The Research on Personalized Recommendation Algorithm of Library Based on Big Data and Association Rules

He Ping*

The library of Chengdu University of Information Technology, The Chengdu University of Information Technology, Sichuan, Chengdu, China

Abstract: In order to provide better information consistent with their preferences features for users, personalized recommendation technology has become an important research field of digital libraries and get more and more attention from searchers. Among them, the large data mining and association rules-based personalized recommendation technology is the focus of research in the field of recommendation. In this paper, these two issues are studied. In order to increase the lending rate of collections, this paper use association rules analyzes for borrowing pattern mining, to obtain library users interests, to analyze different types of readers' purpose library collections, and automatically provide readers with other books related to such book. Through improved frequent pattern growth algorithm, combined with online recommended and offline recommendation method, achieved a more satisfactory recommendation results. Finally, taken experimental analysis and verification for these techniques studies, and future research were discussed.

Keywords: Association rules, big data, personalized recommendation algorithm, recommended library.

1. INTRODUCTION

Definition of digital libraries is still difficult to determine, there is no common definition of norms and foundation. Broadly speaking, the digital library is a computer that can be processed, a collection of information organized and orderly. It can also be seen as a storage that stores digital information. It uses digital technology to organize and manage information resources. Users can easily query the network efficiently retrieve information in order to obtain information services and their users to store and access information regardless of time and geographical constraints [1]. As a result of digital information carrier, with the help of networking and communications technology, the digital library to provide more comprehensive, more detailed, more efficient than the traditional library information services. Digital Library provides fast digital resources to create, indexing and retrieval capabilities to provide resource sharing worldwide.Once created digital resources that can be reused indefinitely without loss, store large amounts of data can really save a lot of manpower and resources, bring a large economic benefits.

Data volume growth reflects the people's information needs. The purpose of the digital library is the service of humanity's information needs. Digital Library for people to find the information needed to provide a convenience.Generally speaking, there are two ways to find the information you need: search and browse. Digital Library of the rapid growth of information makes information retrieval has become very difficult to return too many search results. For example, the query "library" in the ACM Digital Library, the result number 1807625, the query "computer", the number of results returned is 4,137,246, "personalized" query results for 4218 [2]. For a query user, you may browse the result is usually a number of links to ten.

In order to increase the lending rate of collections, this paper use association rules analyzes for borrowing pattern mining, to obtain library users interests, to analyze different types of readers' purpose library collections, and automatically provide readers with other books related to such book. Through improved frequent pattern growth algorithm, combined with online recommended and offline recommendation method, achieved a more satisfactory recommendation results. Finally, taken experimental analysis and verification for these techniques studies, and future research were discussed.

2. BIG DATA

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on" [3]. Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics [4], connectomics, complex physics simulations, and biological and environmental research.http://en.wikipedia.org/wiki/Big_data - cite_note-4.

Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks [5]. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×1018) of data were created [6]. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers" [7]. What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target. Thus, what is considered to be "Big" in one year will become ordinary in later years. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration" [8].

3. PERSONALIZED RECOMMENDATION

Digital Library is a huge library of digital resources, to provide users with quality service is the key to digital libraries. However, the increasing amount of digital resources, so that the digital library has a growing number of users, how massive digital resources to provide efficient access to the mass of users, but also an urgent problem.Existing digital library can not provide satisfactory personalized service for users, especially not needle users personalized information, such as the user's access history, knowledge of the user's level, education level, the user accesses the resource title, content, resources are frequency, impact factor access to resources, etc. for users to recommend information consistent with its preference for features [9]. Personalized recommendation technology digital library environment that take advantage of the user's personalized information, the user needs to transmit the information to the user, and the user to filter out irrelevant information, digital libraries based on user characteristics for different periods, different backgrounds, different purposes users provide different services to meet different needs, personalized information retrieval and information recommendation, improve service quality digital library, improve resource utilization.

Personalized recommendation is defined according to the characteristics of the user's interest, recommended to the user information about its interest. The principle is based on the user information to find its matching model, or find users with similar interests, and then browse through the recommended mutual information. Personalized recommendation technology is a key technology of personalized service.The technology has been in the fields of digital libraries, e-commerce, Web retrieval and other widely used. Different areas of personalized recommendation technology based on the user's personalized information, take the initiative to provide users with the resources of interest.

Let's introduce the research status personalized recommendations in accordance with the rules-based recommendations, based on the recommended content, collaborative filtering recommendation order.

3.1. Rule-Based Recommendation

The recommendation is based on the rule means the rules have been recommended to the user information generated way. This method is applied to more e-commerce network. According to users to browse and purchase log generation rules, by the rules have not figure out what the user to browse or purchase of interest, and then according to the rules of support for these contents sorted and presented to the user. Rule-based recommendation systems, such as IBM's WebSphere, BroadVision, ILOG [10], etc., allows the system administrator based on the static and dynamic characteristics of the user's property rules. The essence of a rule is an If-Then statement. Rules determine how to provide different services in different situations. Rules can be used to create static properties of the user; the user can use to build dynamic information. Advantage of rule-based recommendation system is simple and straightforward. Its disadvantage is difficult to ensure the quality of the rules, and cannot be dynamically updated. As the number of rules, the system will become increasingly difficult to manage.

3.2. Content-Based Recommendation

Recommended content-based is pointing to recommend information user by comparing the similarity approach of user and resource. Due to the need for content-based recommendation matching calculation, so the text field is applied more calculable, as recommended, newsgroups and other news recommendation browse pages.Filtering information based on content filtering systems use similar resources and user interest. Content-based filtering system advantage is simple and effective. Its disadvantage is that it is difficult to distinguish the quality of the resource content and style, and can not find new resources for interested users, and users can find resources have similar interests.

3.3. Collaborative Filtering Recommendation

Collaborative filtering recommendation is through the same or similar interest in the evaluation of resources users recommend to the user information. This method is recommended by the similarity to compare information between users, that is, first classify users. It applies to the text field of computable, but also applied to other areas, such as music, movies, books and so on. Advantages of collaborative filtering recommendation is to discover new information the user may be interested in, but there are no special requirements for the recommended object that can handle complex unstructured objects, such as music and movies. The disadvantage is that there are two difficulties: First, sparsity, that in the early use of the system, because the system has not been enough resources evaluation, the system difficult to use these evaluations to find similar users; the other is scalability, that is, with increase system users and resources, the system performance will be getting lower and lower.

Combining content-based and collaborative filtering methods help to overcome their disadvantages. Has a lot of work to merge the two technologies, in order to obtain a more precise recommendation technology called hybrid recommendation.Mixed recommendation refers both resources by comparing the similarity with each user model, make recommendations based on the content, but also by a similar interest user groups, conduct collaborative filtering recommendation one way. This method has a good recommendation performance. It also uses the recommended content-based and collaborative filtering recommendation two technologies. The combination of these two techniques can overcome some of their disadvantages. In order to overcome the sparsity of collaborative filtering problem, you can use the user to browse through the contents of the resource evaluation of prospective users of other resources. This can increase the density of resource evaluation. Then recommend the use of these evaluations based on content, thus improving the performance of collaborative filtering.

4. ASSOCIATION RULE

Association rule mining is one of the most active research method in data mining. it can be used to find links between things, the first time it used is to find the relationship between supermarkets transaction databases between different commodities.

There is a Wal-Mart anecdote. Wal-Mart analyzed more than a year for the data warehouse of the original transaction data in detail, and finally found that the commodity purchased together with diapers the largest number is turned out to be beer.With data warehouse and association rules, discovered hidden behind the fact: the United States, women are often asked her husband after work to buy diapers for their children, while 30% to 40% of her husband but also the way to buy their own beer drinkers after buying diapers.Based on this discovery, Wal-Mart shelves adjust the position of the diapers and beer on sale together, greatly increasing the sales. Here borrow an example to introduce the lead mining association rules [11, 12].

Table 1. Supermarket transaction data

Transaction Number Tid	Goods Customers Purchased
T1	bread, cream, milk, tea
T2	bread, tea
T3	bread, cream, milk
T4	beer, milk, tea
T5	cake, milk
T6	bread, cream, milk, tea
T7	bread, tea
T8	milk, tea
T9	bread, cake, milk
T10	bread, milk, tea

Definition 1: Assuming that I= {i1, i2,…, im}is the set of m different items, and each ik is called an item. Set I of the project called item-sets. The number of its elements is called the length of the item set. Item-set length k is called k- item-sets. Each item of Cited example is a project. Item-set is I = {bread, beer, cake, cream, milk, tea}. Its length is 6.

Definition 2: each transaction T is a subset of the set of items I in. Each corresponding to a transaction has a unique transaction identification number, denoted by TID. All transaction constitutes a transaction database D. | D |equals to the number of transactions D. Examples cited included 10 transactions, therefore $|D|= 10$.

Definition 3: For item-set X, set that count $(x \subseteq T)$ is the number of transactions in D contain X set of transactions. Then support degree of Item-set X is

$$|D|\, support\left(X\right) = count(X \subseteq T) \,/\, \left|D\right| \tag{1}$$

In the cited example, $X = \left\{bread, milk\right\}$ appear in T1, T2, T5, T9 and T10, so the support is 0.5.

Definition 4: Minimum support is the minimum support threshold of item-sets, denoted SUPmin, representing the lowest importance of interest to users of association rules. The item-sets which support is not less than SUPmin called the frequent sets. Length k frequent set is called k- frequent sets. If you set SUPmin 0.3, support of cited examples $\left\{bread,\ milk\right\}$ is 0.5, so it is the 2-frequent sets.

Definition 5: Association rule is one implication:

$$R \quad oX \Rightarrow Y00 \tag{2}$$

Of which, $X \subset I,\ \ Y \subset I$. Said the item-set X in a transaction occurs, resulting in Y will appear with a certain probability. Users concerned about the association rules, you can use two standards: support and confidence.

Definition 6: Association rule R is the transaction set, the transaction also includes the number of X and Y with | D | ratio, that's

$$support(X \Rightarrow Y) = count(X \cup Y) \,/\, \left|D\right| \tag{3}$$

Support Degree reflects the X, Y simultaneous occurrence probability. Support equal support association rules frequent sets.

Definition 7: For association rule R, credibility is the ratio of X and Y contains transaction number and the number of transactions containing X.

$$confidence(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support\left(X\right)} \tag{4}$$

If the deal reflects confidence contains X, Y, then the probability of containing the transaction. In general, only high support and confidence of association rules is the one user interested.

Definition 8: Association rules set minimum support and minimum confidence are SUPmin and CONFmin. Rule R's support and confidence, and not less than SUPmin CONFmin, then known as strong association rules.The purpose of association rule mining is to find strong association rules to guide business decisions.

These eight definitions contain several important basic concepts related to association rules. Association rule mining has two main problems: the transaction database to find all frequent item sets is greater than or equal to a user-specified minimum support;Use frequent itemsets to generate the desired association rules, based on user-defined minimum confidence screened strong association rules. The researchers are mainly studies the first question. Find frequent set is more difficult. And with frequent re-generate the strong association rules set relatively easy. Algorithm generates frequent sets have more classic Apriori algorithm.

5. EXPERIMENTS AND RESULTS

In order to accurately represent the user's interest, the need for each user to establish a user interest profile (user profile). Description of the user's interest and can be said to be the user's personalized information. In the use of the right of association rule mining algorithms, association rules mined between library users, and how accurately represent user interest. Suppose S is all the rules for users, $S=\{r_1, r_2, r_3,. . .r_n\}$. r_i is a rule mined. Each rule again includes pre-entry and post-entry. $r_i = \{b_i, a_i\}$, b_i expresses the preceding paragraph of the rule r_i, a_i represents the post-entry rule r_i.

Definition 9: user interest model is described as a collection of consisting of triples, namely:

$$U_i = \{(b_1, p_1, w_1), (b_2, p_2, w_2), . . . , (b_n, p_n, w_n)\} \qquad (5)$$

U_i is the i-th user. b_i is the first item of the user i rule. p_1 is in the S ruleset for the user, the page after all entries have a front page item is b, consisting of a collection of rules. w_1 is the weight.

Based on the above ideas, specific steps recommended in this paper are as follows:

Step 1: Get the user's current library browsing mode that get the current page and the top n-1 books the users accessed $p_1, p_2, . . .p_i$, constitute the preceding paragraph of rules.

Step 2: Find a match in the rule set in the preceding paragraph and the current access patterns of the user, that's the k rules:

$$\{p_1, p_2, . . .p_i\} = \{p_{a1}, p_{a2}, . . .p_{ai}\} \qquad (6)$$

After the item rule corresponding to the page p is the recommended page in the current page;

Step 3: If k is not greater than m, then the term k after rule corresponding to all of the pages recommended; if k is greater than m, the page is calculated after k items to recommend rule weights, and then follow from large small order to select the top m page recommendation;

Step 4: In the current window, in addition to display the user's current access to books, but also shows the recommended book topics that are "hot key", when clicked, you can link to the corresponding book URL. The URL corresponds to the book as the next window of the current books.

In order to make the results more reliable comparison, the present study, the July 2011 the first week of school library journal. After processing, the resulting dataset session record total of 3419, including 142 books.

In setting the support threshold, Figs. (**1-3**) were given the recommended coverage, accuracy and comprehensive measure of the situation with the support of the change where the length of foregoing paragraph of recommended rules are 1,2,3,4 and 5.

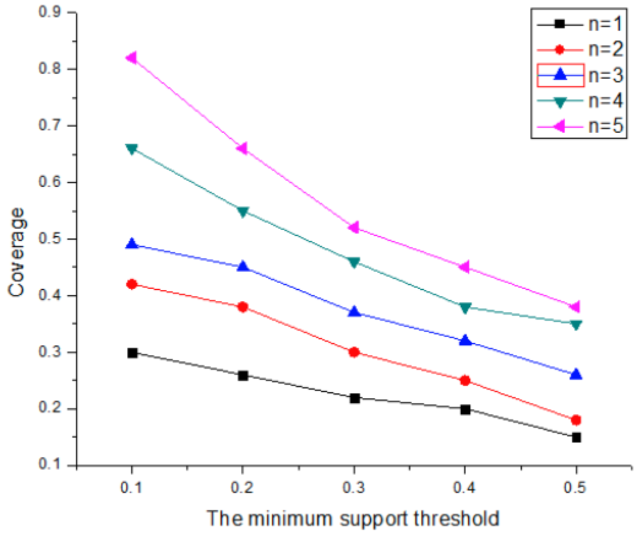

Fig. (1). The coverage with support threshold changes different numbers of rules foregoing paragraph.

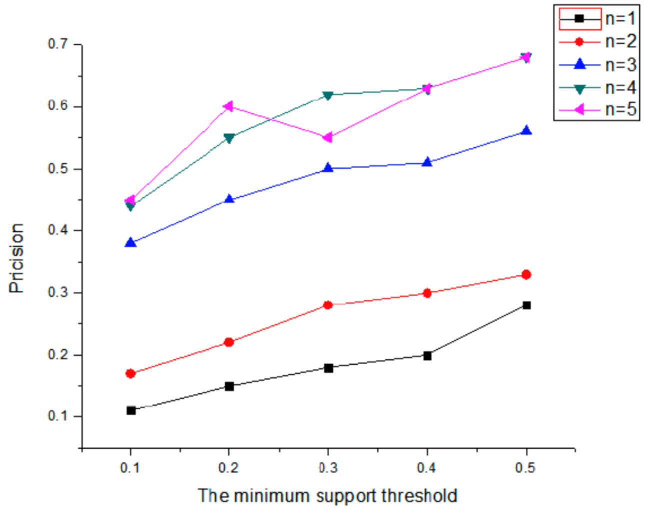

Fig. (2). The precision with support threshold changes of different numbers of rules foregoing paragraph.

The Fig. (**1**) shows, as recommended by the rule in the foregoing paragraph n = 1 has high coverage. And seen from Fig. (**2**), the recommended rules foregoing paragraph n = 5 has a high accuracy rate. However, Fig. (**1**) shows, when n = 4, the coverage is the lowest, while Fig. (**3**) appear when n = 5, the composite measure has a large advantage. Therefore considers longer rule in Foregoing paragraph has the comprehensive advantages recommendation.

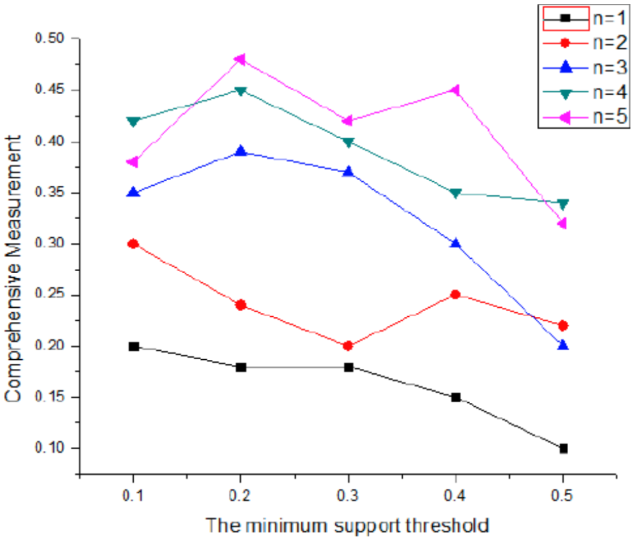

Fig. (3). The changes of comprehensive measurement with support threshold of different numbers of rules foregoing paragraph.

CONCLUSION

The Internet has made the current digital library information resources more abundant, but with the expansion of information, there has been "information overload" and "Information isotropic" phenomenon. In order to provide better information consistent with their preferences features for users, personalized recommendation technology has become an important research field of digital libraries and get more and more attention from searchers. Among them, the large data mining and association rules-based personalized recommendation technology is the focus of research in the field of recommendation. In this paper, these two issues are studied.

First, in terms of collaborative filtering technology to analyze the existing problems, pointing out the differences with the number of users of digital library systems and the increasing number of resource library and professional background of the user, resulting in extreme scoring matrix on the entire project space sparse, making the recommendation to the user's interest in the results of large deviations.

Secondly, In order to increase the lending rate of collections, this paper use association rules analyzes for borrowing pattern mining, to obtain library users interests, to analyze different types of readers purpose library collections, and automatically provide readers with other books related to such book. However, direct use of frequent pattern growth algorithm will produce very substantial number of association rules. Themes were recommended to the user based on these rules, will give the system a great burden, and may cause a lot of repetition of recommendation, generated a lot of computing redundancy. Through improved frequent pattern growth algorithm, combined with online recommended and offline recommendation method, achieved a more satisfactory recommendation results.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] L. Fernandez, J. A. Sanchez, and A. Garcia, "Mibiblio: personal spaces in a digital library universe", In: *ACM DL*, pp. 232-233, 2010.

[2] P.W. Foltz, and S.T. Dumais, "Personalized information delivery: an analysis of information filtering methods", *Communications of ACM*, vol. 35, no. 12, pp. 51-60, 2003.

[3] E. A. Fox, and M. Gary, "Web usage mining:discovery and application of usage patterns from web data", *Communications of the ACM*, vol. 44, no. 5, pp. 36-49, 2011.

[4] M. Elena, "A personalized collaborative digital library environment", In: *5th International Conference on Asian Digital Libraries*, pp. 262-274, 2012.

[5] S. Dumais, and H. Chen, "Hierarchical classification of web content",In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, ACM Press, pp. 256-263, 2004.

[6] D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", in: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Communications of ACM*, vol. 35, no. 12, pp. 71-89, 2010.

[7] J. Callan., "Distributed information retrieval", *Advances in Information Retrieval,* 157-20 self, 2000.

[8] M. Balabanovic, "An adaptive web page recommendation service", *Proceedings of the 1st International Conference on Autonomous Agents*, New York, ACM Press, pp. 378-385, 2008.

[9] H. Sakagami, T. Kamba, and A. Sugiura, "Effective Personalization of Push-type Systems visualizing Information Freshness", In: *Proceedings of the 7th International Conference on World Wide Web, Elsevier Science Publishers B.V*, vol. 30, no. 7, pp. 53-63, 2001.

[10] K. L. Wu, C. C. Aggarwal, and P. S. Yu, "Personalization with Dynamic Profiler", In: *Proceedings of the 3rd International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems, CA, IEEE CS Press*, pp. 12- 20, 2014.

[11] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on web usage mining", *Communications of the ACM*, vol. 43, no. 8, pp. 142-151, 2009.

[12] K. D. Bollacker, S. Lawrence, and C. L. Giles, " Discovering relevant scientific literature on the web", *IEEE Intelligent Systems*, vol. 15, no. 2, pp. 42-47, 2005.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015