International Journal of Intelligent Information and Management Science

Volume 6, Issue 3, June, 2017

http://www.hknccp.org

Publisher: HongKong New Century Cultural Publishing House
Address: Unit A1, 7/F, Cheuk Nang Plaza, 250 Hennessy Road, Wanchai, Hong Kong

---

International Journal of Intelligent Information and Management Science
ISSN: 2307-0692, Volume 6, Issue 3, June, 2017

Contents

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

Shu LI
Hunan City University, Yiyang, 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. 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].

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).

Natural, recommendation system service object of the heterogeneous social network is a user node, its recommended content mainly divided into two categories: referring users to other user node (e.g., recommend "lisi" to "zhangsan") and recommended topics node (e.g., recommend "the weather of Beijing" to "zhangsan"), that is recommendation which focus on each other relations and subscription relationship [7-11]. Need for mutual relationship and subscription recommendation based on different factors: recommended about the relation between mutual concern, through mutual friends, contacts, contacts and other real social information method usually achieve very good effect; to subscribe to the relationship, recommendations based on user's interest, this is the common recommendation, document recommended similar scenario in the recommendation system. Questions about such recommendations, adequate research

study is also carried out by scholar, collaborative filtering and content-based methods are put forward [12].

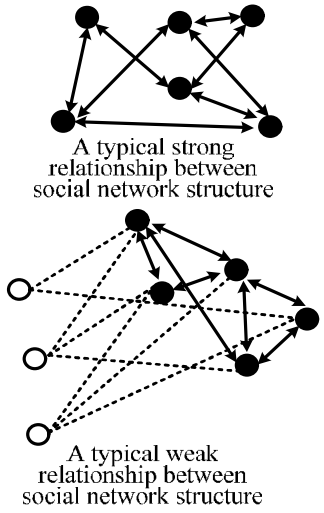


Figure 1. The typical social network structure between strong and weak relationship

Recommended problems on social networking, however, especially about the relationship between the subscription recommendations, different from the traditional recommendation system, one big challenge lies in its extreme data sparse. Mislove et al pointed out that the Internet social Network presents the characteristics of the Scale - free Network, a tiny amount of users have more connection, and a large number of users have only a small amount of connection relationship [13]. As a result of the existence of a large number of topics node, this phenomenon is more pronounced in the weak relationship social networks. 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.

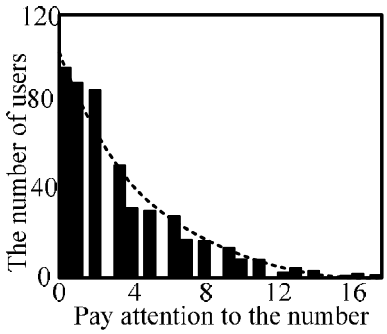


Figure 2. Sampling statistics of sina weibo user attention

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.

At present, there are many researches for recommendation system. In the recommendation algorithm, the main research directions include collaborative filtering recommendation, the recommendations based on the content, the clustering technology, Bayesian network, association rules, etc. Collaborative filtering algorithm is the most popular recommendation technology by far, which makes use of the similarity between user interest to recommend, does not depend on the actual content of the item, But it needs to users' preference information of items, usually in the form of evaluation or grading [14]. However, this classic collaborative filtering method cannot be directly applied to social network friends recommend, because in a social network, it doesn't have the concept of items and scoring. In addition, as a result of the social network data sparse, recommend effect of collaborative filtering algorithms is not good. Other research using the item's content to recommend, according to user's favorite items in the past, to recommend the similar items for users which they used to like. The method of content based on similarity can be well used in friends' recommend of the social network, use of natural language processing technology such as Piao Scott to deal with user's tweets, extract the user's interests, to recommend friends with similar interests. Sakaguchi et al proposed a system based on fuzzy sets of the concept, the system identification of a Twitter user's interest and recommend relevant friends, system uses concept dictionary based on fuzzy sets and word vector that represents the interests of individual Twitter users, using cosine vector measure similarity of users. However, recommendation based on content similarity is too specific, can only recommend the friends who has similar interested with users. A feature on Facebook is "people you may know", it is based on" Friend - of - Friend "algorithm to recommend. The algorithm idea is: if lot of friends of A is friends of B, so A's friend may also is B's. This algorithm can help users find not added strong relationship, economic sociologist Mark Glen, Victor puts forward: Relative to the strong relationship, weak relationship helps to transfer new information. According to the weak relational network like Twitter and sina weibo, the more valuable than strong ties to the recommendations of the weak relationship. Some researchers put algorithm based on content similarity combined with friends' relationship of social network. Hannon et put forward Twit - tomender system, according to the users tweet friends, fans and friends and fans tweet to user modeling, using TF-IDF of Lucene to measure the

weight of keywords. Kim Younghoon etc. using the probability model for collaborative filtering can recommend K friends and K tweets which users are most interested in. Probability model considering the relationship between the tweets information and users, the Kim et al., also propose a prediction algorithm to calculate the parameters of the probability model, and using MapReduce to deal with mass data [15].

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 Recommendain) 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.

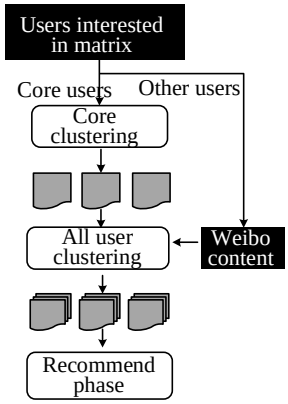


Figure 3. GCCR main processes

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

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.

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.

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 offline calculation generated clustering results and class interest vector, for any new users, only need to go on cluster calculation in and the recommendation process. 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

For N users, M theme, respectively for the user set $y=\{y_1,y_2,...,y_n\}$ with theme set $h=\{h_1,h_2,...h_m\}$. For each user UI, have corresponding interest vector $r_i=(k_1,k_2,...k_m)$, 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} **f** 0, represent the user UI's interest degree to the subject s j, if there is no subscription, the corresponding $k_{ij}=0$.

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

$$V=Y\cup h \tag{1}$$

E for collection of subscribing relationship:

$$E=\left\{e(y_i,h_j)\middle|y_i\in y,h_j\in h,k_{ij}\ \mathbf{f}\ 0\right\} \tag{2}$$

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 λ (usually 10%) is defined as the core. Then the core users set can be defined as

$$y'=\left\{y_i\middle|y_i\in y,des(y_i)\ \mathbf{f}\ \lambda\right\} \tag{3}$$

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

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:

$$y'=\overset{n}{\underset{i=1}{y}}c_i,\ c_i\neq\varnothing,and\ i\neq j,c_i\cap c_j=\varnothing \tag{4}$$

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

$$E_{s_j}(c_i)=\left\{y\middle|y\in c_i\ and\left(y,s_j\right)\in R\right\} \tag{5}$$

So participation q_{ij} meet:

$$q_{ij}=\frac{\left|q_{s_j}(c_i)\right|}{|c_i|}\ \mathbf{f}\ \sigma(\sigma\ \mathbf{f}\ 0,\text{is Intensity threshold}) \tag{6}$$

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

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

$$Amb_{ij}=\begin{cases}\left|c_i-q_{s_j}(c_i)\right|,q_{ij}\geq\sigma\\ \left|q_{s_j}(c_i)\right|,q_{ij}\geq\sigma\end{cases} \tag{7}$$

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

$$Amb_{ij}=\sum_{s_j\in s}Amb_{ij} \tag{8}$$

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

$$Amb=\log\left[\frac{\sum\limits_{c_i\in Clus}Amb_i}{|Clus|}\right] \tag{9}$$

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

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

$$ca_{ij}=\begin{cases}\dfrac{\sum\limits_{y_k\in c_i}a_{kj}}{|c_i|}, & q_{ij}\geq\sigma\\ 0, & q_{ij}\ \mathbf{p}\ \sigma\end{cases} \tag{10}$$

The c_i class interest vector on the theme set s is

$$cv_i=(ca_{i1},ca_{i2},...,ca_{im}) \tag{11}$$

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:

$$diff\left(c_i, c_j\right) = \frac{cv_i . cv_j}{\left|cv_i\right| . \left|cv_j\right|} \tag{12}$$

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

$$dvst = \frac{\sum_{c_i \in Clus, c_j \in Clus} diff\left(c_i, c_j\right)}{\left|Clus\right|} \tag{13}$$

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.

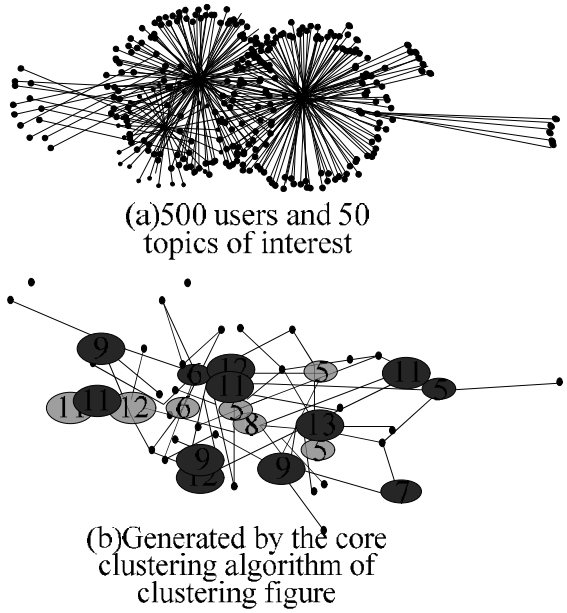


Figure 4. Interest figure contrast before and after clustering

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

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 UI 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:

$$cv_i = \left(ca_{i1}, ca_{i2}, ..., ca_{im}\right) \tag{14}$$

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.

$$dev_{i,j} = \sum_{c_i \in GClus} \frac{ca_{ki} - ca_{kj}}{\left|GClus\right|} \tag{15}$$

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

$$ca_j = \bar{ca}_i + \frac{\sum_{i=1}^{m} dev_{i,j}}{M-1} \tag{16}$$

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 appro-

priate 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. The Experiment and Analysis

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:

$$a = \frac{q(r|R)}{q(r)} = \frac{q(r).q(R|r)}{q(R).q(r)} = \frac{q(R|r)}{q(R)} \qquad (17)$$

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: 5 ~ 10 nodes adjacent or close to the user as a seed. 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. 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. 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.

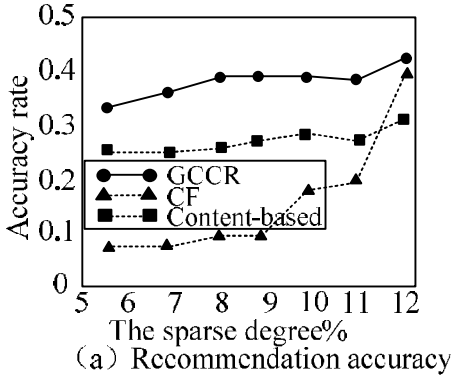
In this experiment, much attention are grabbed - interest groups of different matrix. The final experimental results are the average values of various experimental data. Of which, each set contains about 500 users, 50 theme and nearly 20000 weibo content. Experiment implementation by Python and Java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

Reference of algorithm is checked as: Collaborative Filtering recommendation algorithm based on Top - K similar (Collaborative-Filtering CF); 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.

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:

$$MAP = \left(\sum_{k=1}^{y} AP(k) \right) / Y \qquad (18)$$

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

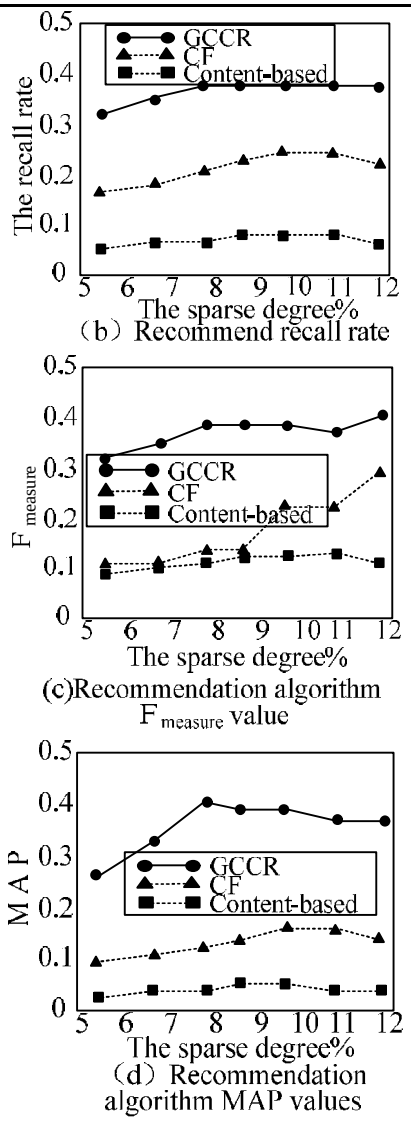


Figure 5. Recommended effect of three kinds of algorithm

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

5 . Conclusions

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.

Acknowledgment

This research work is supported by the Hunan Science and Technology Project (No. 2012FJ6011) and the Construct Program of the Key Discipline in Hunan Province, China.

References

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

[2] H. Z. Guo, Q. C. Chen, C. Xin, et al. "A Length-variable Feature Code Based Fuzzy Duplicates Elimination Approach for Large Scale Chinese WebPages", Journal of Software, 2012, 7(11), pp.262-269

[3] K. Atual, 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] Xiao-bo Li, Quan Zhou, "A lossless data hiding transmission method for satellite remote sensing image based on histogram modification," Journal of Astronautics, vol.34, no.5, pp. 686- 692, May,2013.

[5] 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.

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

[7] 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.

[8] 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

[9] 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.

[10] J.Tian, "Reversible watermarking using a difference expansion," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 8, pp. 890–896, Aug.2003