Personalized Distance Learning System based on Sequence Analysis Algorithm

http://dx.doi.org/10.3991/ijoe.v11i7.4764

Ji-chun ZHAO [1, 2], Shi-hong LIU[2], Jun-feng ZHANG[1, 3]
[1] Beijing Academy of Agriculture and Forestry Sciences, Beijing, China
[2] Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing, China
[3] The Research Center of Beijing Engineering technology for Rural Remote Information Services, Beijing, China

***Abstract*—Personalized learning system can provide users with the most valuable learning resource to them through intelligent recommendation models and algorithms. This paper proposed the classical sequence analysis algorithms, and the Prefixspan algorithm is validated through distance learning platform data. In the event that the minimum support threshold is between 0.003 to 0.004%, test data shows that the performance of the algorithm's accuracy rate is relatively stable and the recommendation effect is satisfactory.**

***Index Terms*—Distance Learning; Sequence Analysis; Personalized Learning**

I. INTRODUCTION

The distance learning platform for farmers was set up by Beijing Academy of Agriculture and Forestry Sciences, which includes front broadcast platform, learning site, learning resource library and learning management system. What's more, the function of video-on-demand, live and expert lectures were existed in the platform. It is possible for farmers to learn in low cost, and they can get agriculture technology knowledge ASAP. At present, the number of registered users reached more than forty thousand, and the video teaching resources has reached more than 9,000 pieces . But it is very difficult for farmers to get their interested learning resources in the platform. So Personalized learning system was developed to solve this problem, which can analyze the user's behavior of individual , then provide them with useful information. The paper researched personalized learning system based on the massive user behavior data in the distance learning platform, and carried out the research of distance learning systems personalization algorithm for sequence analysis.

II. SEQUENTIAL PATTERN MINING

Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data samples where the values are delivered in a sequence. It is usually presumed that the values are discrete, thus time series mining is closely related, which usually considered a different activity. Sequential pattern mining is a special case of structured data mining.

Several key traditional computational problems are addressed in this field. These involve in building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, the problems of sequence mining can be classified as string mining, which is typically based on string processing algorithms and item set mining which is typically based on association rule learning.

III. PREFIXSPAN ALGORITHMS

1 Scan sequence databases, and generate all sequences mode of length 1.

2 Form the corresponding projection database according to the length of the sequence pattern 1.

3 Repeat the above steps on the corresponding projection database until it cannot produces a sequence mode of length 1 on the corresponding projection database

4.The projection for different databases were repeating the process until no new sequence of length 1 mode are set up. the basic principle of Prefixspan algorithm is shown in Figure 1.

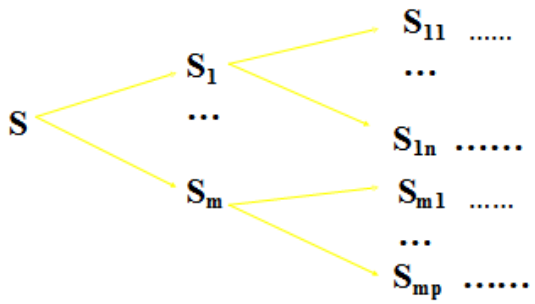

Figure 1. the basic principle of Prefixspan algorithm

IV. THE STEP OF SEQUENCE RECOMMENDATION ALGORITHM

1. The system generate a user sequence timely through processing video every day what users watched .

2. The same record user ID is combined, the specific time of occurrence of each video can be ignored, and generate a sequence database.

3. Call Prefixspan algorithm processing sequence database into a user maximal frequent sequences and obtain frequent sequences credibly and supports.

4.The length of user sequence is obtained by querying the average number of user watching video daily. User sequence database is obtained by intercepting the user specify length of the sequence.

5.Slect user sequence credibly . Sequence has been obtained using the user, and the user frequently sequence table is obtained by fuzzy query, which containing the current user sequences. The credibility value has chose in

these frequent sequences. N sequences are obtained, and N is the number of recommended video.

6. According to users credibility, frequent sequences is filtered . The highest sequence credibility and support video are priority selected to recommend.

V. THE TEST PROGRAMS AND PROCESSES

A. Test data selection

The users' learning records in the distance learning platform are as the test data. The number of the records is about 440,000 .And the user learning time length ,less than 10minutes , is filtered.

B. The test processes

(1) Select the training and test sets

The data set is divided into training and test sets, and the training set is about two-thirds of the record, which is used to generate frequent sequences. The rest ones are test data.

(2) The division of the test set

There are two methods to divide the test data set, as follows.

1) In the test set, each user browsed daily video records by using their own user ID, and being arranged in ascending order of time. This record is divided into two portions t1 and t2. T1 is account for 2/3 of the total browsing video records, which is used to generate a recommendation for video, t2 is account for 1/3 of the total browsing video records, which is used for the evaluation of the recommendation video result.

2) In the test set, each user browsed the daily video records by using their own user ID, and arranged in ascending order of time. This record is divided into two portions t1 and t2. T1 is account for 2/3 of the total browsing video records, which is used for generating a recommendation for video, and t2 is account for 1/3 of the total browsing video records, which is used for the evaluation of the recommendation video result.

2) In the test set, each user browsing the video records daily are identified based on the user ID, and each video is separated by a commas and made up of user sequence. This record is divided into two portions t1and t2.T1is account for 2/3 of the total browsing video records, which is used for generating a recommendation for video, and t2 is account for 1/3 of the total browsing video records, which is used for the evaluation of the recommendation video result.

C. Generating recommended video

(1)The recommended sequence algorithm program is run into training data set, frequent sequences of training is obtained.

(2)T1 data set is using the recommended sequence algorithm program for recommendation , the frequent sequence is obtained from training data set.

(3)The recommendation video set from training set is denoted by RS.

VI. TEST RESULTS AND ANALYSIS

A. Recommendation video number is fixed, The test result and analysis from different minimum support threshold

(1) Test Results

After the data set is divided, a total of training set records is 189,733. In the first method, there are 60989 records and 4988 users. In the second method, there are 61,626 records and 5757 users.

Table 1 shows the test result when recommendation video number is 5

TABLE I. THE TEST RESULT WHEN RECOMMENDATION VIDEO NUMBER IS 5

Recommendation video number is 5				
Minimum support	**Method 1 Precision**	**Method 1 coverage**	**Method 2 Precision**	**Method 2 coverage**
0.003%	33.83%	61.11%	35.46%	61.92%
0.004%	33.63%	51.44%	35.52%	53.53%
0.005%	35.36%	42.30%	39.62%	44.68%
0.006%	38.02%	35.91%	43.63%	38.58%
0.007%	38.74%	30.43%	44.50%	33.61%

The effect curve line of precision is shown in Figure 2, when recommendation video number is 5, and the effect curve line of coverage is shown in Figure 3.

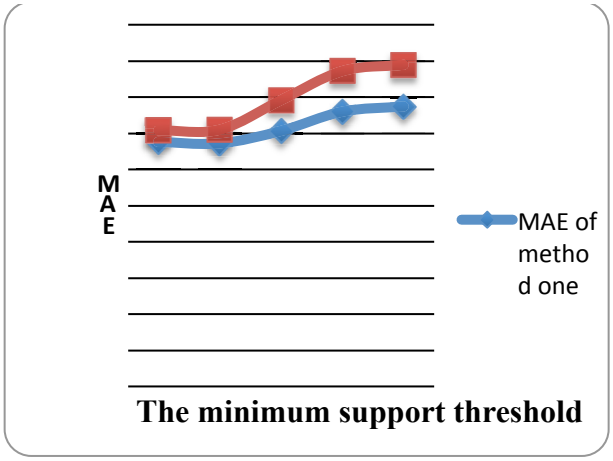

Figure 2. the effect curve line of precision, when recommendation video number is 5

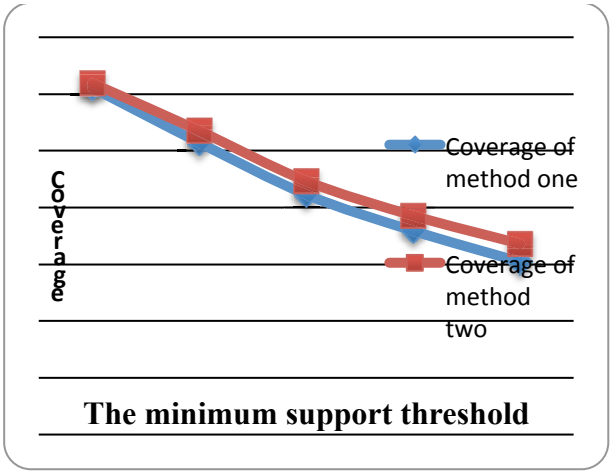

Figure 3. the effect curve line of coverage, when recommendation video number is 5

(2) Analysis on test result

When the recommendation number is the same, the minimum support threshold is changed, and the performance of accuracy and coverage in the method 1 is better than in the method 2.

B. The minimum support threshold is fixed, the test result and analysis under different recommendation video number

(1) Test result

After the data set is divided, the number of training set records is 189,733. In the first method, there are 60989 records and 4988 users. In the second method, there are 61,626 records and 5757 users.

The minimum support threshold is set up to 0.004%, and the recommendation video number is changed 1, 3, 5, 8,10.The precision and coverage is shown in Table 2.

Table 2 shows the test result, when minimum support threshold is 0.004%.

TABLE II. THE TEST RESULT WHEN MINIMUM SUPPORT THRESHOLD IS 0.004%.

Minimum support threshold is 0.004%				
Recommendation video number	**Method 1 Precision**	**Method 1 coverage**	**Method 2 Precision**	**Method 2 coverage**
1	13.76%	51.44%	16.29%	53.53%
3	26.27%	51.44%	27.94%	53.53%
5	33.63%	51.44%	35.52%	53.53%
8	40.69%	51.44%	44.03%	53.53%
10	43.69%	51.44%	47.92%	53.53%

when minimum support threshold is 0.004%,the effect curve line of precision shows in Figure 4, , while the effect curve line of coverage shows in Figure 5.

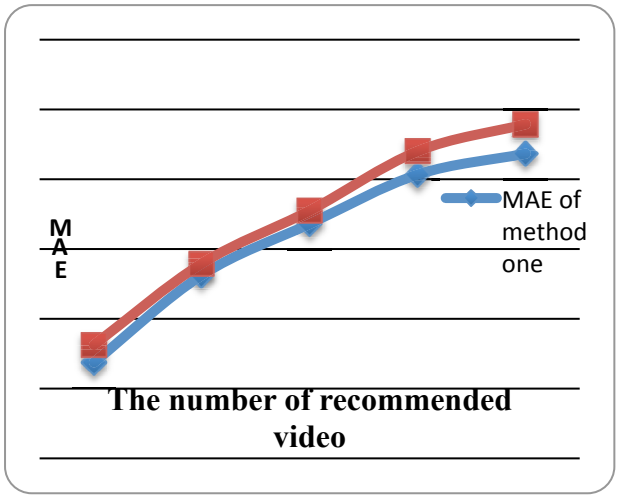

Figure 4. the effect curve line of precision, when minimum support threshold is 0.004%

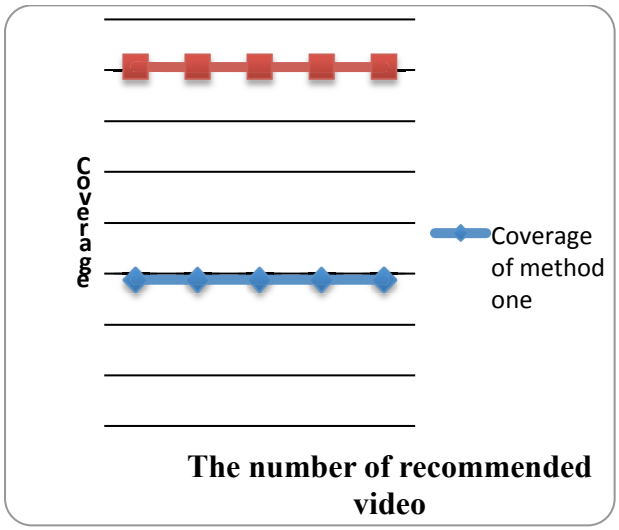

Figure 5. the effect curve line of coverage, when minimum support threshold is 0.004%

(2) The test result analysis

When minimum support threshold is set up, the recommendation number is changed, and the accuracy and coverage performance in the method 2 is better than in the method 1.

VII. CONCLUSION

The accuracy of the prefixspan algorithm will be better while increasing the minimum support threshold. The setup of frequent sequence data will decrease tremendously, if the minimum support threshold increased. So it is recommended that minimum support threshold should not be too large under the guarantee of coverage and frequent sequence data. when the minimum support threshold is between 0.003 to 0.004%, experimental results show that the performance of the algorithm's accuracy rate is relatively stable and the recommendation effect is satisfactory.

REFERENCES

[1] Mabroukeh, N. R.; Ezeife, C. I. (2010). "A taxonomy of sequential pattern mining algorithms". ACM Computing Surveys 43: 1. http://dx.doi.org/10.1145/1824795.1824798

[2] Abouelhoda, M.; Ghanem, M. (2010). "String Mining in Bioinformatics". In Gaber, M. M. Scientific Data Mining and Knowledge Discovery. Springer..

[3] George, A.; Binu, D. (2013). "An Approach to Products Placement in Supermarkets Using PrefixSpan Algorithm". Journal of King Saud University-Computer and Information Sciences 25 (1): 77–87 http://dx.doi.org/10.1016/j.jksuci.2012.07.001

[4] Linden G., Smith B., York J. Amazon.com recommendations: item-to-item collaborative filtering[J]. IEEE Internet Computing 2003, 7(1): 76-80. http://dx.doi.org/10.1109/MIC.2003.1167344

[5] Cheung K.W., Tsui K.C., Liu J.M. Extended latent class models for collaborative recommendation[J]. IEEE Transactions on Systems, Man and Cybernetics (Part A) 2004, 34 (1): 143-148. http://dx.doi.org/10.1109/TSMCA.2003.818877

[6] Balabanovic M. Shoham Y. Fab:content-based collaborative recommendation. Communications of the ACM 1997,40(3):66-72. http://dx.doi.org/10.1145/245108.245124

[7] Des. Keegan, The Future of Learning: From E-learning to M-learning[DB/OL] , http : //learning. ericsson . net/mlearnin92/project_one/thebook/chapterI.html.

[8] Paul Resnick, GroupLens: An Open Architecture for Collaborative Filtering of Netnews,1994

[9] Heckmann D., Schwartz T., Brandherm B. et al. GUMO-the general user model ontology [C]. In: International Conference on User Modeling, Edinburgh, UK, 2005: 28–432. http://dx.doi.org/10.1007/11527886_58

[10] http://en.wikipedia.org/wiki/Collaborative_filtering

[11] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Computer Supported Cooperative Work, pp175-186, Chapel Hill, North Carolina, 1994.

ACKNOWLEDGMENT

In this paper, the research was sponsored by National Science and Technology Support Program(Project No. 2014BAD10B02) , which is construction and application of provincial rural information service platform in developed area, and supported by Distance Education Innovation Team Project of Beijing Academy of Agriculture and Forestry Sciences.

AUTHORS

Ji-chun ZHAO is with Beijing Academy of Agriculture and Forestry Sciences, Beijing, China (e-mail: zhaojichun_0@ 163.com).

Shi-hong LIU is with Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing, China (e-mail: lius@mail.caas.net.cn).

Jun-feng ZHANG is with Beijing Academy of Agriculture and Forestry Sciences, Beijing, China (e-mail: zhangjf@ agri.ac.cn).

Submitted 25 May 2015. Published as resubmitted by the author 25 june 2015.