Intelligent Recommendation of Friends in Web Community

Jie Qiu [1], Jun Qiu [2]

[1] Financial Accounting Research & Development Center Chongqing University of Technology, Chongqing, China
[2] School of Science Chongqing Jiaotong University, Chongqing, China
qiujie_cn@163.com, Qiujun1@hotmail.com

Abstract - Web community is an imporatant internet application. An intelligent recommendation system has been designed in this paper to look for friends with same interests in web community. By this way, the web community can satisfy users' needs effectively.

Index Terms - web community; intelligent recommendation; data mining

1. Introduction

Web community is the popular internet application nowadays. It is an integrated communication virtual society with BBS, discussion group, chatting room, making friends and Blog, etc. Web community supplies rich interactive tools to users for their communication with friends. Web community grows rapidly recent years and it becomes one of the most popular applications with lots of users in internet.

In web community, people with same interests come together to some virtual theme society to discuss and share experiences and feeling. One reason why people are attracted by web community is to make friends. But most web communities lack function to find friends with same interests. So it is very important to build an intelligent friends' recommendation system which can make web community more friendly and human. On one famous web community in China, we use intelligent recommendation technique to design that system which greatly satisfies users' need to make friends.

2. Intelligent Recommendation Technique

Intelligent recommendation technique actively sends information needed to users. It can recommend accurate information such as commodities, friends and web contents to users in order to help those users find interesting information quickly. Intelligent recommendation technique changes the way how to get information and it transforms passivity into initiative. Intelligent recommendation technique makes use of different science fields synthetically such as data mining, database, statistics, artificial intelligence and social network analysis, etc. In order to supply intelligent recommendation of friends in web community, some kinds of technique have to be used, such as data mining, web data mining, database and web programming, etc.

A. Data Mining

Data mining makes use of modern analysis technique to find the relationship between model and data within mass data, in order to acquire effective, original, underlying meaning and understanding model [1]. Data mining is an interdisciplinary branch of science combining machine learning, statistics, neural net, database, pattern recognition, rough set and fuzzy mathematics, etc. Data classification and data clustering are frequently used methods in data mining.

The aim of data classification is to build classification function or model (usually called classifier) by use of training set after we have predefined classification. Data mining is to find the classification concept description which delegates the complete information of the whole set and describes the real essence of this classification. From this description, we can build the classifier which map unknown data into predefined set in order to classify unknown data. There are different techniques to achieve data classification such as Vector Space Model [2], K-Nearest Neighbor, naive Bayesian classifier, Decision Tree and neural network, etc.

Data clustering inscribes data in different sets according to their similarity. Data in same set are similar to each others and data in different sets are alien from each others. Unlike data classification, data clustering has no predefined classifications and it is typically a way of machine learning[3]. The techniques of data clustering include Hierarchical Clustering Methods which can be divided into decomposition method and agglomerative method according to clustering direction, Partitioning Clustering Methods whose typical examples include K-Means algorithm and K-Modoid method, Heuristic Clustering Methods, Density-based Clustering Methods which include DBSCAN clustering, OPTICS clustering and DENCLUE clustering, Grid-based Clustering methods whose typical algorithms include Sting algorithm, Clique algorithm and Wave-Cluster algorithm, Model-based Methods and Neural net-based Methods[4], etc.

B. Web Data Mining

The mining data could be structured data such as data in relational database or semi-structured data such as web pages, text, graph and image. The data mining in web community aims at semi-structured data, called web data mining. Web data mining makes use of data mining techniques to draw and conclude useful information from web pages and services. According to different mining objects, web data mining could be divided into three categories:Web Content Mining,Web Structure Mining,Web Usage Mining[5].

Compared with traditional data and data warehouse, information in web is unstructured or semi-structured, dynamic

© 2013. The authors - Published by Atlantis Press

and confused[6].It is difficult to draw needed information from web pages directly, so all those data in web have to be dealt with to become logical form or relational forms which are easily analyzed

3. System Design for Intelligent Recommendation of Friends

Intelligent recommendation of friends is the way to classify web community users into different sets according to their attribute and behavior analysis and then recommend relative friends. It can improve the success rate to find friends and avoid to looking for friends randomly.

All web users want a good interaction with web community except to be interested in the web contents. Making friends and communication is the most important part of interaction. Traditionally, it is almost a random behavior to make friends. For example, people sometimes add author into friend list when they are reading some papers or look for friends by searching online with some qualifications. Those traditional ways are intentional behavior to look for friends, but the friends found are not always matching. The system in this paper is try to recommend matching friends actively and improves the successful rates to look for friends with same interests to web users. It is a useful compliment to the traditional ways.

To achieve accurate intelligent recommendation, the first thing is to classify all data accurately. Manual work can achieve accurate classification, but the work load is too huge to afford. Another way is to make use of machine computing, but the result is not always accurate enough although lots of computing resources have been consumed. To make use of the advantages between human and computer, the system in this paper combines with the manual label supplementary classification and computer data mining to classify data. The principle for system design is to coordinate human being and computers.

A. Manual Classification and Label

Web community classifies text objects such as web pages by predefined directory structure. When web pages are submitted, they are posted into some classification, where those web pages have some common points. But directory classification itself could not satisfy the requirements for text classification and search, for instance, a web page could not be found in international relation channel when it is in military channel. To handle these problems, in the process to design web community, manual label has been used to enhance text objects classification. Manual label let web users play initiative and achieve preliminary classification for text objects and web users. To web pages, manual label uses tag classification technique in web 2.0 to describe text. Tag technique, namely social bookmark, is a flexible, accurate and user-defined classification technique. To describe different subjects, web users can freely define some key attributes (tags) which can classify those subjects with different profiles. For accurate classification of users, manual label combines with user predefined attributes setting and tag label. User predefined attributes include gender, age, education, career, interests and address. Users could be classified according to these attributes. Besides that, users themselves could add new tags from which users could be classified.

B. Application for Data Classification and Data Clustering

By use of directory classification, user attributes definition and tag technique, some objects classification could be accomplished, and these methods help to achieve intelligent recommendation. But for accurate recommendation, data mining has to be used for deep analysis. In data mining, data classification and data clustering are two major techniques. Data classification makes use of vector space model (VSM) to classify text objects and analyzes users' custom on web contents usage to find users interests. Data clustering uses improved aggregation hierarchy clustering method to group users with similar interests. The system will recommend users with friends according to user predefined attributes, tags classification and data mining in users' web custom log.

C. Steps to Achieve Intelligent Recommendation Technique

Accomplishment of intelligent recommendation needs background data analysis and mining. To avoid affecting response speed of web community, data analysis and mining is accomplished in a special server instead of web community server. This special server includes application server and database server. Application server deals with data needed for intelligent recommendation and database server saves original data and data analysis results. Original data come from web community. After analysis, some results will feed back to web community database. To achieve intelligent recommendation, seven steps are needed: data manual classification, data transforming, data treating and preprocessing, classification eigenvector reformulation, data mining, data feedback, intelligent recommendation.

Data manual classification: make uses of predefined classification, tags and users' attributes to manually classify data, and the rough results will be the initial start of intelligent recommendation.

Data transforming: newly added data such as users, texts, tags and logs will be transferred from web site online database to particular data analysis server.

Data treating and preprocessing: treat data such as texts and logs as weighted eigenvectors for further analysis and save these eigenvectors into database. Word segment should be finished for text before computing eigenvector.

Classification eigenvector reformulation: measure if the classification eigenvector and tag eigenvector need to be reformulated. If needed, calculate and revise eigenvector and make it more accurate for classification.

Data mining: treat text with classification and clustering, and treat users' custom logs with clustering. Save results into data analysis database.

Data feedback: send part of results to web community online database to supply intelligent recommendation.

Intelligent recommendation: according to users' relative setting, finish intelligent recommendation.

4.Key Section in the Design of Intelligent Recommendation System

To achieve accurate intelligent recommendation, users' data logs have to be analyzed. Data mining on users' data logs require proper classification for web community contents. There are many text contents in web community, so classification on text contents is the precondition for intelligent recommendation. After that, accurate data mining on usage custom for users can be obtained to confirm users need.

In the process to design intelligent recommendation system, key parts include text words segment, eigenvector calculation, text similarity calculation and users clustering.

A. Text Words Segment

The system uses VSM method for text classification. Before classification, text should be represented to eigenvector. To represent text as eigenvector, words segment should be finished at first. Because there is no blank between Chinese words, some ways are applied for machine word segment.

By now, different ways for Chinese word segment have been studied. There are two major methods among them. The first one is based on statistics; the second one is based on dictionary words matching. The advantages of dictionary words matching are accuracy, fast-speed and high efficiency; but this method is unable to recognize words not in dictionary. Word segment based on statistics does not need dictionary and it is achieved according to probability. The advantage of this method is that it can recognize any kinds of new words; the disadvantage is that it requires mass calculation and the segment results are not accurate.

Based on the advantages and disadvantages between dictionary and statistics, our system combined these two methods integrated.

Firstly, dictionary words matching method is used for first text scan. Then words in dictionary would be split and unidentified words would be marked.

Secondly, unidentified words would be divided into separate-word. Each separate-word would be calculated to get probability. Then separate-word would be composed by two, three, four until eight words. Each words composition would be analyzed according to statistics segment. After that new words would be obtained and added into dictionary for future use.

The third step is to identify "stop-word". The stop-word refers to some auxiliary words without real meaning. There are some Chinese auxiliary words which could be neglected without influence the original text. In our system, more than two hundred "stop-word" have been identified.

At last, words identified from text and its frequency of occurrences would be recorded.

B. Eigenvector Calculation

On the basis of accurate words segment, the further work is to calculate eigenvectors. Eigenvectors include text eigenvector and classification eigenvector. We use formula below:

$$V=V(P_1, W_1, P_2, W_2, \ldots; Pn, Wn) \quad (1)$$

Here, V represents eigenvector; P_i represents key word; W_i represents weight for key word. The accuracy of eigenvector is the basis of accurate text classification. To reduce the expense of eigenvector calculation, elements number in vector should not be too big. Less elements requires more accurate eigenvector otherwise eigenvector could not express text accurately. In our system, elements number in one vector is no more than 20. For better eigenvector chosen, the weight of word expression and its frequency from words segment result should be adjusted. We use inverse document frequency (IDF) adjusting key words to get rational word frequency. The word frequency weight for each word in dictionary will be adjusted by IDF. The formula for IDF is below:

$$IDFi=\log\ (T/Ni\ ) \quad (2)$$

T represents total number of sample text documents; Ni is equal to how many times the word i appears in sample text documents. If the word never appears in any sample documents, Ni is set to 1. 2000 documents collected from all documents classifications have been used as samples to calculate IDF in our practical web community system. After that, the word frequency weight would be adjusted according to IDF. The modified formula is:

$$W\ i=TFi*\ IDFi \quad (3)$$

Wi refers to modified word frequency weight; TFi refers to original word frequency weight after text words segment.

Then, the top 20 words with biggest word frequency weight will be chosen as text eigenvector.

To each classification predefined by our system, 10 documents will be manually chosen to draw eigenvector. At first each document will calculate its own eigenvector. Then the results from all documents will be gathered to choose top 20 words with biggest word frequency weight as eigenvector representing this classification.

As to user defined tag, it will be regularly calculated because it could not be calculated in advance. To each tag without eigenvector, 10 documents with this tag will be randomly chosen to calculate the tag eigenvector.

C. Text Similarity Calculation

After text eigenvector, predefined classification eigenvector and tag eigenvector have been calculated, the similarity of one given text could be compared to predefined classification and tag for further accurate classification. Our system uses vector space model (VSM) for text similarity calculation. Because document belongs to some classification when it is submitted in web community, it is not necessary to compare text with each predefined classification eigenvector and tag eigenvector to reduce calculation expense. What we need is to compare the eigenvector similarity among the text and the known classification and tag related to the text. Because the number of total classifications and tags in web community is huge, the operating efficiency would be improved thousands of times to compare text with its related

classification and tag only. We use cosine theorem to calculate similarity and the formula is below.

$$Cos(\theta) = (Vi*Vc) / (||Vi||*||Vc||) \quad (4)$$

Vi represents eigenvector of text i; Vc represents eigenvector of classification or tag; Vi*Vc means inner product between Vi and Vc; ||Vi|| means the size of Vi; ||Vc|| means the size of Vc. In this formula Vi and Vc adopt to its own word frequency weight, so it is 1*N or N*1 vector. If one word in Vi is not in Vc, it would be neglected in calculation; In other case, if one word in Vc is not in Vi, its word frequency weight in Vi would be set to 0.

After calculation, the similarity would be saved in related database for further use.

D. User Clustering

For user clustering, improved aggregation hierarchy clustering algorithm would be used. The traditional aggregation hierarchy clustering algorithm will calculate the similarity of any two users. To a web site with more then ten million users, comparing times would be combination number of n (n>ten million). If no measure is taken, the web server can not afford so many times of computing. So the conception of clustering subspace is imported. Clustering subspace is the collection of the sub sample which has similar character. The subspace only clusters users in the same space. By this way, the users can be effectively compared and it greatly improves the efficiency of aggregation hierarchy clustering. All users in the web site are divided into some small users' clusters by using clustering subspace. The number of users in each clustering subspace is less than 1000. The partition of clustering subspace is according to the characteristic combination labeled by people, such as user' occupation, address, hobby, age, education and tag and so on. The system can set different standard of characteristic combination of clustering subspace. The users who have the same characteristic combination are grouped in the same clustering subspace. If the number of users belonging to one clustering subspace is more than 1000, it will randomly choose 1000 users. For the users in the same clustering subspace, user's logs of web community contents usage can form users' eigenvector. The major elements of users' eigenvector are depended on the proportion of articles in each classification by user browsing and writing. The following formula is used to express user eigenvector.

$$V=V(P1, W1; P2, W2; \ldots; Pn, Wn) \quad (5)$$

V represents users' eigenvector; Pi represents article' classification; Wi represents the proportion of each classification articles occupied by all articles used by users. In our system, the weight of one article wrote by user is five times to article read by user. Then the eigenvector of users would be saved in eigenvector table. After all users' eigenvectors within same group have been accomplished, similarity of every two users could be calculated. The way to calculate similarity of users is similar to way of text. At last, using aggregation hierarchy clustering algorithm carries on users clustering: through establishing and renovating the user similarity coefficient matrix, unite the most similar two kinds, until all clustering objects are united to one. When it is finished, the result would be saved in database as the basis for recommending friends to users.

5. Conclusion

As a practical application, intelligent recommendation of friends has very important practical meaning to improve the competition of web community. This paper designs a feasible intelligent recommendation plan which has been applied in a practical web community. This system has just been applied a short time, so its practical effect needs more time to verify and some parts of this design needs more effort to optimize.

6. References

[1] Song Xu-dong, Zhang Tong-xue, Liu Xiao-bing, "Research of domain-oriented data mining system," *Application Research of Computers*, vol. 25, no. 5, pp. 1432-1433,2008

[2] Zhou Yan-tao, Tang Jian-bo, Wu Zheng-guo, "Method of multi-topic Web text classification based on VSM," *Application Research of Computers*, vol. 25, no. 1, pp. 142-144,2008

[3] He Ling,Wu Ling-da,Cai Yi-chao, "Survey of Clustering Algorithms in Data Mining," *Application Research of Computers*, vol. 24, no. 1, pp. 10-13,2007

[4] Yang Bo, Liu Da-You, Liu Jiming, Jin Di, Ma Hai-Bin, "Complex Network Clustering Algorithms," *Journal of Software*, vol. 20, no. 1, pp. 54-66,2009

[5] Dai Dong-bo, Yin Jian, "Web Personalized Recommendation Service Based on the Combination of Web Usage Mining and Web Content Mining," *Computer Engineering and Applications*, vol. 41, no. 18, pp. 162-165,2005

[6] Zhang Pei-yin, "Personalized Web Page Recommendation System Based on Web Content and Log Mining," *Application of Computer System*, no. 18, pp. 9-11,2008