User Algorithm based on Social Network

Hean LIU, Zhike KUANG
Hunan City University, Yiyang Hunan 413000, CHINA

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. 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

Different from the traditional social network, due to the unidirectional of the weak relationship, Future communication network is a IP-based seamless blending heterogeneous system in which various wireless access networks can coexist. As heterogeneous wireless network resources have diversities, to achieve unified management of radio resources, and to improve the utilization of radio resources become the research focus in the current radio resource management [4-6]. Grid technology is a kind of technology that can effectively and safely manage and share a variety of resources connected to network, and can provide corresponding services [7-9].

The theme node, on the other hand, as news publishers, subscriptions by a large number of user node, and the relationship between initiative and two-way focus number is far less than the number of subscription. Figure 1 (a) shows a typical social network structure based on the strong relationship, presents the homogeneity in the network node. Figure 1 (b) for a typical heterogeneous weak relationship social network from sina weibo (node for the user, a white point node for the theme, dotted lines for one-way subscription relations, solid line for mutual relationship).

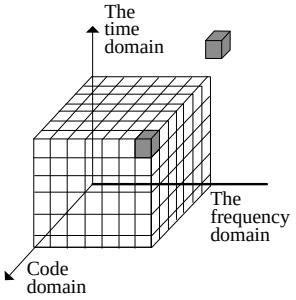


Figure 1. The typical social network structure between strong and weak relationship

As shown in Figure 2, according to the sina weibo statistics sampling of 500 users and 50 theme, only 20% users have subscribe relationship to the concern of more than 10% theme, and focus on the theme number less than 5% of the total number more than half of the proportion. For such a sparse data, such as collaborative filtering simple method based on binary relation cannot achieve the ideal effect of recommendation.

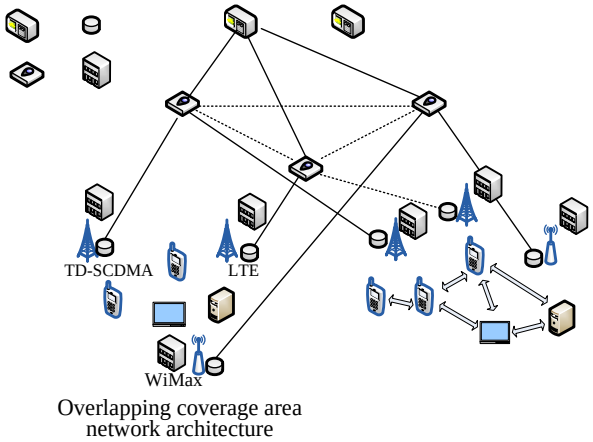


Figure 2. Sampling statistics of sina weibo user attention

In addition, new users are added to the social network often face the cold start problem. New users tend to show little interest in tendencies, and recommendation method based on content generally do not have enough variety, recommendation results can quickly converge to a collection of small range, and losing the possibility that more users interested content may have recommendation.

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.

2. GCCR Framework

GCCR 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 3.

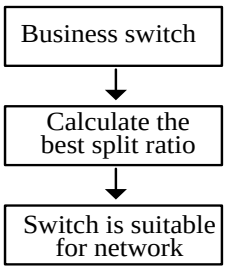


Figure 3. GCCR main processes

2.1. 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:

$$y' = \overset{n}{\underset{i=1}{y}} c_i, c_i \neq \varnothing, and\ i \neq j, c_i \cap c_j = \varnothing$$

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

$$s_{m'}^{i,k} = \frac{u_{i,m}(e_m d - In\square)}{(\lambda_i + \gamma_i)u - In\square}$$

So participation q_{ij} meet:

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

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 :

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

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

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

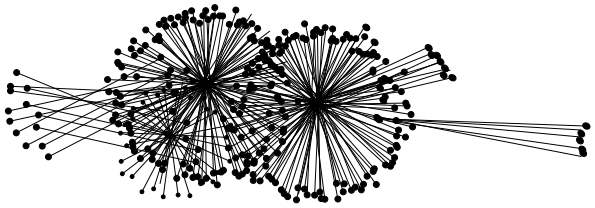
The degree of *Clus* on user of set $y"s$ global fuzzy is the theme set s

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

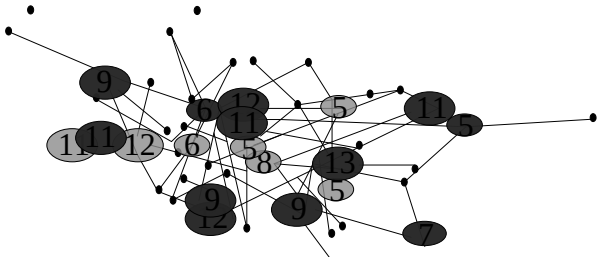
The exponential here is to ensure that changes linearly ambiguity with the clustering of global growth trend.

2.2. All user clustering

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.



(a)500 users and 50 topics of interest



(b)Generated by the core clustering algorithm of clustering figure

Figure 4. Interest figure contrast before and after clustering

Define the feature vector of user y_i is v_{y_i}, $v_{y_i} = (Tweet\, s_i)$. Core clustering $Clus_j$ characteristic

vector for $VClus_j$, have $VClus_j = (Tweet\, s_m)$, $y_m \in Clus_j$.

2.3. Recommend stage

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

$$c_{im} = s_i \beta_{im}$$

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 .

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

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

$$J_{i,m} = l_{i,m} * \frac{\beta_{i,m}}{c_{im}}$$

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.

3. The Experiment and Analysis

3.1. 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:

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

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 calculate 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.

3.2. Recommended effect

In the experiment, several algorithms under the condition of facing sparse data sets, its ability to produce recom-

mendations 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. Compaarison 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 5 (a) shows the recommend accuracy of the algorithm under the different levels of data sparseness.

Figure 5 (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.

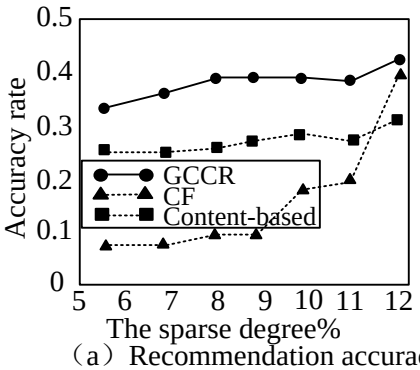
In order to better compare recommendation quality of algorithm, we introduce the $F_{measure}$ and *MAP* (Mean Average Precision) two index. Among them, the $F_{measure}$ harmonic Mean for accuracy and recall rate, the higher this value is, the better comprehensive performance of the recommend algorithm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

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

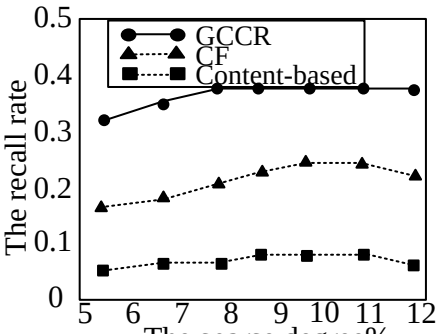
At the same time, the recommendation results' *MAP* of a user group produced can be defined for AP (Average Precision)' Average value of each user recommendation results, the higher the value, suggests that the better the overall recommend quality of recommendation algorithm:

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

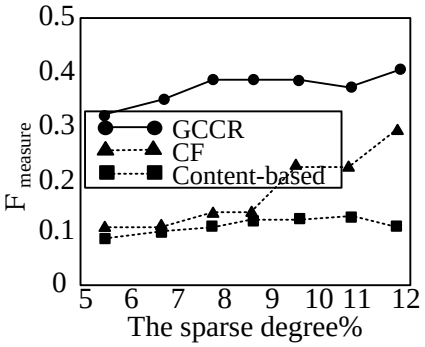
The AP value indicates that the average accuracy of the result of the recommendation from a user. Three algorithms' $F_{measure}$ value and *MAP* respectively as shown in figure 5 (c) and is shown in figure 5 (d).



（a）Recommendation accuracy



（b）Recommend recall rate



(c)Recommendation algorithm $F_{measure}$ value

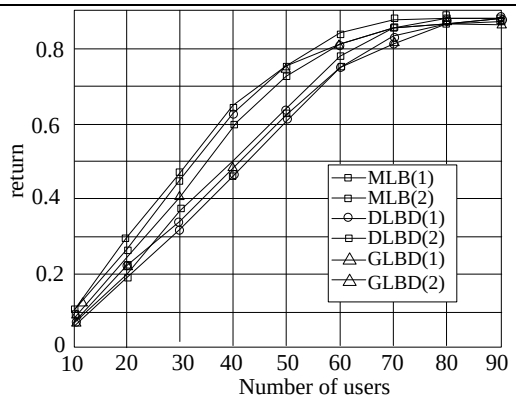


Figure 5. Recommended effect of three kinds of algorithm

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.

3.3. Influence of various parameters on the effect of recommendation

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 7, 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.

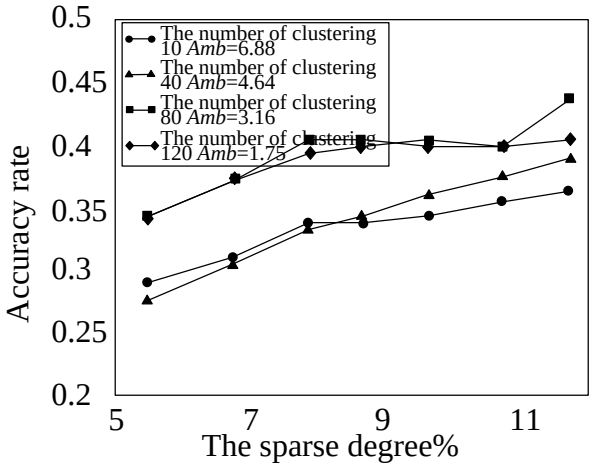


Figure 6. Recommend accuracy under different fuzzy degree

Difference index reflects average interest difference degree between clustering, it increases with the increase in the number of clustering, we can see in figure 8, 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.

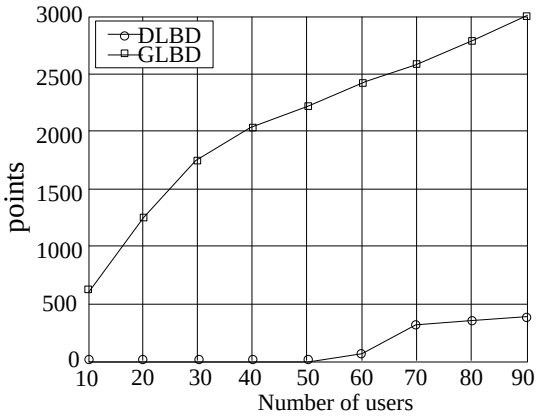


Figure 7. Recommend diversity under different diversity index

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.

3.3. Relation Intensity Threshold σ

σ defining strong focus on relationships minimum coverage in a cluster, σ 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 σ value is, the need

more members meet in such the attention to this topic. If the smaller the σ value is, for clustering decision condition is loosening. Figure 9 shows the σ value influence on predicting accuracy.

σ define strong focus on relationships' minimum coverage in a cluster, such as Tian makes σ 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.

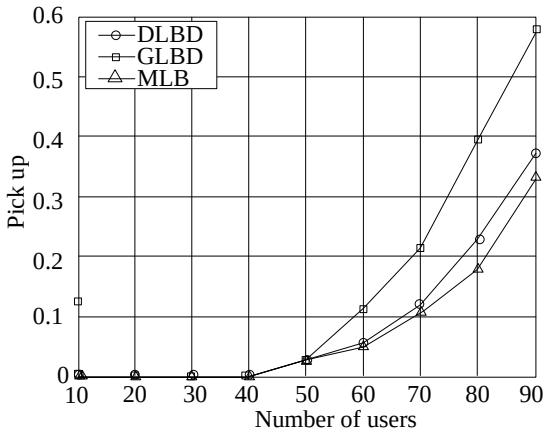


Figure 8. Recommendation accuracy under different σ values

4. 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.

References

[1] Li Wan, Jie Yang. Advanced Split BIRCH Algorithm in Reconfigurable Network. Journal of Networks, Vol 8, No 9 (2013), 2050-2056

[2] SHENG L, GANG W, STONES DS, et al. T-code: 3 Erasure Longest Lowest-Density MDS Codes . IEEE Journal on Selected Areas in Communications, 2010, 28(2):289-296.

[3] N. P. Ramaiah, "De-duplication of Photograph Images Using Histogram Refinement", Recent Advances in Intelligent Computational Systems, 2011, pp.391-395

[4] A. V. Sreedhanya and K. P. Soman, "Secrecy of cryptography with compressed sensing," International Conference on Advances in Computing and Communications, pp. 207-210, 2012.

[5] C. Wengert, M. Douze, H. Jegou. "Bag-of-colors for Improved Image Search", In Proc. of the 19th ACM international conference on Multimedia, 2011, pp.1437-1440

[6] CHEN T W., SHEN G W., XU B H., et al.H-Code : A Hybird 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,2011:782-793.

[7] FENG G.-L., DENG R., BAO F., SHEN J.-C.. New Efficient MDS Array Codes for RAID, Part I: Reed Solomon Like Codes for Tolerating Three Disk Failures. IEEE Trans. Computers, 2005,54(9): 1071-1080.

[8] X. Zhang, Y. Ren, G. Feng and Z. Qian, "Compressing encrypted image using compressive sensing," 2011 7th International Conference on Intelligent Information Hiding and Multimedias Signal Processing (IIH-MSP), pp. 222-225, 2011.

[9] XIANG H L, JI W S. Summary of Research for Erasure Code in Storage System. Journal of Compute Research and Development,2012, 49(1): 1-11.