TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.1, January 2014, pp. 353 ~ 360
DOI: http://dx.doi.org/10.11591/telkomnika.v12i1.4002

E-commerce Website Recommender System Based on Dissimilarity and Association Rule

Li Feng Zhang*, Shu Wen Yang, Ming Wang Zhang
Faculty of Geomatics, Lanzhou Jiaotong University
Gansu, Lanzhou, 730070, China
*Corresponding author, e-mail: zhanglf@mail.lzjtu.cn

Abstract

By analyzing the current electronic commerce recommendation algorithm analysis, put forward a kind to use dissimilarity clustering and association recommendation algorithm, the algorithm realized web website shopping user data clustering by use of the dissimilarity, and then use the association rules algorithm for clustering results of association recommendation, experiments show that the algorithm compared with traditional clustering association algorithm of iteration times decrease, improve operational efficiency, to prove the method by use of the actual users purchase the recommended, and evidence of the effectiveness of the algorithm in recommendation.

Keywords*: Dissimilarity, clustering, association rules, the electronic commerce recommendation*

1. Introduction

With the rapid development of electronic commerce, people's life and social environment have changed, the network users who purchase goods through the website of electronic business have become a universal phenomenon. As a business site operators, the need to keep browsing the web user, need to visitors into buyers, need to increase website sales varieties for customers multiple choices. Research on recommendation system for e-commerce website received more and more attention and research [1].

At present, as research in the field of mainstream is improved recommendation algorithm [2], mainly concentrated in the collaborative filtering algorithm, based on the content of the algorithm and the hybrid algorithm [3]. In this algorithm, hybrid algorithm application research is more, mainly concentrated in the association rules, genetic algorithm, neural network algorithm used in fusion [4].

The research content of this paper is the use of hybrid clustering and association rules in real-time performance, bad accuracy and recommendation of poor results, and presents a novel dissimilarity clustering and association rules algorithm. Most algorithms for clustering and association rules are used to cluster and association recommendation forms, but for sparse data and massive data, clustering and Association recommended joint algorithm in real-time and effective would decline, based on dissimilarity clustering and association algorithm the new algorithm, not only sparse matrix data can improve the recommendation accuracy, but also can improve the mass data recommend results, as the final recommendation to provide better service.

2. Web E-Commerce Recommendation System Overall Design

Although the research on recommender system is hot, but most researchers mainly focus on the recommendation algorithm, very few people can from the point of view of the system recommendation system discussed the construction [5]. In this paper, first from the construction of the overall design of the system of recommendation, in order to be updated in real time data processing, make electronic commerce website recommendation, this paper proposes the following recommendation system model.

Received June 24, 2013; Revised August 22, 2013; Accepted September 18, 2013

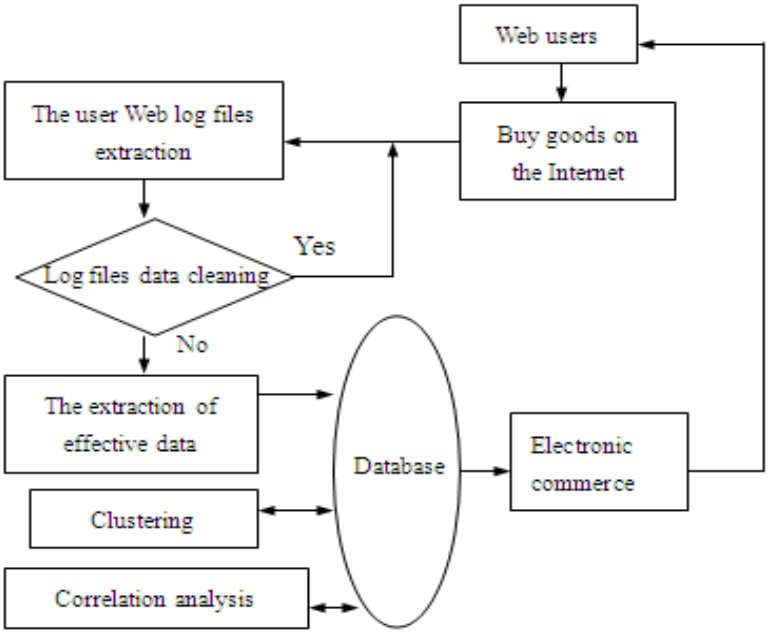

Figure 1. Web E- Commerce Recommendation System Model

The first part is the data collection. From the web site users purchase data log data is extracted, and the logs were effective data extraction, namely for data cleaning.

The second part is the data processing. After cleaning the effective data respectively dissimilarity clustering and association rules of calculation, and the use of database for data access and processing, by clustering and association recommendation, the data stored in the database.

The third part is the association recommendation. When a new user in Web data, it can be in your web site log data based on dissimilarity clustering and association algorithm, to purchase the recommended.

3. Web E- Commerce Website Access User Data Extraction and Cleansing

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2] [6]. The discussion can be made in several sub-chapters.

As the electronic commerce website accessible to the user, at the site of residence and to purchase items, or click browse merchandise will leave record, and stored in the web log server in [6] [7] [8].

According to the web log user purchase commodity record, set up user purchases incidence matrix, as shown in Table 1.

Where $U_i(i=1,2,\ldots,N)$ means users to access the site; $P_j(j=1,2,\ldots,m)$ expressed in electronic commerce website commodity number.

In order to facilitate the user purchase commodity clustering, firstly users to purchase goods in the matrix, it follows the rule of 1:

$$M=\begin{cases}0, & U_iP_j=0\\ 1 & U_iP_j=1\end{cases},(i=1,2,...n,\ j=1,2...m)$$

The shopping establishment matrix M:

$$M = \begin{bmatrix} 0 & 1 & 1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 1 \\ & & \cdots & & \\ 0 & 1 & 1 & \cdots & 0 \end{bmatrix}$$

Table 1. Electronic commerce website users purchase matrix table

Types of goods / User	P1	P2	P3……Pm
U1	0	1	1……0
U2	1	0	0……1
U3			……
……	……	……	……
Un	0	1	1……0

4. Recommendation Algorithm Description

4.1. Recommendation Algorithm Analysis

Recommend the result accuracy, data sparsely and computation complexity of these problems, leading to real-time recommendation is not easy to solve, and the recommended time is an important evaluation index, is the recommended algorithm improvement.

In this paper the dissimilarity dynamic clustering algorithm, which is used in common clustering algorithm and K- means clustering algorithm in an improvement. In the K mean clustering algorithm, the initial number of clusters is arbitrarily assigned; do not accurately reflect a sample set of accurate clustering number, at the same time, the algorithm iteration time's larger, cluster for a long time, not suitable for real-time recommendation.

While the dissimilarity degree clustering algorithm, for clustering the sample space, advanced dissimilarity cluster, and the cluster number as dynamic initial clustering number, then the sample space and the number of cluster are compared, thereby dividing the whole sample space, this algorithm has the advantage of reducing the clustering process iterative times, reduce the clustering time, increase the real-time recommendation.

Figure 1 shows, this recommendation system website, for new customers, only the purchase items associated with different results, so as to customers recommend products. The clustering results, using association rules algorithm for purchases and correlation analysis, to produce different patterns of association, and the results stored in a dedicated database.

4.2. Recommendation Algorithm

(1) based on the dissimilarity of the initial clustering

Dissimilarity is the characterization of object similarity degree, to a group of data dependence can usually use the dissimilarity matrix representation of object, the dissimilarity usually use the object I and object J dissimilarity quantification of $D(I, J)$ said, usually non - negative. Two objects nearer, the value is close to 0; the two object is different, the larger value of [9] [10], and have established:

$d(i, j) = d(j, i), d(i, i) = 0$

Based on the principle of calculating dissimilarity, matrix M turn to the dissimilarity matrix D by dissimilarity calculation [11].

$$D = \begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ & \cdots\cdots & & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

Define dissimilarity matrix is reflexive and symmetric, Where $d(I,I)=0,\ d(I,I)=0,\ D(I,J)$ $(\mathrm{i}=2,3,\ldots,N;\ \mathrm{j}=1,2,\ldots,N)$ said two element matrix D two element dissimilarity

$$d(i,\ j)=\frac{f_{00}+f_{11}}{f_{00}+f_{11}+f_{01}+f_{10}}$$

$F_{00}=x_0$ and y_0 the number of attributes; $F_{01}=x_0$ and Y_1 the number of attributes; $F_{10}=x_1$ and y_0 the number of attributes; $F_{11}=x_1$ and Y_1 the number of attributes.

The definition of $C_1,C_2,\ldots,C_n$ representation class, D_{pq} said C_p and C_q samples the dissimilarity

(2) The choice of matrix $D(t)$, when $t=0$ maximum element, assuming D_{pq} corresponding to class C_p and C_q merged into one category, denoted as $C_m=\{x\,|\,x\in C_p, X\in C_q\}$

(3) Calculate the new class and other classes of dissimilarity

$$D_{\mathrm{mk}}=\min_{i\in c_m,\ j\in c_k} d_{ij}=\min\{\min_{i\in c_p,\ j\in c_k}\mathrm{d}_{\mathrm{ij}},\ \min_{i\in c_q,\ j\in c_k}\mathrm{d}_{\mathrm{ij}}\}=\min\{\mathrm{D}_{\mathrm{pk}},\mathrm{D}_{\mathrm{qk}}\}$$

The $D(T)$ Q line and the P line P and Q columns are combined into a new column, new ranks should be C_m, obtained by matrix $D(t+1)$.

(4) If all the samples have been clustered into one class, then stop the algorithm, Otherwise $t=t+1$.

(5) Set difference threshold, choice clustering, resulting in data sets clustering center vector $C_1,C_2,\ldots,C_n$

Get cluster data set $G=\{C_1,C_2,\ldots,C_n\}$.

4.3. New Sample Clustering

When the sample space data elements increases, produce element and use of dissimilarity the clustering results are calculated, the specific steps are as follows:

(1) The sample space elements and the cluster center were distance calculation:

$$d(i,\ j)=\sqrt{(s_{i1}-c_{i1})^2+(s_{i2}-c_{i2})^2+\quad\cdots\quad(s_{in}-c_{in})^2}$$

Where $\mathrm{i}=1,2,3,\ldots,N$

(2) Arranged a distance threshold α, If $d_{\min}(i,\ j)<\alpha$, S_i belong to the same cluster, or generate new clustering center, and S_i will belong to the cluster of C_{n+1}.

(3) Repetition step 1 and step 2, until find all of the samples.

4.4. Recommendations Based on Association Rules

Through the results of clustering, using Boolean association rules frequent item sets algorithm, for each clustering results in association rule mining.

Association rule is of the form $A\geq B$ implication, in which $A\subset I, B\subset I$ I means sample set, and support is included in A and B affairs percentage; confidence is transaction contains A also contains B percentage, the following type [1] [12].

$$\sup\ port\,(A=>B)=P(A\cup B)=\frac{\sup\ port\_count\,(A\cup B)}{M},$$

$$confidence\ \ (A=>B)=P(B/A)=\frac{P(A\cup B)}{P(A)}=\frac{\sup\ port\_count\,(A\cup B)}{\sup\ port\_count\,(A)}$$

5. Experimental Analysis

5.1. Experiment Platform Construction

In the experiment, the experimental operating environment Inter (R) Core (TM) i5-2410M 2.6GHz 2GB memory, experimental software MATLAB2007b. Experiments using UCI KDD ARCHIVE website provides access log data, using the Web log users to access data to construct a matrix M, on two kinds of algorithm in the running time and accuracy comparison. The Web log user data set has a total of 92697 users, by eliminating the Web log user data length of more than 7 recorded sessions, select one of the 2000 user as experimental data, two kinds of algorithm evaluation.

5.2. Experiments and Verification

The user log data preprocessing, establish each shopping network in 7 following the experimental data source, respectively based on K-means clustering Association Rules Recommendation Algorithm Based on dissimilarity clustering and association rules recommendation algorithm, the experimental data analysis. Use of the running time of the algorithm and the recommendation algorithm accuracy was two kinds of algorithm to verify. Verification steps are as follows;

1. Two of the running time of the algorithm, using MATLAB2007b software to 100 and 200 and 400 and 800 and 1600 in five groups of data comparison test.

2. Accuracy and effectiveness verification, using two kinds of recommendation algorithm, the first of 2000 groups of data of 1800 groups of data are two kinds of clustering and association recommendation, then using the 200 groups after the purchase data algorithm, calculation of two kinds of algorithms for the 200 group of users and the accuracy of recommendation.

Recommendation accuracy verification calculation method is: if the user Ui purchase P1, respectively, based on dissimilarity clustering association rules and K-means clustering algorithm of association rules for users to recommend commodities $\{\mathrm{Mi}\}$ and $\{\mathrm{Ni}\}$, where $\{\mathrm{Mi}\}$ is based on dissimilarity clustering algorithm of association rules recommendation commodity set, where $\{\mathrm{Ni}\}$ is K-means clustering association rules the algorithm recommended commodity set. Then the recommended two sets and 200 groups of users in the purchase of goods and the purchase of P1 after other actual commodity sets are compared, to validate two kinds of algorithm accuracy.

5.3. Experimental Results and Analysis

The experimental results show that the K-means algorithm in clustering, the algorithm's time complexity bounds for $O(m*n*k*t)$, where t is the number of iterations, K clustering number, a sample of n points, m sample dimension. While the dissimilarity algorithm in clustering the time complexity of the algorithm the upper bound of $O(m*n*(m-1))$, where n is m number samples, sample dimension. From Figure 1 and Table 1 can be seen when the data set is less sample, and the sample with lower dimension, two kinds of algorithms in computing time is about the same, but with the increase in the number of sample data set, the algorithm K-means, the iteration number increased, the time complexity is increased, so the running time becomes large, and the dissimilarity algorithm in data quantity increases, the time complexity changes slowly, the running time of small changes, thus validating dissimilarity algorithm in the algorithm of real-time K-means clustering algorithm.

Table 1. Clustering for 8 of two kinds of algorithm operation schedule

Data set	100	200	400	800	1600
Dissimilarity degree clustering algorithm	0.001821	0.003718	0.007154	0.014426	0.028836
K-means clustering algorithm	0.003987	0.004421	0.010754	0.022453	0.051234

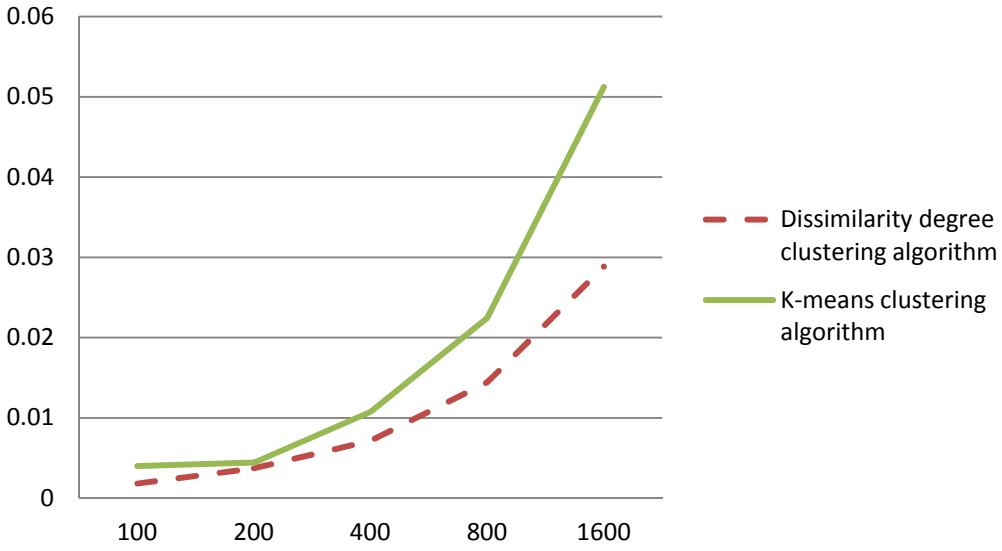

Figure 2. Clustering for 8 of two kinds of algorithm operation time diagram

Table 2. Two different clustering algorithms recommend accuracy table

The number of cluster	3	5	8	10	12	14	1800
Dissimilarity degree clustering algorithm	0.2	0.35	0.4	0.6	0.7	0.8	0.975
Difference threshold	0.8	0.7	0.6	0.5	0.4	0.3	0
K-means clustering algorithm	0.17	0.29	0.34	0.49	0.64	0.84	0.97
Class interval	1.56	1.17	0.78	0.61	0.49	0.32	0

In order to verify the two recommended algorithm validity and accuracy, using different dissimilarity threshold and the number of clusters of two algorithms are recommended validation. From Table 2 and Figure 3 can be seen in the data, the number of clusters and dissimilarity threshold is low, two algorithms of the recommendation accuracy rates are relatively low, and when the number of cluster and diversity is higher, the recommendation precision becomes larger. From Figure 3 we can seen, when the number of cluster becomes very large, recommendation accuracy to increase to more than 95%, but the data clustering and dissimilarity threshold setting is not desirable in a way. Because the WEB data mining, from large amount of data to find consistent with the data set characteristics clustering number, if the number of clusters is very large, close to the data set, lost the significance of data mining.

From Figure 3 we can see that, when the dissimilarity threshold in between 0.3 and 0.7, clustering number from 6 to13, clustering effect is good, can reflect the whole data set data characteristics, and provides a better basis for recommendation. At the same time, we can see that the two kinds of algorithm in recommendation accuracy rate have obvious difference, dissimilarity degree algorithm is better than the K-means algorithm.

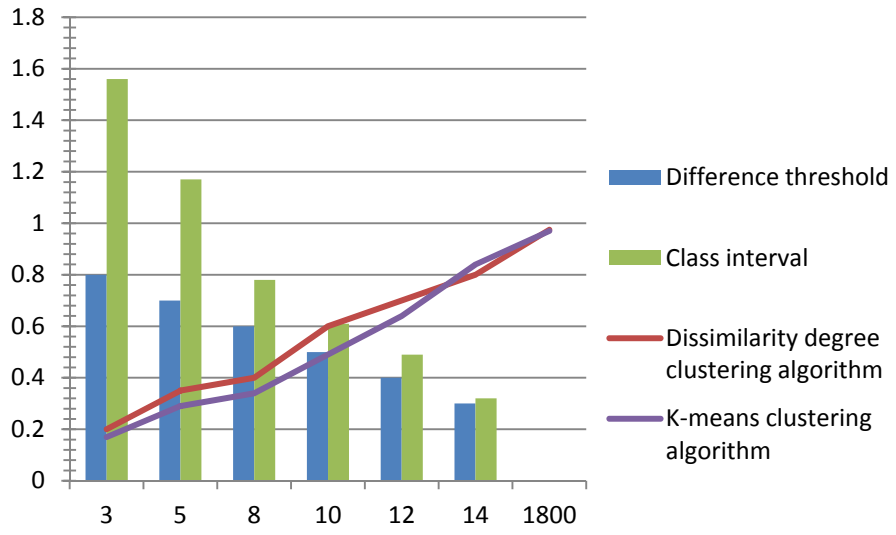

Figure 3. Two different clustering algorithms and the accuracy of recommendation

6. Conclusion

This article proposed a recommendation system model based on dissimilarity clustering and association rules, using dissimilarity algorithm reduces the computation complexity of the clustering process, reduce the clustering of the running time, and improves the real-time recommendation. At the same time, using the real data show that this algorithm is compared with k-means algorithm can improve the recommendation the validity and accuracy of this algorithm, but the WEB users log data as the limited data, users purchase behaviour is not more than 7 times, to purchase behaviour is greater than 7 times the purchase if the use of the algorithm, the is the next step of the research problem.

Acknowledgement

The authors also gratefully acknowledge support from Youth Foundation of Lanzhou Jiaotong University (No. 2013002).

References

[1] Shen Si. Based on association rules and Muti-Agent personalized information recommendation system. *Journal of Library and information work*. 2009; 53(4): 111-114.
[2] Wang Hongyu, Zhao Ying, Dang Yue Wu. Turn based algorithm of Party e-commerce recommendation system design and study. *Technology of Library and Information Science*. 2009; 174(1): 80-85.
[3] Wang Zhongzhuang, Deng Lundan, Shi Wenbing. Data mining technology in the electronic commerce recommendation system. *Microelectronics and computer*. 2007; 24(4): 197-199.
[4] Xin Li, Ting Li. E-commerce System Security Assessment based onBayesian Network Algorithm Research. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(1): 338-344.
[5] Wang Hongyu. Commerce recommendation system design. Anhui: University of Science & Technology China. 2007. Ph.D. thesis.
[6] Jin Wei, Sun Yan, Zhang Zhijun. Web information retrieval technology research and application of association rule mining algorithm. *Journal of intelligence*. 2007; 1: 39-42.
[7] Huang Yihong, Zhuang Yue-ting. Based on a novel competitive neural network WEB log mining. *Journal of computer research and development*. 2003; 40(5): 661-667.
[8] Guo Weiye, Zhao Xiaodan, Pang Yingzhi. Data mining clustering method based on SOM neural network. *Information Science*. 2009; 27(6): 874-876.
[9] Zhou Huan, Huang Liping. Neural network based on SOM means clustering algorithm. *Computer application*. 2009; 27(6): 51-52.

[10] Xiao Qiang, Qian Xiao-dong. *WEB user clustering Algorithm Analysis Based on the D-SOM*. 2011 2nd International Conference on management Science and Engineering (MSE2011). Chengdu, China. 2011; 1: 390-395.
[11] Wang Hui, Ming. Non-redundant Associations From the Frequent Concept Sets on FP-tree. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(7): 3604-3610.
[12] Zhao Mingqing, Jiang Changjun, Tao Shuping. Based on the equivalent dissimilarity degree matrix clustering. *Computer Science*. 2004; 31(7): 183-184.