Application Research of Information Recommendation Based on Association Rules in Agricultural Information Network

RuPeng Luan, Qian Zhang*, JunFeng Zhang, Feng Yu, Xin Liu

Beijing Academy of Agriculture and Forestry Sciences, Institute of Agricultural Scientech Information

Abstract

In this paper, we used association recommendation to achieve page recommended for Beijing 12396 agricultural information networks, and improved accuracy, algorithm performance and execution efficiency of the recommendation. The contrast experiments proved that these improvements optimized the effect of recommendation from different aspects. So it can provide users with better agricultural information service.

Keywords: association rules; Apriori algorithm; web data mining; information recommendation

1. Introduction

Beijing 12396 Agricultural Information Network (12396 website) is established by Beijing municipal government, which provide agricultural science and technology information service for farmers. At present, 12396 website has a lot of historical counseling records and information data, but similar question will be repeatedly consulted and user's residence time is too short. There is an urgent need of optimization: improve user experience and the efficiency of getting useful information, and provide users with more personalized agricultural knowledge and information.

Associated recommendation is able to discover the relationship between affairs and recommended related information for users; it is often used for website optimization and transformation: A personalized association recommendation algorithm combine with website structure and content mining was proposed, which meet the individual needs of the network teaching system [1]; An integrated architecture of association recommendation was also proposed, which applied to academic resource retrieval system [2]; How to integrate a variety of related recommended in the digital library literature search platform was explored [3], thereby helping users to improve precision and recall rate and user's satisfaction.

In this paper, we first introduced web log mining and analyzes the status of 12396 website; then gave user behavior-based associated with recommended strategy and its on-line effects; again improved accuracy, algorithm performance and execution efficiency of the recommendation, and compare effect of recommended functional and it's improvement; finally, summarized the work done in this paper.

2. Association analysis and 12396 website analysis

2.1. Web log mining and association analysis

Web log mining refers to extract implicit valuable knowledge from access records of website which left by users access. It can find characteristics and law of users' access, provide basis for website construction and improvements, establish adaptive site and personalized service. The major steps of data mining are data processing, pattern recognition and pattern analysis, of which the first two are hot research at home and abroad [4].

Association analysis is a common technique for pattern recognition. It is described as if there is a certain correlation between two or more things then one of the things can be predicted by other things [5]. With purpose of excavating hidden relationship between data, it's frequently used Apriori and FP-Tree algorithm [6-7].

2.2. 12396 website analysis

By analyzing server logging of 12396 website, we can find: the user access mainly focused on 3rd grade page, which has 63.81% visitor and 92.32% user; user distribution of 3rd grade page is very uneven; dynamically generated page need to focus on maintenance and tracking.

The problems exist in the website mainly are: unpopular pages rarely appear in system log and have very weak links with other pages; there is no consistent standard of the website's URL naming rules; some URL path of server log record is incomplete.

3. The applications of behavior-based recommendation

The 12396 website applied web log mining which based on association analysis; it provided a list of association recommended 3rd grade page for the site. Associated pages were displayed in form of a content title hyperlink in target page, so as to convenient for users to access more resources, and provide users with a more comprehensive consulting solutions and information services.

3.1. Data Preprocessing

Data cleaning and path supplement: design a regular expression to extract effective access records of 3rd grade page from log files, and filter independent access records to reduce data size; study URL naming rules by renaming the URL; do path supplement for log record URL, so as to it has the complete information element; achieve session identification according to user residence time and it's access order. Below is an example of a figure area and related caption.

3.2. Behavior-based association recommendation

According to extracted web logging, association rules between 3rd grade pages were discovered by associated recommendation algorithm, and a recommendation list was generated. For each of 3rd grade page, use Apriori algorithm found many pages with strong association rules in log record; use function to get title for each URL; recommended pages in form of a hyperlink of content title will link to target page. We realize this function with C# programming.

3.3. Problem of this recommendation

Although recommendation had been basically achieved in 12396 website, but several problems still exist:

Cold start problem: for new pages, it cannot generate associated pages due to small amount of access;

Rare item problem: Volkswagen page will appear in most pages recommended

list, and uncommon page will never appear in other pages' recommended list;
Page type span is too large, so that pages content correlation was reduced;

4. Improvements of this association recommendation

According to the problems exist in 12396 website; we improved it from three aspects: recommended accuracy, algorithm performance and execution efficiency.

4.1. Improve the accuracy of the recommendation

Improve the accuracy of recommendation function by combining category recommendation and behavior recommendation.
(1) Category recommendation
Identify target 3rd grade page’s category through its URL, limited associated recommended in pages with same category. For each category has a list of categories heat list. For first-time visited pages its heat list is directly used as recommended list; for other pages, categories recommended list is used as a supplement of behavior associated recommended list.
(2) Combination strategies with behavior-based recommendation
Identify page category through analysis it’s URL and get heat of category. Generate behavior associated recommend URL list of target page. Target page's recommendation list is combined list.
This strategy does not exist repeat recommend or recommended; It also ensure that all pages have a recommended list and comprehensive user behavior with categories recommended; Higher related pages will be more forward in list order.

4.2. Raise recommendation algorithm performance

Apriori algorithm carries out mining association rules only in static transactional database [8], but lot of data mining work actual based on data witch changes with time. Thus it is important to research dynamic association recommendation algorithm and achieve dynamic recommended for individuals.
In order to solve the inherent "rare items" problem of Apriori, we propose a new association recommendation algorithm: [9-10] minimum confidence was made for each page, so that all rare pages can generate recommendation list; filter data set to page-related data sets, which greatly reduced algorithm size and improved algorithm efficiency. Due to time overhead of dynamic algorithm from unrelated visits influence, bigger amount of log data to demonstrate its superiority.

4.3. Improve execution efficiency of recommendation

In order to improve the efficiency of 12396 website recommendation and reduce flow occupied in site visits peak, a daily preprocessing scheme was designed: Build category recommendation and page recommendation pretreatment table, and update database at 12:00 every day.
Read category recommendation result, read or generate the behavior recommended results, combined the two results as this page recommendation list.
In the preparatory process one page can generate recommendation only one time every day, the other time just call for the records in database. Various category recommendations can be performed only one time every day, relevant page only calls the results of its pretreatment.

5. Effect contrast experiment

Behavior-based recommendation formally launched in 12396 website on February 4, 2012. A variety of improvements of recommendation were basically completed by mid to late 2012. The following content analysis the effects of recommendation with history access data in 12396 website.

5.1. Affect of recommendation function in 12396 website

Contrast log data without recommendation, with it and after its improvements. Each group has 61 days date. Various statistics data are as follows:

Tab. 1: Access statistics data in 12396 website

Variety of indicators	Nov.1-Dec.31	Mar.1-Apr.40	Aug.1-Sep.30
User visits	9392	54096	60785
Visits	4086	17763	14581
average pages viewed	2.347	3.089	4.226
average bounce rate	40.79%	31.35%	23.75%
Mean residence time	13.068%	17.418%	20.482%
new visitors proportion	89.03%	90.87%	90.63%
Return visit proportion	7.47%	7.40%	8.89%
longer stay user proportion	17.31%	25.53%	31.38%

Conclusion: The visits number of 12396 website greatly improved after association recommended functions on line, and nearly all types of statistical indicators have been enhanced. This shows that associated recommendation well enhance the user experience of 12396 website. The improved association recommended makes user visits, average pages viewed, and residence time indicators to further enhance. This shows that improved recommendation more appealing users, and it improve the stickiness of the site's users.

5.2. Accuracy contrast of recommendation

Combining behavior-based recommendation and category-based recommendation, and develop recommended strategies. Statistics 20 recommended list of 3rd grade page, the comparison of simple recommendation and comprehensive recommendation is shown in Table 2.

Tab. 2: Correlation of recommendation list

list number	simple recommendation	integrated recommendation
No.1	35%	80%
No.2	30%	70%
No.3	25%	75%
No.4	25%	60%
No.5	25%	60%
No.6	15%	65%
No.7	15%	50%
No.8	20%	50%
No.9	20%	45%
No.10	25%	40%

Conclusion: The performance of comprehensive recommendation is significantly higher than simple recommendation. The former take both 3rd grade page category and user behavior into account, which made recommended results more accurate, especially for rare pages and new pages.

5.3. Efficiency comparison of improved algorithm

The following are data sets experiment and contrast for two kinds of algorithms. The data sources from web access log of 12396 website. With software environment: Windows XP, Matlab R2011 and Microsoft SQL2005, and mainly use C# programming. The efficiency comparison of improved algorithm as show in Figr.1:

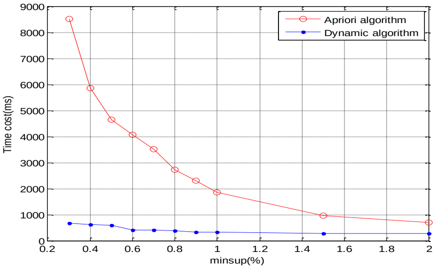

Fig. 1: Algorithm execution time under dif ferent minsup

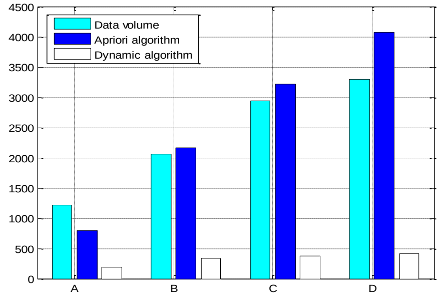

Fig. 2: Algorithm execution time under different data volume

Conclusion: The improved algorithm can improve the efficiency and accuracy of dynamic database of recommended. This is a new association rules discovery algorithm applies to web log mining; it enables real-time dynamic database mining. The algorithm can find minimum rules between rare pages for its support targeted pages. Experiments show that this algorithm have advantage performance on the timeliness, accuracy and time overhead than Apriori algorithm.

6. Conclusion

This paper uses association recommendation to achieve page content recommended of 12396 website and improved it from three aspects. Contrast experiments proved that the improvement optimized the effect of recommendation from different aspects. This study was supported by Youth Fund of Beijing Academy of Agriculture and Forestry Sciences Institute. The associated recommendation of 12396 websites should be considered follow-up Hadoop distributed processing, content-based association recommended, website log Statistical Analysis and other functions.

7. References

[1] LIU QingHua, JIANG Hua. Association Rule for Recommendation Approaches Based on Web Mining [J]. Communications Technology, 2008, 5: 86-93.

[2] Cen Yonghua, Den Sanhong, Wang Hao. Technologies and Applications of Association Recommendation in Academic Resource Retrieval Website [J]. Library and Information Service, 2009, 6:41-45, 99.

[3] Ji Yonghui. Research on Retrieval Results Clustering and Relevant Recommendation in Digital Library [J]. New Technology of Library and Information Service, 2008, 2:69-75.

[4] Li XueZhu, Zhou GuoXiang. Web Log Mining based on Improved Fuzzy Clustering Algorithm [J]. Computer Development & Applications, 2009, 22(4): 7-9, 20.

[5] YI Zhi, WANG Lin-lin, WANG Lian. Research on Web personalized recommendation based on correlation analysis of association rule [J]. Journal of Chongqing University of Posts and Telecommunications, 2007, 19(2): 234-237

[6] CAI Hao, JIA Yu-bo, HUANG Cheng-wei, HUANG Zhi-qiang. Improved method for session identification in web log mining [J]. Computer Engineering and Design, 2009, 30(6):1321-1323, 1390.

[7] Wu Zhixia. Research and Application of Web Log Mining [M]. Science Mosaic，2010,6:43-45.

[8] Bing Liu. WEB DATA MINING. Tsinghua University Press, 2009.

[9] RuPeng Luan, SuFen Sun, JunFeng Zhang. A Dynamic Improved Apriori Algorithm and Its Experiments in Web Log Mining[C]. The 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012, 6.