International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.257-264
http://dx.doi.org/10.14257/ijdta.2016.9.6.26

Sequential Association Rules Based on Apriori Algorithm Applied in Personal Recommendation

Wang Yonggang

School of information science and technology, Zhengzhou Normal University, China
wyghaha@163.com

Abstract

The association rules is a data mining technology that is to find the regularity from the massive data. Based on the imperfection of the traditional association rule mining algorithm ,and in consider of the electronic commerce time latitude, this text proposed temporal association rules algorithm, which is also applied to the field of electronic commerce. Through analyzing the record of consumer of some category customers purchased during the last month from a website, dig out the regularity of customers choosing their products, to help sellers recommend products the users may be interested in, and developing a reasonable marketing strategy.

Keywords*: Sequential association rules data mining personalized recommendation e-commerce Apriority algorithm*

1. Introduction

In recent years, Internet business has developed rapidly. At the same time, in order to attract more consumers, business platform provides all sorts of goods, coupled with producing a large amount of transaction data on a daily basis, which makes the problem of information overload is becoming increasingly serious, therefore has caused a bad impact on the user's availability efficiency. Visibly, it is practically significant to analyze and mine the massive online transaction data and user's browsing behavior.

Data mining is a kind of new data processing technology; it is an effective tool for data analysis to combine the complex algorithm which deals with the massive data with the traditional data analysis methods. Apply the data mining technology to the field of electronic commerce, enterprises can better manage the relationship between them and the customers. They can optimize the website design, and recommend the goods the users interested in to stimulate consumption. [1]

In the current fierce competition of the electricity business, a good electronic business recommender system can give users constantly surprise and interest. Through turning the website visitors into buyers to increase the amount of users and existing user's activity and loyalty， sellers could improve the conversion rate of the electronic commerce website. The powerful data processing ability of computer and the rapid popularization of electronic marketing mode lead the personalized recommendation system of commodities in the field of electronic commerce to the possession of a good development and application prospect. But with the development of electronic commerce, increasing data, and accompanied by a series of invalid random repetitive data makes the recommendation system faces the problem of recommendation quality, including the lack of diversity, individual defects and lower accuracy, etc. The association rules mining algorithm is an important part of recommendation algorithm, It can dig out the relationship between the users and commodities through the analysis of different transaction in the database to help businesses to improve the service quality of an enterprise, and to help improve the consumer experience of the users, reducing cost and promoting the consumption of the users.

The Apriori algorithm is the most commonly used one in association rules model, compared with other algorithms, it owns high reliability and efficiency of the algorithm. The

ISSN: 2005-4270 IJDTA

related mining towards the consumer behavior of a large number of users in the database and commodity metadata finds the strong correlation between them, which achieves a win-win situation for not only provides the convenience to customers but also increase the enterprise profit. so, the research and improvement of temporal association rules based on Apriority algorithm in the application of personalized recommendation system is very important to the theoretical research value and practical guiding significance.

2. Journals Reviewed

Mining association rules is a model to dig out the relationship among the massive data sets, which is the most active algorithm in the field of data mining, first proposed by Agrawal etc in 1993, it is digging out the problem of association rules existed in customer transactions database, and first used in the analysis of shopping basket, that is to dig out the customer's purchase preference through analyzing the transaction record in the supermarket basket database.

Association rules can reflects the mutual dependence and relevance of an event and other events ,and it can be thought as the most common form of local mining modes without the guidance of the learning system .If the relevance exists among two things and more above, we can predict the appearance of some thing with the other things that are related .

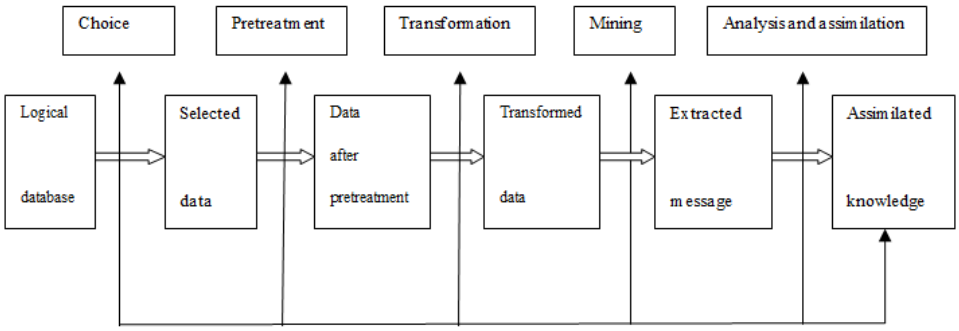

Figure 1. The Steps and Process of Data Mining

One of the most typical examples is the "beer diaper" story, which is based on to be carried out a lot of research towards the mining problem of association rules by many scientists and researchers, such as Zhu Qingxiang ,etc propose a recommendation model of multi-source matrix weighted association rules in order to improve the accuracy and efficiency of personalized recommendation systems [2]; Sun wen jun presents association rules mining algorithm based on statistic t with t statistical level taking the place of the original confidence, and making the mined association rules more credible . Zhang Tongqi etc improve accordingly the mahout FP-growth association mining algorithm, and then achieve a recommendation system based on association rules algorithm, including using the support and confidence of the association rules algorithm to mine strong association rules, and taking advantage of aging degree and interest degree to overcome the transaction without distinction of defects [3].

The methods above have advantages and disadvantages of each; there is no time latitude research. On the basis of previous studies, temporal association rules based on Apriority algorithm is mined through analyzing the historical transaction data in this paper, and it shows us the relevance between different users and goods of different time, which is convenient to recommend specific products to customers in a certain period of time.

3. Introduction of Personalized Recommendation Based on Association Rules

3.1. Referral System

The main function of personalized recommendation is to recommend customers goods or information they interested in according to their buying habits and preferences. With the development of electronic commerce, website has accumulated a massive user related data, such as transaction data, browsing data, registration data, etc, these data contains the characteristics of user's interest and purchase behavior, recommendation system can analyze these user's behaviors and information to obtain useful message. Personalized recommendation system is a kind of advanced business intelligence platform basing on large data analysis and mining. The system is mainly used in two aspects, one is to determine user's hobby towards commodity, another one is to recommend users the goods they may be interested in [4].

3.2. Recommended System Workflow

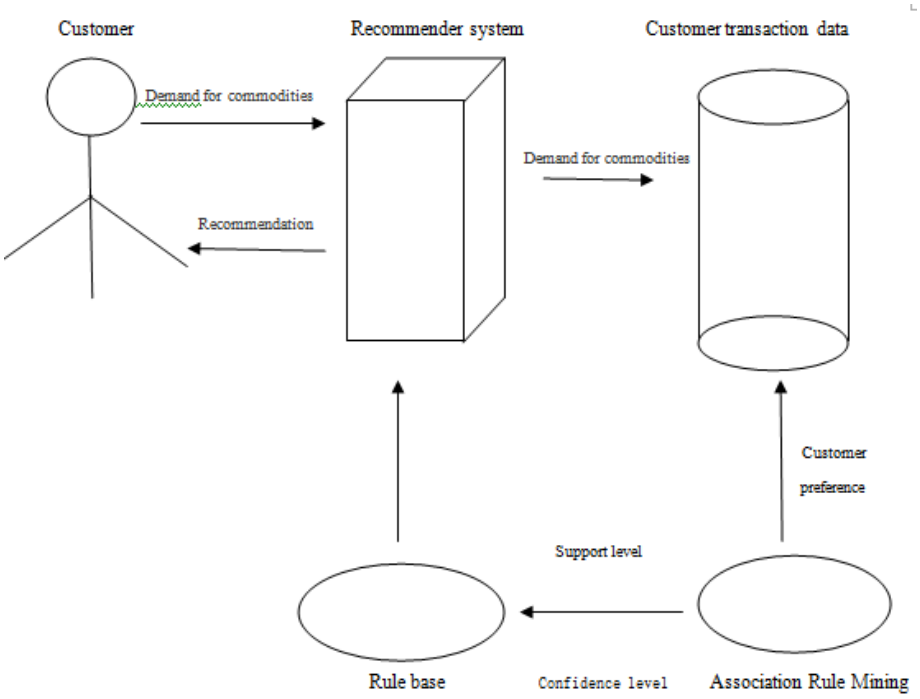

Figure 2. Workflow of Personalized Recommendation

(1) Data collection. Processing, customer registration, browsing, trading and other records, these data must be cleaned, integrated and converted before entering the database, and converted into a data type that is required for association rule mining.

The personalized recommendation system is a system that is completely based on user's data, so data collection is not only the first step of personalized recommendation system workflow, but also the premise and foundation of system work. Now there are three kinds of main data collection ways: explicit, implicit access and heuristic access.

Explicit access refers to users inform the business with information initiatively, which requires users to provide information needed by the system initiatively, including given sample set, their own view of the project and detailed objectives, etc. Explicit acquisition is the most direct and simple way to obtain user's information, which can accurately reflect the user's needs at the time with much more specific, objective, accurate and

comprehensive information. But shortcomings of explicit access are also quite obvious. Certain percentage of users is unwilling to provide their information to system because of trust issues. At the same time, explicit way has disadvantages such as diversity, poor flexibility, unguaranteed instantaneity and operability. The possibility of obtained negative data might also be aggressive to the effective information also exists.

Implicit access is a way operated without human behavior, track user's behavior only in the case of not bothering the users to collect user's data to ratiocinate their information. Implicit access will not bother the user's normal activity or produce any interference. Implicit way, however, there is also obvious flaws in it, the acquired data contains a bunch of redundant data and irrelevant information, which can not truly reflect the actual interests of the users, on the contrary, it increases the cost and complexity of the learning algorithm process.

Heuristic access method needs to provide users with heuristic information, such as authority's advice, professional term extraction, so as to realize the reuse of professional knowledge and improve the quality of obtained users interests.

(2) Form the rule base. On the basis of the data after treatment, mine model according to the association rules. In the case of setting appropriate confidence and support, a certain set of rules can be formed, and stored in the rule library.

(3) Purchase recommendations and sales management. Find out the related goods in the rules library, and then design a recommendation algorithm in the recommendation system, with this we can recommend automatically the users commodities related. Recommendation algorithm almost determines the type and the quality of personalized recommendation system, and it can set the algorithm target according to the collected user's information and the established user's model, and it can also calculate the recommender results towards the specific users, so as to recommend the related products to the users automatically .

3.3. Basic Concepts of Sequential Association Rules Mining

(1). Set: a collection of more than one item, items can be a property, and can also be a category or goods, such as {beer and diapers} is a 2 - set.

(2). Sequence item sets: the set of the corresponding time series items in the item set I is called a sequential item set, namely < (I), T>, where I is the set, and T is the time point of the time series.

(3). Support: A and B satisfies the probability of $T_b - T_a = t$ and the appearance's frequency at the same time in all affairs, i.e.

$$\mathrm{Support}(\mathrm{A} \rightarrow \mathrm{B}:\Delta\mathrm{T}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B}:\Delta\mathrm{T}) = \sigma(\mathrm{A} \cap \mathrm{B}:\Delta\mathrm{T})/\mathrm{N}$$

Among the algorithm that $\sigma(\mathrm{A} \cap \mathrm{B}:\Delta\mathrm{T})$ indicates that the transaction A and B appears at the same time and fulfils the number of times of the $T_b - T_a = t$ in the database, N represents the total size of the database.

(4).Confidence: said the probability of B also occurs and satisfies $T_b - T_a = t$ under the occurrence of A, that is to mean the ratio of probability between the A and B occurs simultaneously and satisfies $T_b - T_a = t$ in the transaction database and the probability of A occurs .the formula is expressed as:

$$\mathrm{Confidence}(\mathrm{A} \rightarrow \mathrm{B}:\ \Delta\mathrm{T}) = \mathrm{P}(\mathrm{AB}:\ \Delta\mathrm{T})/\mathrm{P}(\mathrm{A}) = \sigma(A\ \mathrm{I}\ B:\Delta T)/\sigma(A)$$

The $\sigma(A)$ above only represents the transaction of A

(5).The same maximum temporal association rules: it presents the highest Rule that A and B occurs at the same time and satisfies $\Delta\mathrm{T} = t$ in the database, such as there are 3 records in a database, respectively: : {<A,35> , <B,39>} ,

{<A,11>,<B,15>},{<A,23>,<B,29>}, among them when meet and support the temporal association rules you need to buy A, and then buy B at the time of $\Delta T = 4$ .There are two records, namely the former two, then we can call the rule $A \rightarrow B$: $\Delta T = 4$ the same maximum temporal rule for the sequential mode， and the support value is 2. [5]

3.4. Fundamental Algorithm

Association rules mining algorithms generally contain apriori algorithm , CARMA algorithm , FP-number-frequency algorithm and so on. The article herein is mainly discussing the Apriori algorithm of association rules mining algorithms. The algorithm is the first one analyzed by the association rules and also the most wildly used and impressed of all the algorithms. It uses the iteration to produce frequent item sets, the procedures of which can be described as: Firstly, scanning the transactional databases to discover the frequent item sets that could support the minimum value in the database; Secondly, finding all the rule modes of time series of each frequent item set discussed above, and selecting the qualified rule modes of time series by taking the advantage of the support of the maximum same time series. [6]

The first step of the Apriori association mining algorithm is to excavate the frequent itemsets, and this is also the most important step. The significant characteristic of Apriori algorithm is beginning with a single item, and then followed by reduction in sequence. The search process is simple according to the nature of Apriori algorithm, but the frequent itemsets is a certain difficulty in the process. First, the large number of clients and objects is the most basic requirements, and this amount sometimes could be equal to the memory of the computer, and even beyond the scope of computer storage. Then, with the increase of items, the number of frequent itemsets will also grow exponentially. This paper argues that it would be better if the algorithm is scalable.

It should be noted essentially that the timing sequences which is gained by scanning the databases is not the fake timing sequences. The definition of the fake timing sequences is that the timestamp of the timing sequence is not the maximum one, moreover the value needs to meet ΔT >0.

3.5. Algorithm Code and Instructions

Algorithm: Apriori+ algorithm
the main algorithm 。Main ( )
Transfer （）algorithm{
Input: original object database D
Output: transformed time series database
}
Apriori algorithm
{
Input: transformed time series data table
Output: filter out the frequent item sets according to the support
}
(2) Find ( ) algorithm: find the maximum sequential association rules
{
Input: Frequent item sets obtained from above
Then according to $\rightarrow$ split the rule into the left L and right R, extracting the sequence of items< L,T_{Li} >,< R,T_{Ri} > that is corresponding to the L and R of the frequent item;
If#< L,T_{Li} >=0, to return to the previous step
Else

$<L,T_{Li}>-<R,T_{Ri}>=<(L,R),\Delta T_i>$.

If $\Delta T_i > 0$

Save

Else Return to the second step

Output: The maximum time series association rule that meets the minimum support

}

4. The Application of Personalized Recommendation Based on Time Series Correlation Analysis in a Shopping Network

4.1. Data Interpretation

The purpose of this study is to apply the algorithm of association rules to the large amounts of data. Due to the availability of data, this paper analyzed the user's transaction data under the underwear category from a shopping network at 2016,January, each transaction data including records number, goods sub categories and names, Table 1 are part of the transaction data.

Table 1. Example of Transaction Data

buyer number	commodity name	purchase quantity	price	time	classification
1	bamboo fiber men's vest MZ-1003	3	22	2016/1/3	vest series
2	top US men's mercerized wool one-piece 8501#	200	45	2016/1/4	warm clothes
2	top US men's warm clothing 8508#	200	45	2016/1/4	warm clothing
3	men Luosilai autumn clothes suit 3205#	200	58	2016/1/19	Men's Suit
3	top beauty women's jacquard soft warm suit 1228#	300	100	2016/1/19	warm suit

This statistical data shows that this data contains a total of 81137 records, involving 9481 kinds of goods, and most commodities appear more frequent, so we can carry out the association rules mining.

4.2. Algorithm Implementation Process

(1) the original data contains some errors, redundant or null values and between them the conflict may come into. Therefore, it is necessary to clean data. Here first filtered out attributes of both price and quantity, second we only choose the date regards of the time selection, setting the time format as data format, with A, B, C, D, E, F, respectively representing T-shirt, safety trousers, warm vest, warm keeping shirt, warm pants, warm suit. The Transfer algorithm is used to convert the raw data into the data table, which is needed in the following table:

Table 2. Part Time Series Data Set Dealt

Buyer number	The goods and date of purchase			
1	A,3	B,5		
2	B,12	C,15	D,19	
3	A,11	B,13	D,29	E,29

4	A,11	C,22	D,26	F,29
5	B,4	C,5	D,11	E,18
6	A,13	B,9	D,18	F,26
7	A,14	B,16	D,26	

(2)Defining the minimum support value as "3", then transferring Apriori function to get a frequent two itemset {A,B}(example for the chart above)

(3)Rescanning the databases, and then finding the record of the corresponding timing itemset to constitute a temporary data chart

(4)Transferring the Find function, and then discovering there are three records available, and the support values are all more than the minimum support value, which are filtrated as results.

(5)Setting the minimum confidence as 60%, according to the definition of confidence of time serial association rule, we can know that the result of the example herein should be 3/4, equaling 75%, which is greater than the minimum confidence, so the result may be saved. The rule may be used to indicate that customer who buys product A will buy product B in two days.

4.3. The Analysis of Result of Model

(1) Products of the same genre own strong relevancy and high probability of purchasing as package deal, which is because the website talked herein is a B2B platform. On website of this kind, customer groups aim at wholesale merchants ， for example （briefs and boxers) who own high support value and confidence and also sale the products of the same genre.

(2)Some products share strong relevancy with products of the different genres, such as briefs with bras, loungewear sets with bras, loungewear sets with boxers and leggings with bras. As a fact of that, when arranging the goods shelves, you could put shelves of related goods contiguously or recommend them as a package deal to customers when promotions begin so that you could enjoy an increasing per customer transaction.

5. Conclusion

Personalized recommendation system plays a crucial role in the area of electronic commerce, which is a very important data-mining technology. In this paper, we discuss and research the personalized recommending system based on sequential association rules mining. Taking an online shopping customer sales data as an example, the system is empirically analyzed. According to the sequential association rule model, we can dig out the relevance and timeliness existing behind some of the merchandise and provide technical support of reasonable collocation to enterprises in the process of merchandising, so as to make the corresponding recommendation strategies. For instance, if a customer purchases a pair of bra, we can recommend underwear to them, and also we can adopt promotion means such as "Buy one get one free" or "Discount for buying 2" to improve the user's portfolio purchase rate.

Although we have made empirical analysis towards the availability of the temporal sequence association rule based on the transaction data of a shopping website, there still exist some deficiencies due to the complexity of the mode and the limitation of my ability:

Firstly, we just studied the single Booleans type of the related data in the association mode and didn't further discuss the other aspects of the association mode.

Secondly, a large number of candidate data set have emerged in the process of association rule mining, using the Apriori algorithm, which may affect the efficiency of the operation. How to increase the efficiency of the algorithm could be the direction of the future study.

References

[1] J. Shengyi, "Business data mining and application case analysis", Beijing: Publishing House of electronics industry, **(2014)**, pp. 284-287.
[2] H. Huiru and Z. Qingxiang, "The application of technology in the personalized recommendation based on the multi-source weighted association rules", no. 1, **(2015)**, pp. 83-187.
[3] Z. Tongqi, "Research on the integrated e-commerce recommendation system based on association rules and user's preference", Beijing: Beijing University of Posts and Telecommunications, **(2015)**.
[4] L. Jing. "Research on CRM based on data mining technology", Modern electronic technology, vol. 38, no. 11, **(2015)**, pp. 126-129.
[5] I. J. B. Schafer, J. Konstan and J. Riedl, "E-Commerce Recommendations Applications", Journal of Data Mining and Knowledge Discovery, vol. 5, no. 1-2, **(2001)**, pp. 115-153.
[6] J. Han and M. Kamber, "Data Mining Concepts and Techniques", High Education Press and Morgan Kaufmann Publishers, **(2001)**.

Author

Wang Yonggang, He was born in Henan in the March of 1982, the Master of software engineering of Zhengzhou University, and the PhD in education leadership and management of East China Normal University, he also is a Graduate tutor of the school of information science and technology of Zhengzhou Normal University, and his main research direction is e-commerce and education management.

