Apriori algorithm in the improvement and implementation of e-commerce based on data mining

Gan Tao[*]

Zhoukou Normal University, Henan Province, China

Received 1 March 2014, www.cmnt.lv

Abstract

With the rapid development of e-commerce business traffic increases rapidly. The traditional technology and infrastructure, it is hard to meet the increasing data management and rational utilization. E-commerce businesses overcome difficult, using data mining tools of implicit rule algorithm in the data mining, look for opportunities. Study a recommendation system using data mining algorithm is improved Apriori algorithm of association rules technology the most classic. Experimental results show that the improved Apriori algorithm efficiency is improved, the processing time in support of smaller more obvious.

Keywords: data mining, the Apriori algorithm, the electronic commerce

1 Introduction

With the development of the science & technology, the computer networking technology develops rapidly. At present, computer networks have become an indispensable modern information technology in our lives, which promotes the continuous development of the electronic commerce to some extent. The development of the electronic commerce also continually improves the requirements of the high concurrency and the massive data access in the electronic commerce system. As users continuously recognize the electronic commerce system, its business traffics are also increasing very fast. Faced up with the explosive growth of the business traffics, the traditional infrastructures is already hard to manage the increasing data, which influences the development of the electronic commerce in a certain degree. If the big data processing technology based on the cloud computing is introduced into the electronic commerce, its special advantages of the technology can have a certain effect on the electronic commerce and then promote its sustainable development.

2 The electronic commerce and the big data processing

The definition of the electronic commerce refers to the trading activities and the related service activities with the means of the electronic transaction in the Internet, Intranet and VAN, Value Added Network. That is, the electrification and the network of the traditional commercial business in each link. At first, the electronic commerce is divided into the broad and the narrow electronic commerce. The definition of the broad electronic commerce means the commercial activities with the use of all electronic tools, and the definition of the narrow electronic commerce means the commercial activities with the use of the Internet. No matter it is the concept of the broad electronic commerce or the narrow electronic commerce, the electronic commerce contains two aspects: 1.The electronic commerce cannot leave the Internet. If it leaves the Internet, it is not the electronic commerce. 2. It is a kind of the commercial activities with the use of the Internet. Narrowly speaking, the electronic commerce means commercial trading activities conducted on a global scale with the use of the Internet and other electronic tools, including telegraphs, telephones, broadcasts, TV, faxes, computers, computer networks, mobile telecommunications and others, that is, all commercial activities on the basis of the computer networks, including the summation of the related each behaviours, such as the commodity and the service's suppliers, advertisers, consumers and intermediaries. The electronic commerce understood by people refers to the narrow one. With the development of the business traffics, the data scale and number of the electronic commerce continuously increase so that the electronic commerce system appears great changes in a certain degree. Faced up with the large scale of the database, the traditional data processing technology cannot meet the requirements of the big data processing. The big data processing technology with its high fault tolerance, low cost, high extension and other advantages has become the most popular data processing technology in the electronic commerce system.

3 Impact and analysis of the big data processing on the electronic commerce

3.1 THE POWERFUL INFORMATION RETRIEVAL FUNCTION

The richness of the commodities directly affect the competitiveness of the electric commerce, while the massive numbers of the commodities, the complex classification

[*] ***Corresponding author's*** e-mail: gantaozknu@163.com

system, the complicated & non-structured commodity attribute data and others require IT infrastructure to have sufficient flexibility and the powerful retrieval ability. The super-large scale computing ability and the big data processing ability provided by the cloud platform framework can offer a powerful personalized information retrieval function, that is, the intelligent massive information retrieval can be done in terms of users' individual differences, individual interests and demand features. In addition, the function cans highly efficient return back to the retrieval results of the high recall ratio and the precision ratio. What's more, it can also realize the information push service, the hot information push, information recommendation and other new information retrieval service.

The technological advantages of the cloud computing can make the information retrieval and service solves the problems of the long-existing human natural language understanding and knowledge inference. Making full use of the functions of the depth data mining and the knowledge discovery can accurately and rapidly analyze the processing of users' information behaviours, understand users' natural language expressions and conduct the corresponding intelligent retrievals so that the information and products in accordance with users' requirements can be obtained. The speed and the precision of users' service should be improved and the customers' satisfaction should be maximized.

The powerful information retrieval function and the service function in the cloud computing technological advantages can solve the problems of the long-existing human natural language understanding and knowledge inference. Customers' information can be analyzed rapidly and precisely, customers' language can be understood and the data information of customer's requirements can be retrieved by making full use of the functions of the depth data mining and the knowledge discovery so that the services' precision and speed can be improved. In this way, customers can rapidly obtain the needed information, but all of these work just can be done in the database network system built in the static network environment so that it just adapts to the specific fields cannot properly adapt to the uncertain requirements of the database resources. Therefore, the database network system based on a dynamic dataset should be set to improve the managing level of the data resources, the obtaining of the precision and the inquiry ratio of the resources and the managing efficiency of the database resources.

3.2 THE PRECISED ANALYSIS TO THE DATA

The real-time big data analysis is becoming the core competitiveness of the electric commerce, and the value of the big data lies in the information analysis and utilization. The cloud computing can collect, store, analyze and process the massive data and the big data in the shortest time so that the information analysis ability of the enterprises can be improved and it is possible to have the real-time massive data mining and the big data depth analysis to the electric commerce. Taobao shopping produces massive trading time, commodity price, purchasing numbers and other data, shareholders 'years, occupations, addresses and other individual information with millions of transactions every day. Taobao shopping can precisely rank all kinds of shopping's and conduct the intelligent recommendation of the individuality in these massive data. In this way, the users' behaviour data can be analyzed and the individual information and commodities needed by the electric commerce users can be obtained so that the precision marketing can be carried out. The merchants can carry out the production and the inventory project according to the historical information and "the Taobao Index "so that the customers can obtain the commodity information in accordance with the individualized requirements and the customers' satisfaction can be improved.

The real-time big data analysis in the electric commerce field can improve its marketing competition. The detailed analysis to the data information and the rational utilization of the related information are the main value of the big data. The adoption of the cloud computing can effectively collect, store, analyze and handle the massive data in the shortest time. The information analysis of the enterprises and the data processing ability to an extent can be improved so that the electric commerce can real-time mine the massive data and analyze the data deeply.

3.3 THE RAPID FLEXIBLE PROCESSING ABILITY

The rapid flexible processing ability is the mainly purpose pursued in the electric commerce system. The rapid flexible processing ability can effectively solve the sudden traffics; customers browse requests and large numbers of the orders. In the meantime, the services should be continuously expanded and the storage equipment of the data should be increased according to the increasing of the business volume and customer requests.

The cloud storage platform based on the cloud computing technology has the unlimited massive storage, the super-large scale computing and other resources so that TB class and PB class can be stored and processed. The enterprises are unnecessary to install the hardwires for rapidly deploying and applying the system at a low price and realizing its flexibility so that the management & control capacity of the resources can be improved and the optimal utilization can be promoted. The cheap and rapid application system can be widely used in many enterprises. For example, Taobao and Tmall adopt the method to increase its sales. In this way, the big data processing can influences the electric commerce and improve the rapid flexible data processing ability and its operation efficiency.

3.4 THE INFORMATION SECURITY OF THE CLOUDING

The key sustainable development of the service is the information security in the electric commerce enterprises. The enterprises have big data with the core assets in the fierce marketing competition. The data is too sensitive and complex that it becomes the offensive targets and the revealing risks of the enterprise information privacy can be accelerated. The depth data analysis technology in the big data can make the hackers be more précised in attacking big data. The information security is the most critical sustainable safeguard of the e-commerce enterprise businesses. The big data has become the core assets of the country and the enterprises in the big data era and the big data will become the commanding height of the competition in future. The data is usually more complicated and sensitive, and easier to become the apparent targets of attacking the networks so that the revealing risks of the enterprise information privacy can be accelerated. What's worse, the depth big data analysis technology can make the hackers be more précised in attacking big data. Although the electric commerce cannot prevent the external data from mining the individual information, all social network sites can open the real-time data produced by users from different degrees. The external data providers can obtain users' information systems by collecting, monitoring and analyzing these data. The normal safety schemes and measures cannot meet the requirements of the non-liner increasing demand of the data in the data era so that the users' privacy security problem becomes rather apparent.

The big data processing technology can comprehendsively, precisely and real-time monitor the network anomalies and aggressive behaviours. In addition, the users' information can be done the real-time safely and preventive analysis, quality safe class and risk. According to the specific situations, the corresponding safety schemes and measures should be worked out for searching for the attacking source, avoiding the hackers 'invasion of the network and ensuring the information's safety.

4 Application step of data

Application data is broadly divided into the following steps: **a)** data collection, verification and filtering; **b)** classification and stored in the data warehouse; **c)** data mining association in order to find the data implied by law and data between; **d)** data modelling and parameter adjustment; **e)** data-based application development and decision support. The following examples to illustrate.

1) The American Medical website WebMD EDM regularly send to the user information based on pregnancy pregnant female users fill remind mothers precautions that point in time, you need to intake of nutrition, prenatal physiological changes and ideas ready to do a good job, postpartum recovery, the baby's upbringing and health, among others.
2) 1 shop use of big data analysis to customers send personalized EDM. If the customer had viewed a merchandise store on the 1st website without buying, there are several possibilities: a) stock, b) the price is inappropriate, c) are not wanted or not wanted brand merchandise, d) just look advise the customer when the customer if viewed in the arrival of the goods out of stock now; if there were goods and the customer did not buy because the price is very likely to be caused, at the time the merchandise markdowns notify customers; at the same time when the introduction of the goods and similar goods or associated warm inform customers. In addition, by digging cyclical buying habits of customers, at the customer's buying cycle when approaching a timely reminder to the customer.
3) Taobao launched in 2012 Taobao time machine. The application by analyzing customer self-registration for the behaviour of users since told with humour and vivid language customer Taobao's growth, the statistical behaviour of the user similar to the preferences of other users on the after analysis of the customer to understand their preferences and predict their behaviour, and so on. With vivid presentations and personalized data and customers closer distance.
4) Google's Ad sense for search process and their attention to each customer's site for data mining. And sites within its coalition tracking the whereabouts of customers, the launch customer interest and potential match's ads on affiliate sites, precision marketing, improve the conversion rate.
5) Amazon in recent years launched a FDFC (Forward Deployed Fulfilment Center) concept, in order to accelerate the speed of delivery to customers. Amazon's fulfilment center is divided into two levels: FC and FDFC, where FC varieties more complete, but FDFC physical location closer to the target market, but the species targeted market focus to accommodate a hot commodity, most of the demand by customers FDFC to meet, cannot meet the long tail of goods by FC to meet. Most merchandise so customers can FDFC needed to more efficient and cost-effective logistics to complete. Since the hot commodity is changing with time and season, so what products will be stored in FDFC decisions are dynamically adjusted, and this decision is based on customer demand analysis and forecasting.

Various examples of applications are difficult to be exhaustive, but the trend is clear: the value and potential of big data can no longer be underestimated. But not all companies can be in big data really digging this gold mine gold. Only those foresight and vision, attention system, will to invest, attract excellent business analysis and system personnel will spoils.

5 Apriori algorithm and improvement

Application data is broadly divided into the following steps: **a)** data collection, verification and filtering; **b)** Classification and stored in the data warehouse; **c)** data mining association in order to find the data implied by law and data between; **d)** data modelling and parameter adjustment; **e)** data-based application development and decision support. The following examples to illustrate.

5.1 APRIORI ALGORITHM

The Apriori algorithm used in the system will be introduced in detail as follows.

The pseudo-code in the Apriori algorithm is as follows:

```
$Input: Ts\ DataBase\ D$
$Input: Minimum\ support\ threshold\ min\_sup$
$Output: Frequent\ pattern\ L$
$L1 = \{Large\ 1\ - itemsets\};$
$for(k = 2; L_{k-1}\ != NULL; K++)$
{
$C_k'\ = join(L_{km} >\ L_{kn});$
$C_k' = pruning(Ck');$
$C_k' = A - gen(L_{k-1});$
$for\ ts\ t \in D$
{
$C_t = minset(C_k, t);$
$for\ all\ candidates\ c \in C_t$
$++c.count;$
$L_k = \{c \in C_t \| c.count \geq min\_sup\}$
$\}for\ all\ minset\ s \subseteq L_k$
{
$ifconf(s \Rightarrow Lk - s)min - conf;$
$printf("s \Rightarrow L_k - s");$
}}
```

FIGURE 1 The pseudo-code in the Apriori algorithm

5.2 APRIORI ALGORITHM IMPROVEMENT

The biggest problem in the Apriori algorithm is that the source database should be repeatedly scanned once. The algorithm not only produces a mass of candidate item sets, but also enlarging the load capacity of the system, delaying the handling time and consuming a large amount of the main memory space. Therefore, the candidate item sets should properly distribute the internal memory. The technology and the method of realizing the fast scanning to the large-scale database system is to improve the efficiency of the management rules. It is very important to effectively extract the association rules from the massive data facing up with the mass of the database.

According to the operations of the above examples, the rule can be obtained. If there is no N frequent item sets in each database, the N+1 frequent item sets cannot be included. In the meantime, the support level in the N item sets is irrelevant to the item sets which are smaller than the N item sets', it is unnecessary to consider it.

It is unnecessary to scan the segments which are smaller than the N item sets. The processing number of the data can be reduced.

The auxiliary table F should be set up, and the thing number and the segment length in the table F should be recorded. In addition, the auxiliary table should be updated according to its orders and the item sets which is smaller than the segment length and the item sets which are not included should be deleted.

The improved Apriori algorithm should scan the auxiliary table during the operation, and the record which does not exist in the auxiliary cannot be scanned. If the produced Ln frequent N item set is larger than its segment length, the record not included in the frequent item sets cannot exist in the auxiliary table, and the auxiliary table should be updated. The auxiliary table should be scanned at first and then the records not included in the auxiliary table can be skipped so that its efficiency can be improved.

5.3 THE REALIZATION OF THE IMPROVED APRIORI ALGORITHM

The following example illustrates the application of the improved Apriori algorithm. The things database is as shown in the Table 1 and the minimum support level is assumed as 4.

TABLE 1 The things database

Things number	Item set	Things number	Item set	Things number	Item set
T_1	A,B,C,D,E	T_4	A,B,C,E	T_7	A,B
T_2	---	T_5	E	T_8	---
T_3	A,C,E	T_6	C,E	T_9	A,B,C,D,E

1) Scan the database. The appearance of each item should be counted and then the C_i candidate item sets can be formed, as shown in the Table 2. The internal memory should be put in the data table so that the initial auxiliary table F1 can be produced, as shown in the Table 3.

TABLE 2 C1 candidate item sets

Item sets	Support count
{A}	5
{B}	4
{C}	5
{D}	2
{E}	6

TABLE 3 F1 auxiliary table

Things number	Length of the field	Things number	Length of the field
T1	5	T6	2
T2	0	T7	2
T3	3	T8	0
T4	4	T9	5
T5	1	-	-

2) Delete the item sets whose C_i candidate item sets are smaller than the defined support level 4, as shown in the Table 4. The F1 auxiliary table should be updated and the records whose segment length is smaller than 1 and the records which are not included in the L1 frequent 1- item sets should be deleted so that the F2 auxiliary table can be obtained, as shown in the Table 4.

TABLE 4 L1 frequent 1- item sets

Item sets	Support count
{A}	5
{B}	4
{C}	5
{D}	-
{E}	6

3) Connect the L1 frequent 1- item sets and the C2 candidate item sets can be produced, as shown in the Table 5. Through scanning the data table and the F2 auxiliary table, the L2 frequent 2- item sets can be obtained by deleting the item sets, which cannot meet the requirements of the minimum support level. As shown in the Table 6. The F3 auxiliary table can be obtained by deleting those records whose segment length is less than 2 and which are not included in the L2 frequent 2- item sets, as shown in the Table 7.

TABLE 5 C2 candidate item sets

itemsets	Support count
{A,B}	4
{A,C}	4
{A,E}	4
{B,C}	3
{B,E}	3
{C,E}	5

TABLE 6 L2 frequent 2 - item sets

itemsets	Support count
{A,B}	4
{A,C}	4
{A,E}	4
{C,E}	5

TABLE 7 F3 auxiliary table

Things number	length of the field
T_1	5
T_3	3
T_4	4
T_9	5

4) Connect the L2 frequent 2- item sets and the C3 candidate item sets can be produced, as shown in the Table 8. Through scanning the data table and the F3 auxiliary table, the L3 frequent 3- item sets can be obtained by deleting the item sets which cannot meet the requireements of the minimum support level, as shown in the Table 9. The algorithm operation should be concluded and all frequent item sets should be obtained, as shown in the Table 10.

TABLE 8 C3 candidate item sets

Item sets	Support count
{A,B,C}	3
{A,B,E}	3
{A,C,E}	4
{A,B,C,E}	5

TABLE 9 L3 frequent 3 item sets

Item sets	Support count
{A,C,E}	4

TABLE 10 Frequent item sets

Item sets	Support count
{A,B}	4
{A,C}	4
{A,E}	4
{C,E}	5
{A,C,E}	4

The operation can be accelerated and its efficiency can be obviously increased by improving the algorithm for the data in the data table is directly put in the internal memory by the system. The speed of inquiry can be accelerated for it is unnecessary to scan the previous database table again in each scanning and it just needs to visit the internal memory directly. The join of the auxiliary table can also acelerate the improvement of the efficiency. The auxiliary table can reduce the visits which is irrelevant to the records so that the number of the records in the data table can also be obviously reduced and the time can also be reduced.

5.4 THE COMPARATIVE ANALYSIS BETWEEN THE BEFORE ALGORITHM AND THE AFTER ALGORITHM

In order to verify the improved algorithm, the same software is installed in a host computer. The before improved Apriori algorithm and the after improved Apriori algorithm should be tested in the same support level so that the time comparison can be finished and the improving effect can be judged. If the nine item sets are assumed as (T1, T2, T3, T4, T5, T6, T7, T8, T9) and there are 4000 data, the operating time should be calculated in the same time complexity, as shown in the Table 11.

TABLE 11 The operating time table between the before algorithm and the after algorithm in the same support level

Support count	2	2.5	3	3.5	4	4.5	5
Before improved Apriori algorithm	49s	36s	23s	14s	12s	9s	7s
After improved Apriori algorithm	33s	24s	14s	11s	9s	7s	6s

6 Conclusion

When the support level is larger and there are less algorithm operating times, the auxiliary table and the application efficiency of extracting the algorithm is low and the advantages of the algorithm are not obvious. When the support level is smaller, the affair, which is not conformed to segment length, should be deleted, the processing time can be reduced and the database table can be reduced so that the advantages of time in the figure are obvious. In summary, the efficiency of the after improved Apriori algorithm can be improved to a certain extent and the processing time is more evident during a smaller support level. When there are a large amount of the data, the corresponding auxiliary F can also be enlarged and the internal memory space can be reduced correspondingly. In the meantime, the processing time of extracting the algorithm can be influenced so that the algorithm needs to be further improved.

References

[1] HadoopMapReduce. http://hadoop.apache.org/docs/r0.20.2/mapred_tutorial.html 2013

[2] HDFS. http://hadoop.apache.org/docs/hdfs/current/hdfs_design.html 2013

[3] Apache Hadoop NextGen MapReduce (YARN). http://hadoop.apache.org/docs/r0.23.0/hadoop-yarn/hadoop-yarn-site/YARN.html 2013

[4] Manyika J, Chui M, Brown B, et al.2011 Big data; the next frontier for innovation, competition, and productivity

[5] http://www.199it.com/archives/category/inter-net/electronic-commerce

[6] Garg S K, Versteeg S, Buyya R 2013 Future Generation Computer Systems **29**(4) 1012-23

[7] Nurmi D, Wolski R, Grzegorczyk C, et al 2009 the eucalyptus open-source cloud-computing system Cluster Computing and the Grid *2009 CCGRID'09 9th IEEE/ACM International Symposium on. IEEE* 124-31

[8] Moreno-Vozmediano R, Montero R S, Llorente I M 2013 Key Challenges in Cloud Computing: Enabling the Future Internet of Services *Internet Computing IEEE* **17**(4) 18-25

[9] Toet A, Hogervorst M A, Nikolov S G, Lewis J J, Dixon T D, Bull D R, Canagarajah C N 2010 Towards Cognitive Image Fusion Information Fusion 11(2) 95-113

[10] Li G, Yang W, Weng T 2007 A Method of Removing Thin Cloud in Remote Sensing Image Based on the Homomorphic Filter Algorithm Science of Surveying and Mapping **32**(3) 47-8

[11] Liu Y, Bai J 2008 Research on the Cloud Removal Method of Remote Sensing Images Geomatics & Spatial Information Technology 3 120-2

Author

Gan Tao, April 13, 1978, Henan Province, China.

Current position, grades: associate professor of Zhoukou Normal University.
University studies: economic information management.
Scientific interest: the small and medium-sized enterprise management.