2014

ISSN : 0974 - 7435 Volume 10 Issue 22

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(22), 2014 [14002-14008]

Research on a dynamic social network recommendation approach

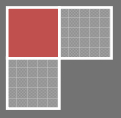
Hangjun Zhou[1, 2,*], Zhongli Liu[1], Sha Fu[1], Guang Sun[1], Zhanhong Xiang[1]
[1]Department of Information Management, Hunan University of Finance and Economics, Chang Sha, Hu Nan, 410205, (P.R.OF CHINA)
[2]School of Mechanical Engineering, Nanjing University of Science and Technology, Nan Jing, Jiang Su, 210094, (P.R.OF CHINA)
E-mail: zhjnudt@gmail.com

ABSTRACT

Owing to the rapid proliferation of Internet service technologies, the development of social network analysis is ever-increasingly important in recent years. As large datasets in social networks becomes available, recommendation plays a more and more important role in our daily lives. Recommendation approaches automatically prune large information to recommend the most relevant data to users by considering their preferences. Recent studies demonstrate that the efficiency of social networks could be exploited by improving the performance of recommendations. In this article, a novel recommendation approach is proposed to effectively extract dense subsets from sparse data set of micro blog social network, and cluster the whole user group into categories based on content similarity to produce better recommendation results. Through groups of reasonable experiment implementations with real data crawled from micro blog social network, the performances of this new proposed approach and other classical existing recommendation approaches are evaluated and compared by various measurable parameters. The experimental results demonstrate that the proposed approach could greatly improve the recommendation accuracy rate, recall rate and comprehensive measurable indexes when compared with other studied recommender algorithms. On the other hand, the computation overhead of the proposed approach is smaller than that of the other ones.

KEYWORDS

Social network; Recommendation algorithm; Collaborative filtering; Content similarity; Clustering.



INTRODUCTION

With the rapid development of Internet and mobile network technologies, the social network is prevailing more and more all over the world and has tremendously increasing number of favorite users in every year. The well-known representative social network instances are Twitter, Facebook, Sina Micro Blog, LinkedIn, and so on. Except strong relationship among people, the weak relationship is becoming high frequently occurred and widely reflected in those representative social network instances.

In weak relationship of social network, the user nodes, usually as message subscribers, give one-way attention to a large number of topic nodes. This kind of one-way subscription relationship is often based on user inclination of interests for different types of themes. Meanwhile, as news publishers, the theme nodes are subscripted by a large number of user nodes, and the relationship between initiative and two-way focus number is far less than the subscription number[1].

As we know, recommendation system of the heterogeneous social network is for user nodes, where the recommended content mainly is divided into two categories: recommended users and recommended subscription topic nodes. Many scholars have conducted researches from multiple aspects in this area. Among the existing recommendation algorithms, the main research directions include collaborative filtering recommendation, the recommendations based on the content, the clustering technology, Bayesian network, association rules, etc. Among them, collaborative filtering algorithm is the most popular recommendation technology by far, which uses the similarity between user interests to recommend. It merely needs user preference information of items usually in the form of evaluation or grading [2]. 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. Moreover, as a result of the social network sparse data, the recommending effect of collaborative filtering algorithms is not very effective on handling the sparse data of social network[3]. For the methods of content based on similarity, they can be well applied in friends' recommend of the social network by using natural language processing technology to decompose user tweets and extract the user interests for recommending friends with similar interests. Nevertheless, recommendation based on content similarity is too specific to recommend the friends who have different interests. Thus, further algorithms are proposed based on content similarity combined with friends relationship of social network. Twit-tomender system is given according to the users tweet friends, fans and user modeling, by using TF-IDF of Lucene to measure the weight of keywords[4]. In addition, the probability model for collaborative filtering is put forward to recommend K friends and K tweets that users are most interested in through considering the relationship between the tweets information and users. Furthermore, a prediction algorithm is given to calculate the parameters of the probability model, and using Map-Reduce to deal with mass data[5].

To solve the problem of effectively recommending topic nodes to users and process sparse data in heterogeneous social network classes, in this paper, a novel recommendation approach is proposed with Graph-Based Two-Stage Clustering algorithm. The approach can form a dense data subset from social network spars data set and use graph method to shift dense data into relative interests with similar core clustering. In following step, the approach identifies the content features of cluster seeds and forms a data set from whole user data with content similarity. After that, the topic theme is recommended to relative users based on the clustering results.

The rest of the paper is organized as follows. The problem modelling is defined in section 2. Then, the dynamic recommendation approach is described in section 3. In section 4, groups of experiments are implemented to evaluate the results. Conclusions are stated in section 5.

MODELLING THE PROBLEM

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 $k_{ij}\succ 0$, represent the user UI's interest degree to the subject *j*, if there is no subscription, the corresponding $k_{ij}=0$.

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:

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

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

$$y'=\left\{y_i\middle|y_i\in y,des\left(y_i\right)\succ\lambda\right\}$$

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

GRAPH-BASED TWO-STAGE CLUSTERING ALGORITHM

In this section, a dynamic Graph-Based Two-Stage Clustering (GBTSC) algorithm is proposed. Through the dense interest matrix m' constructing core interest figure G_m (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:

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

So participation q_{ij} meet:

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

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:

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

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.

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

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

The c_i class interest vector on the theme set s is

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

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:

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

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

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

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, who has published micro blog for Origin Tweet s_i, it's first to the original micro blog data preprocessing, such as removing the emoticons in micro blog and the "@" someone's information, and so on. To get users to post the plain text micro blog content Tweet s_i, 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 editing distance algorithm to calculate the characteristic vector similarity. Editing distance was originally used to measure the similarity between the strings. For 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 uses a single word in automatic segmentation of a sentence as the basic editor unit. In addition, the algorithm considers the edit operation cost and sentence length influence on similarity and puts forward the new block switching operation. 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 of 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 $GClus$.

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

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

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.

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

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

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

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, K value is given for the user having concerned topics or half of that number.

In the case of online recommendation, for the need to recommend the user, you can extract the content characteristic vector and use the process of classification in all user clustering processes. Each user would be assigned to the appropriate clustering by applying the predictive vector CV' of clustering to the recommendation. Apparently, the whole process, in addition to user classification process, needs real-time calculation and 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. Usually, the practical user clustering number is very small in the actual case, which also ensures the online recommendations efficiency of the algorithm. For the result of clustering and recommendation, when the number of new users increases in large scale, there would be the impact on the interest distribution adjustment.

EVALUATION ON THE EXPERIMENT RESULTS

In the experiments, users expect review rate is calculated to obtain the interest degree of user in the topic. In micro blog system, the meaning of the calculation could be understood as a user on particular topic comment content or forwarded by the potential probability. The index turns after with user evaluation of rate regulation. The approximate probability formula can use the following conditions:

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

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

Experimental data are grabbed through open platform of Sina micro blog 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 micro blog to simulate the formation process of online community based on weak relation step by step, with the opening of the seed users, and then obtain local samples of the heterogeneous social network features. The 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, to crawl users nodes adjacent with the current users; Or with the method of breadth first, to 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, to obtain detailed attention, forward and review data, then to calculate "user expectations review rate" according to the formula above, after that the end user - topic interest matrix is obtained.

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 micro blog 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 is applied with the user-based manner, and user similarity is computed using Pearson correlation coefficient, then finally the eventual recommendation results are recommended with the Top-K method. In the Content-based algorithm, the similarity of theme is calculated with the Chinese sentence similarity. In the experiment, half of theme is used as the training set, and the other half theme would be used as a test set.

In the followings, the accuracy and recall rate under the optimal parameters of several algorithms are evaluated. It is always set that the focus number of recommended number is equal to the number of training focus. For the CF and the Content-based algorithm, Top-k number is set for 10. Accuracy can be defined as the ratio of the number of recommendation hits and the total recommended amount; the recall rate can be expressed as ratio of recommendation hit number and the focus number in test set.

Figure 1 shows the recommend accuracy of the algorithm under the different levels of data sparseness. In the case of the extreme data sparseness (density is lower than 10%), the accuracy of the collaborative filtering algorithm is the worst, gradually ascending when the density is more than 10%. The recommended method based on content is not sensitive to sparse data, but accuracy is maintained at high level. With the comparison of CF and Content-based algorithms, the GBTSC algorithm could always maintain higher recommendation accuracy even under the condition where the data are very sparse. Apparently, in the scenario of this kind of micro blogging heterogeneous social network with the prevalence of sparse data, usually less than 10%, GBTSC has obvious enhancement compared with those traditional methods.

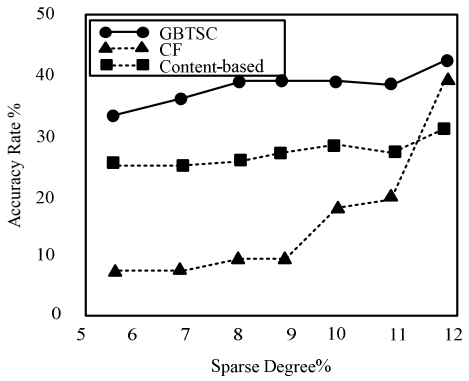


Figure 1 : Accuracy rates of recommendation algorithms

The comparison of the recall rates of different algorithms is demonstrated in Figure 2. In the implementation of the experiments, the recall rate of GBTSC algorithm keeps in a stable higher level than the other two algorithms with sparse data set. However, the CF and Content-based methods are influenced a lot by the data sparseness so as to have poor recommendation quantity and quality. Especially, the recall rate of content-based method is the lowest one in the figure because of its small number of generating recommendation results.

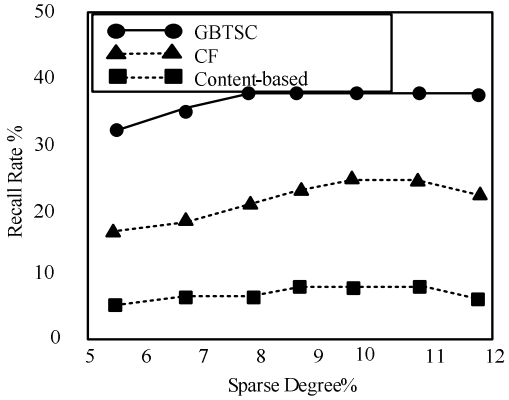


Figure 2 : Recall rates of recommendation algorithms

For the purpose of better comparing the recommendation effectiveness of these algorithms, the index $F_{measure}$ is introduced. The $F_{measure}$ is the Harmonic Mean for the recommendation accuracy and the corresponding recall rate. For this index comparison, the higher this value of a recommendation algorithm is, the better the comprehensive performance of the recommend algorithm is. The calculation of $F_{measure}$ could be defined as follows:

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

Figure 3 depicts the $F_{measure}$ values of different recommendation algorithms in different circumstances of the various data sparse degrees. It can be seen that the GBTSC algorithm performs better than the other two algorithms in every experimental conditions.

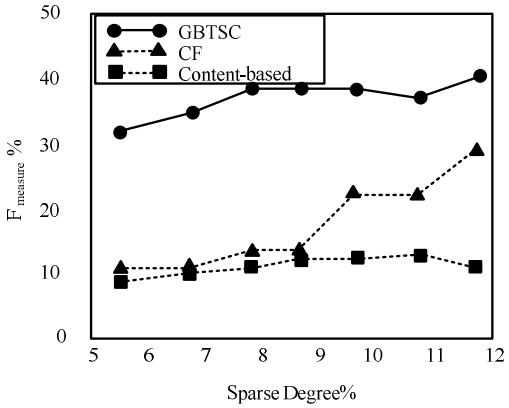


Figure 3 : The $F_{measure}$ of recommendation algorithms

On the other hand, another index *MAP* is introduced as Mean Average Precision for the recommendation results of a user group. *MAP* means average value of each user recommendation result. Similarly, the higher this value is, the better the overall effectiveness of a recommendation algorithm is. The calculation of *MAP* could be defined as follows:

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

The *MAP* value shows the average accuracy of the recommendation result from a user. Figure 4 gives the *MAP* values of different recommendation algorithms with cases of various data sparse degrees. The experimental results display that the *MAP* values of GBTSC algorithm are always higher than those of the other two algorithms with all data spare degrees.

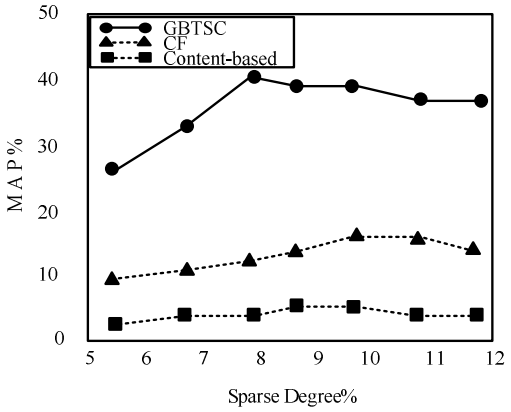


Figure 4 : The MAP of recommendation algorithms

CONCLUSIONS

As the social network, such as micro blog, obtains greatly increasing attention, it is necessary and urgent to put forward effective algorithms to solve the recommendation problems generated in the field of social network with intermittently emerging new situations. In this paper, a novel GBTSC recommendation approach is proposed, which could extract dense subsets from sparse data set of micro blog and cluster the whole user group into categories on the basis of content similarity for effective recommendation. The groups of experiments are implemented reasonably to evaluate the performances of different recommendation approaches. The experimental results demonstrate that the GBTSC recommendation approach outweighs the other ones with better accuracy rate and recall rate, and also performs better when referred to verify the comprehensive indexes of Harmonic Mean and Mean Average Precision.

ACKNOWLEDGMENT

This research work is supported by the Hunan Science and Technology Project (No. 2014GK3042) and Scientific Research Fund of Hunan Provincial Education Department (13C093). It is also supported by the Hunan Science and Technology Project (No. 2012FJ6011) and the Construct Program of the Key Discipline in Hunan Province, China.

REFERENCES

[1] Doan Anhai, Raghu Ramakrishnan, Alon Halevy; Crowdsourcing systems on the world-wide web, Communications of the ACM, **54(4)**, 86-96 **(2011)**.

[2] Sarwat Mohamed, Jie Bao, Ahmed Eldawy, Justin Levandoski, Amr Magdy, Mohamed Mokbel; Sindbad: A location-based social networking system, In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 649-652 **(2012)**.

[3] Bao Jie, Yu Zheng, Mohamed Mokbel; Location-based and preference-aware recommendation using sparse geo-social networking data, In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, 199-208 **(2012)**.

[4] Hannon John, Kevin McCarthy, James Lynch, Barry Smyth; Personalized and automatic social summarization of events in video, In Proceedings of the 16th International Conference on Intelligent User Interfaces, 335-338 **(2011)**.

[5] Hong Liangjie, Aziz Doumith, Brian Davison; Co-factorization machines: Modeling User Interests and Predicting Individual Decisions in Twitter. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 557-566 **(2013)**.