according to the formula above, calculate "user expectations review rate", the end user - topic interest matrix is obtained.

In this experiment, much attention are grabbed - interest groups of different matrix. The final experimental results are the average values of various experimental data. Of which, each set contains about 500 users, 50 theme and nearly 20000 weibo content. Experiment implementation by Python and Java code. The code runs on the MacBook Pro Mc990, Python version 2.7, the JDK version 1.7.

Reference of algorithm is checked as: (1) Collaborative Filtering recommendation algorithm based on Top - K similar (Collaborative-Filtering CF); (2) based on  $K$  neighbor recommendation algorithm of topic Content (Content-based) similarity. The control algorithm of machine learning based on open source libraries Apache Mahout and implementation. The Collaborative-Filtering algorithm Collaborative Filtering for user-based, user similarity computing using Pearson correlation coefficient, the final recommendation results use the Top-K recommended. In the Content-based algorithm, the similarity of theme calculates with the Chinese sentence similarity. In the experiment, we will use half of theme as the training set, and the other half theme will be carried out experiment as a test set.

#### 4.2. Recommended Effect

In the experiment, several algorithms under the condition of facing sparse data sets, its ability to produce recommendations results are different, Table 1 shows the different data sparseness degree cases, several kinds of algorithm will have the largest number of recommended results contrast (in this experiment we limit the maximum recommended amount shall not exceed the number of non zero interest value test set).

**Table 1. Comparison of the Number of Recommended Results**

Sparsity/%	Required to be recommended number	Actual recommended number / required to be recommended number/%			
		CF,Top-5	CF,Top-10	Content-based	GCCR
5.56	727	7.2	39.9	16.5	99.5
6.77	640	8.3	42.0	18.4	99.6
8.64	522	8.4	43.9	20.9	100.0
11.81	170	8.9	45.8	19.01	100.0

Obviously, in the case of extreme sparse data, CF and the method based on content cannot produce enough recommend result, the recommended ability of GCCR method is little influenced by the data sparseness.

Then, we compared several algorithms accuracy and recall rate under the optimal parameters. We always take the focus number of recommended number is equal to the number of training focus, for the CF and the Content - -based algorithm, we take Top - k number for 10. Accuracy can be expressed as the ratio of the number of recommend hits and the total recommended amount, the recall rate can be expressed as ratio of

recommended hit number and the focus number in test set. Figure 3 (a) shows the recommend accuracy of the algorithm under the different levels of data sparseness.

As you can see, in the case of the extreme data sparseness (density is lower than 10%), the accuracy of the collaborative filtering algorithm is worst, gradually ascending when the density is more than 10%. The recommended method based on content is not sensitive to sparse data, but accuracy maintained at high level. GCCR algorithm can maintain high recommend accuracy under the extreme condition that the data are sparse, and the data density increase can also gradually increase, this has to do with this algorithm's realization of data intensive in the process of Figure. Thus, under the scenario of this kind of micro blogging heterogeneous network, due to the prevalence of sparse data (usually less than 10%), GCCR has significant improvement compared with the traditional method.

Figure 3 (b) shows the contrast between recall rates of different algorithms. As you can see, GCCR algorithm in sparse data set, the recall rate remain in a stable high. While the CF and Content - -based method, are greatly influenced by the data sparseness sex, have poor recommended the quantity and quality, of which the Content - -based method make the recall rate is very low due to the small number of produce recommendation results.

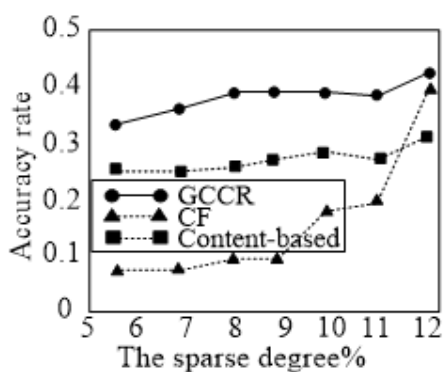
In order to better compare recommendation quality of algorithm, we introduce the  $F_{measure}$  and  $MAP$  (Mean Average Precision) two index. Among them, the  $F_{measure}$  harmonic Mean for accuracy and recall rate, the higher this value is, the better comprehensive performance of the recommend algorithm is:

$$F_{measure} = \frac{2 \cdot precision \cdot recall}{(precision + recall)}$$

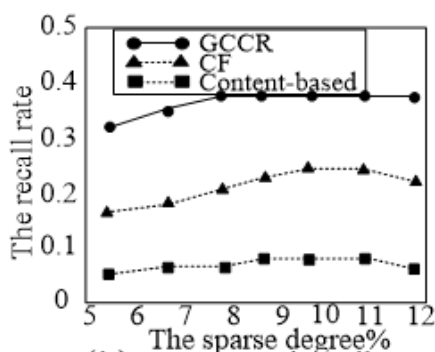
At the same time, the recommendation results'  $MAP$  of a user group produced can be defined for AP (Average Precision)' Average value of each user recommendation results, the higher the value, suggests that the better the overall recommend quality of recommendation algorithm:

$$MAP = \frac{\sum_{k=1}^y AP(k)}{Y}$$

The AP value indicates that the average accuracy of the result of the recommendation from a user. Three algorithms'  $F_{measure}$  value and  $MAP$  respectively as shown in Figure 3(c) and is shown in Figure 3 (d).

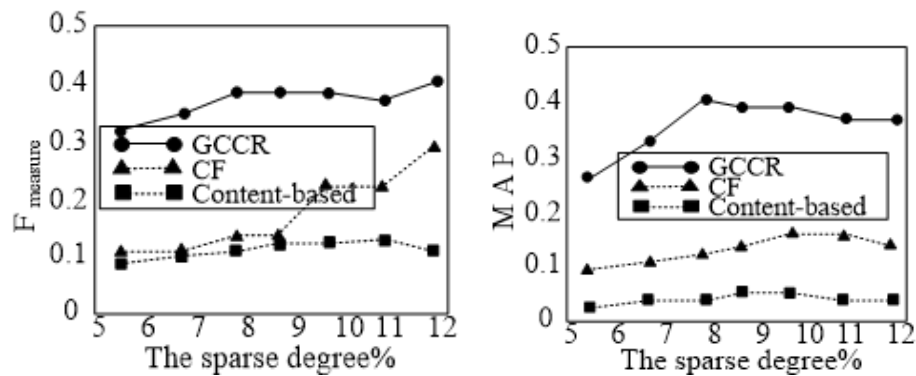


(a) Recommendation Accuracy



(b) Recommend Recall Rate





(c) Recommendation Algorithm  $F_{measure}$  Value (d) Recommendation Algorithm MAP Values

**Figure 3. Recommended Effect of Three Kinds of Algorithm**

As you can see, GCCR algorithm has a larger advantage, no matter on the recommended comprehensive performance and recommends quality.

#### 4.3. Diversity

50 topics are divided into text, fashion, jokes, information and other 16 categories by us through the manual annotation. Due to the cold start scenarios, new users usually focus on the limited topic, thus the interest scope may only reflect in several categories of subject, recommendation algorithm in order to avoid too quick convergence, we hope that we can produce more diversified recommended result within a certain range, so as to improve the recommended possibility of cross category. We will train the attention of a category subject which does not exists concentrate and generate the recommendation definition of the category subject as "recommended across categories, at the same time, such recommendation results called "across the category" in the hit of a test set.

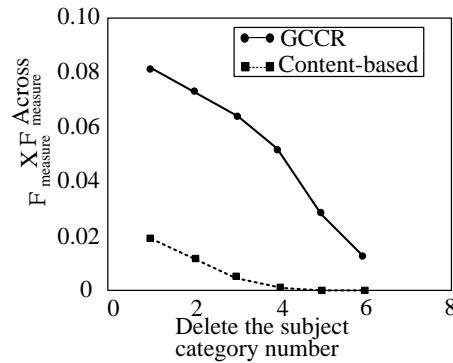
In the experiments, through targeted deletion in the training data set on one or several kind of theme attention information, to simulate the characteristics of the data under cold start scene, to test the algorithm's ability of across class recommendation. Table 2 shows a set of results in the experiment.

**Table 2. Across Class's Recommendation Results**

Algorithm	Removed number of categories	Recommended number	Recommended number of cross-category	Number of hits	Number of cross-category hit
Content-based	1	272	72	60	22
	2	186	100	49	36
	3	135	89	40	28
GCCR	1	746	104	323	43
	2	982	198	329	81
	3	1261	216	277	85

GCCR algorithm in simulation of the cold start scenarios, generating cross class recommended quantity and quality are significantly better.

In Figure 4, we adopt the product of the across recommended  $F_{measure}$  value and overall recommendation result  $F_{measure}$  in recommendation results to measure diversity of recommendation.



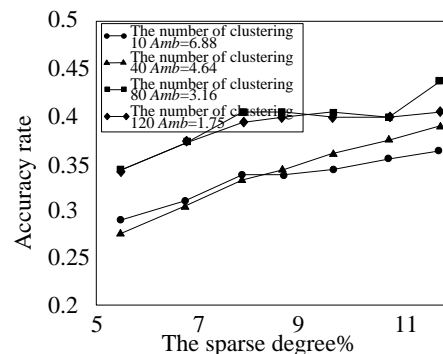
**Figure 4. Recommend Diversity Comparison**

Compared with the traditional recommendation method based on content, GCCR can produce across categories' recommendation result of higher quality, it comes from the fuzzy clustering by generated the GCCR. And the method based on content in the case of lack of more subject categories, has extremely low recommend diversity, lead to rapid convergence recommendation results.

#### 4.4. Influence of Various Parameters on the Effect of Recommendation

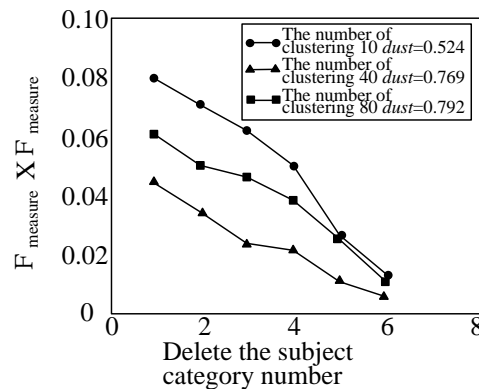
##### Ambiguity and Diversity Index

Ambiguity is one of among the members of the cluster, for subject degree of attention to the differences of measurement, expressed in  $A_{mb}$ . We can see in Figure 5, the current global fuzzy degree of clustering results decreased with the increase of the number of clustering, this is because when clustering become small, it will be easier to form a strong focus on relationships. At the same time, the overall effect of the recommendation algorithm improved with the decrease of ambiguity, and with the increase of density of data set, the gap is more obvious. When clustering number too much, however, the accuracy recommendation will be reduced, this is due to too small clustering makes interest become sparse matrix.



**Figure 5. Recommend Accuracy under different Fuzzy Degree**

Difference index reflects average interest difference degree between clustering, it increases with the increase in the number of clustering, we can see in Figure 6, with the increase of diversity index, with DVST represents. The diversity of recommended effect gradually reduced. Clustering, at a number of 10,  $DVST = 0.769$  is the minimum value, the more the recommendation diversity, when clustering number is 81, DVST has the minimum value, the smaller clustering makes recommended diversity decreased significantly at this time. This can be understood as the interest differences increases between clustering, clustering internal interest more consistent, produce recommendations of across categories is more difficult.



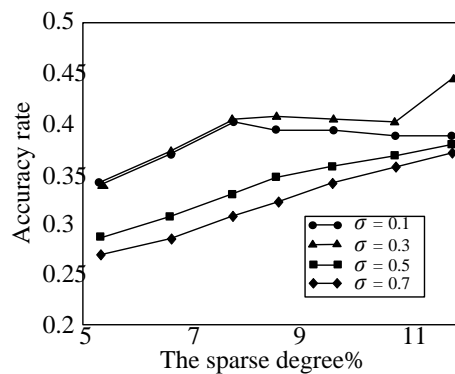
**Figure 6. Recommend Diversity Under different Diversity Index**

From the experiment results above, choosing the optimal number of clustering, recommend diversity and accuracy need to be considered at the same time. The influence of more fuzzy clustering number is that each cluster degree is reduced, while improves the accuracy of recommendation narrowed the prediction range of interest, thus reducing the possibility of producing recommendation across class. On the contrary, the less the number of clustering can provide a wider recommendation scope, thus improve the recommendation effect of cold start. As a result, the number of clustering depends on the recommended requirements. In practice, in the absence of a clear tendency, we choose the clustering result which makes the product between diversity index and the fuzzy degree reaches the maximum value.

#### Relation Intensity Threshold $\sigma$

$\sigma$  defining strong focus on relationships minimum coverage in a cluster,  $\sigma$  values determines the confidence degree of clustering interest in the process of clustering. When we need presumptive a clustering is interested in a particular subject, if the greater the  $\sigma$  value is, the need more members meet in such the attention to this topic. If the smaller the  $\sigma$  value is, for clustering decision condition is loosening. Figure 7 shows the  $\sigma$  value influence on predicting accuracy.

$\sigma$  define strong focus on relationships' minimum coverage in a cluster, such as Tian makes  $\sigma$  value is 0.5, and in GCCR implementation, facing of more sparse data sets, the relative ease strong relation judgment conditions (optimum when the  $\sigma = 0.5$ ), and makes the cluster formed by the Figure has more than zero interest value, which can achieve better recommendation results. But when there are low intensity threshold, the cause of the recommended effect decline is the class interest for judgment is too fuzzy.



**Figure 7. Recommendation Accuracy Under different  $\sigma$  Values**

## 5. Conclusion

In order to solve the data sparse and cold start problem that exist in the weibo heterogeneous social network, the recommendation algorithm GCCR based on the diagram, similar in content and hybrid clustering is proposed in this paper in this paper, GCCR in extremely sparse data sets with high accuracy, at the same time, under the scenarios of cold start can provide diversity of recommended results, thus to avoid the problem of recommended results fast convergence. Finally, the effect of the algorithm is verified by real data sets, and the influence of various parameters on the recommendation results is analyzed.

## Acknowledgments

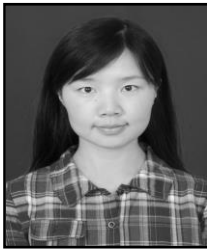
This research work is supported by the National Science Foundation of China (Grant No.61105064), the National Science Foundation of China (Grant No. 61203311), the National Science Basic Research Plan in Shaanxi Province of China (No.2011JM8007), the Scientific Research Projects Supported by Shaanxi province education department (NO.2013JK1183), Middle-aged and Young Teachers' Scientific Research Funds of Xi 'an Post and Telecommunications University (ZL2013-24)

## References

- [1] L. Wan and J. Yang, "Advanced Split BIRCH Algorithm in Reconfigurable Network", Journal of Networks, vol. 8, no. 9, (2013), pp.2050-2056.
- [2] H. Z. Guo, Q. C. Chen, C. Xin and X. L. Wang. "A Length-variable Feature Code Based Fuzzy Duplicates Elimination Approach for Large Scale Chinese WebPages", Journal of Software, vol. 7, no 11, (2012), pp.262-269.
- [3] K. Atual and J. Weissman, "ViDeDup: An Application-Aware Framework for Video De-duplication", In Proc. of the 3rd USENIX conference on Hot topics in storage and file systems, (2011), pp.7-7.
- [4] X. B. Li and Q. Zhou, "A lossless data hiding transmission method for satellite remote sensing image based on histogram modification," Journal of Astronautics, vol. 34, no. 5, May (2013), pp. 686- 692.
- [5] Sheng L., Gang W., Stones D. S., Xiao G. L. and Jing L., "T-code: 3 Erasure Longest Lowest-Density MDS Codes", IEEE Journal on Selected Areas in Communications, vol. 28, no. 2, (2010), pp. 289-296.
- [6] L. Zhihan, L. Feng, S. Feng and H. Li, "Extending Touch-less Interaction on Vision Based Wearable Device", Virtual Reality (VR), IEEE, (2015).
- [7] Z. Mengxin, Z. Lv, X Zhang, G. Chen and K. Zhang, "Research and Application of the 3D Virtual Community Based on WEBVR and RIA", Computer and Information Science, vol. 2, no. 1, (2009), pp. 84.
- [8] C. Wengert, M. Douze and H. Jegou, "Bag-of-colors for Improved Image Search", In Proc. of the 19th ACM international conference on Multimedia, (2011), pp.1437-1440.

- [9] Chen T. W., Shen G. W. and Xu B. H., “H-Code: A Hybrid MDS Array Code to Optimize Partial Stripe Writes in RAID-6”, Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium, Anchorage, Alaska, USA, IEEE Press, (2011), pp.782-793.
- [10] J. Tian, “Reversible watermarking using a difference expansion”, IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 8, August (2003), pp. 890–896.
- [11] L. An, X. Gao, C. Deng and F. Ji, “Reversible watermarking based on statistical quantity histogram”, in Proc. Adv. Multimedia Inform. Process, LNCS 58791, (2009), pp.1300-1305.
- [12] Feng G. L., Deng R., Bao F. and Shen J. C., “New Efficient MDS Array Codes for RAID, Part I: Reed Solomon like Codes for Tolerating Three Disk Failures”, IEEE Trans. Computers, vol. 54, no 9, (2005), pp. 1071-1080.
- [13] Y. Geng, J. He, H. Deng and K. Pahlavan, Modeling the Effect of Human Body on TOA Ranging for Indoor Human Tracking with Wrist Mounted Sensor, 16th International Symposium on Wireless Personal Multimedia Communications (WPMC), Atlantic City, NJ, (2013) Jun.
- [14] Y. Geng, J. He and K. Pahlavan, “Modeling the Effect of Human Body on TOA Based Indoor Human Tracking”, International Journal of Wireless Information Networks, vol. 20, no. 4, (2013), pp. 306-317.
- [15] X. Zhang, Y. Ren, G. Feng and Z. Qian, “Compressing encrypted image using compressive sensing”, 2011 7th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), (2011), pp. 222-225.
- [16] Xiang H. L. and Ji W. S., “Summary of Research for Erasure Code in Storage System”, Journal of Compute Research and Development, vol. 49, no. 1, (2012), pp.1-11.
- [17] P. Lu, Z. Xu, X. Lu and X. Liu, “Digital image information encryption based on Compressive Sensing and double random-phase encoding technique”, Optik-International Journal for Light and Electron Optics, vol. 124, (2013), pp. 2514-2518.

## Author



**Li Pei**, she received her M.S. degree in Control science and engineering from Xi'an Jiao Tong university in Xi'an, China, She is currently an associate professor in College of computer at Xi'an university of Posts and Telecommunications. Her research interest is mainly in the area of network security, intelligent information processing. She has published several research papers in scholarly journals in the above research areas and has participated in several books.

