Advances in Computer Science Research (ACSR), volume 61
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

Research On Novel Model of Data Mining Based on Improved Association Rules and Clustering Algorithm

Qing Tan
College of Information Technology, Luoyang Normal University, Henan Luoyang, 471934, China
edutanqing@163.com

Keywords: Apriori algorithm; Decision tree; Association rule; Clustering; Data mining

Abstract. Apriori algorithm is one of the most effective algorithms for mining frequent itemsets of Boolean Association rules. Decision tree is a method to analyze and summarize the attributes of a large number of samples. The frequent itemsets are used to generate the association rules, and the strong association rules are generated according to the minimum confidence set by the user. The paper presents research on novel model of data mining based on improved association rules and clustering algorithm. Finally, the effectiveness of the proposed algorithm is verified by experiments.

Introduction

The information on the Web website is a larger, more complex data cube. If each of the site information on the Web as a data source, and it is these data sources are heterogeneous, because each site information and organization is not the same. Want to use this massive data mining; first of all, we must study the integration of heterogeneous data between sites [1]. It is possible to integrate the data from these sites into a unified view to obtain the desired. Secondly, we must solve the problem of data query on Web, because if the data is not available, the analysis, integration and processing of these data will be impossible.

Concurrency requires high and strict requirements for transaction integrity, security. The difference between OLTP and OLAP: users and system orientation: OLTP customers, OLAP and market oriented; data content: OLTP management system of the current data and history data management OLAP; database design: an OLTP system using entity relationship (ER) model and application oriented database design, OLAP system typically the star and snowflake model; view: OLTP system is mainly concerned with the current data within an enterprise or department, and unified data mainly focus on the summary of the OLAP system; access mode: OLTP access are short of atomic transaction, and most of the OLAP system access is read-only operations, although many may be complex queries.

Particle swarm optimization algorithm (PSO) since its fast convergence has been widely used in various aspects, if the algorithm is applied to a simple clustering is obviously not realistic, but only the PSO and K-means algorithm will affect the clustering performance. Therefore, it is in the case of not affecting the merits of the algorithm itself to maximize the ability to enhance the clustering, the foundation for the acquisition of association rules.

Decision tree is a process of classifying data by a series of rules. The decision tree is divided into two kinds: classification tree and regression tree. The basic principles and methods of decision tree data mining are studied through decision tree experiments.

Firstly, the transaction database is divided into several non overlapping sub databases, and the frequent itemsets mining is carried out [2]. Finally, all the local frequent itemsets are combined as the candidate itemsets of the whole transaction database. Scan the original database to calculate the support of candidate set. The algorithm generates the frequent itemsets of the whole transaction database and only needs to scan the database two times. Based on hash technology, by using hash technology, DHP (Direct-Hush and Prune) 5 can be used to generate more candidate itemsets when generating candidate sets. So each candidate set is more approximate to the frequent set. This technique for 2 candidates pruning is especially effective. On the other hand, DHP technology can effectively reduce the size of

each scan database. The paper presents research on novel model of data mining based on improved association rules and clustering algorithm.

Typical Example of Association Rule Mining and Frequent Items

A typical example of association rule mining is the analysis of shopping basket. Market analysts need to find out from a great deal of data about the relationship between the different goods that customers put in their shopping baskets. If a customer buys milk, how much is it possible for him to buy bread? What group or collection of customers most likely to buy at the same time shopping? For example, there are 80% customers to buy milk also to buy bread, or 70% of the customers who buy a hammer also buy nails, which are extracted from the shopping basket in the data association rules. The analysis results can help managers to design different store layouts.

Correlation tests were conducted in the sample elements of the whole system, mainly constitute the interrelationship between the elements that is to check whether these defects affect each other, if the mutual influence, and has a very important position in the system, impact on system security is high, the system failure of the obvious signs [3]. This stage is the main examination results of second stage data accumulation of large data based on Data Mining Based on dynamic, real-time reflect the focus of inspection, improve the performance of inspection and targeted.

Data mining is a kind of deep level data analysis method. The data analysis has a history of many years, but in the past data collection and analysis for the purpose of scientific research, in addition, due to the limited computing power for complex data analysis is limited to the large volume of data. Now, due to the implementation of business automation in various industries, the business sector has generated a large number of business data, the data is no longer for the purpose of analysis and collection, but because of the commercial operation of the. The analysis of these data is no longer just for the sake of the research, but also for the business decision to provide the real valuable information, and then get profit.

Select the frequent items in Trans and sort them in the order in L. The list of frequent items set after Trans is [p|P], where p is the first element, and P is the list of the remaining elements. Call insert_tree ([p|P], T) [4]. The process is as follows. If T have children N make N.item-name=p.item-name, N counts increased 1; otherwise, create a new node N, the count is set to 1, link to its parent T, and through the chain link to node structure with the same item-name node. If P is not empty, then recursively call insert_tree (P, N), the formula is as follows, as is shown by equation(1).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-a}{b}\right) \tag{1}$$

In fact, the evaluation of data cleaning is to evaluate the quality of the data after cleaning, and the process of data quality evaluation is a process of optimizing the data value by measuring and improving the comprehensive features of the data. The difficulty of data quality evaluation index and method lies in the meaning, content, classification, classification and quality evaluation of data quality.

Clarans algorithm is a random search method based on it is proposed based on the Clara algorithm, it is different with the Clara algorithm: Clara algorithm to find the best process center in the sample is unchanged, and the Clarans algorithm in each iteration process using sampling is not the same. The advantage of this method is that on the one hand, it improves the clustering quality of Clara; on the other hand, it expands the range of data processing [5]. It has better clustering effect, but its computational complexity is O (N2), therefore, one of its shortcomings is still low, is not sensitive to noise data, but sensitive to the order of input data, only clustering convex or spherical boundary.

Web information is frequently changed, such as news, stocks and other information is updated in real time. And this high change is also reflected in the dynamic link and random access on the page. Second, users on Web are hard to predict. Users have different knowledge background, interest and purpose. Finally, the data environment on Web is high noise. Studies have shown that the data of a Web site may

have no more than 1% of the information that is relevant to a particular mining theme. These variables are also Web data mining must face the problem.

Markov chain analysis is method. The length of the system call sequence t (S1, S2),... (St), calculate the probability of St+1 (St+1S1) in the normal system mode under the next system call P,... (St). Take the sequence here (S1, S2),... St, Si) state of all time (1 I t) to the St+1 transition probability weighted average value; the calculation formula is as follows [6].

$$\begin{aligned}\bar{v}_i(m) &:= \left[\bar{v}_i^T(m,1), \bar{v}_i^T(m,2), \cdots, \bar{v}_i^T(m,M)\right]^T \\ &:= \left[v_i^T((m-1)M+1), v_i^T((m-1)M+2), \cdots, v_i^T(mM)\right]^T\end{aligned} \tag{2}$$

The input timing mode in accordance with the rules "," Fu split L and right R left set. 1) Locate the temporary data input in the I table records, L and R extraction set in the corresponding record in the temporal item (set) < > and < >. 2) Pseudo sequential set discriminating operation of < >, if the output is 0, it returns second step). 3) Timing set subtraction: < >-< >=< (L, R), which, if > >0, save the results.

There are some scholars proposed the combination of two algorithms to use, but the method has some limitations in the data mining of educational management, so the research on Clustering Algorithm in a study of a new algorithm k-means algorithm and PSO algorithm are combined organically, which may be related to the two the adaptation algorithm, through the final research after solving algorithm, obtained by clustering algorithm to deal with multi-dimensional data management is part of the problem in the process of association rules [7].

Then, prune (pruning) step, i.e. for any C, C, Ck, Ck, delete all those (k-1) dimensional subset is not Lk-1 in the project set, get the candidate itemsets Ck. Equation(3) is expressed as: forallitemsetc, Ck, forall (k-1) - sofcif (s does not belong to the subset of Lk-1 thendeletecfromCk) the set of all k-1; said: Ck=={X, Ck'|X dimensional subset in Lk-1}.

$$PV(x) = \frac{1}{y_2 - y_1} \sum_{y=y_1}^{y_2} I(x, y), \quad x \in (x_1, x_2) \tag{3}$$

Let I be a set of data items, D is associated with the collection of data in D in each transaction T (Transaction) is a non empty subset of I, namely TI, each transaction has a corresponding identification number, called TID (Transaction ID). If A and B for the project set, and a A B= defines support association rules AB (Support) D in A and P transaction also includes the probability of B (A, B), the confidence level (Confidence) for A D is included when the transaction also includes the percentage of B, P (B|A) is the conditional probability.

A relatively complete set of time series modeling theory and analysis method, the mathematical method of these classical established stochastic model, such as autoregressive model, autoregressive moving average model, ARIMA model and seasonal adjustment model, prediction of time series. Because a large number of time series are non-stationary, the characteristic parameters and data distribution change with time. Therefore, only through the training of a certain historical data, the establishment of a single neural network prediction model, but also can not achieve accurate prediction tasks.

Novel Model of Data Mining Based on Improved Association Rules and Clustering Algorithm

The ID3 attribute selection is used sub tree information gain (here can be defined in many ways, ID3 is the use of entropy (entropy (entropy) is a kind of purity measure), is) the change in entropy value. While C4.5 is using the information and it is gain rate. That's a lot more. General rate is used to balance the variance as the role of almost, for example there are two runners, a starting point is 10m/s, the 1s 20m/s; another one is 1m/s, the speed up after 1s for 2m/s. If you count the difference, then the two gaps is very large, if the use of speed increase rate (acceleration) to measure, the same is true of the 2.

CUR E is a very novel hierarchical aggregation algorithm, using intermediate strategy between centroid and based on the representative object method based on it, choose a fixed number of data space

and representative points to represent a cluster, and these points will be multiplied by an appropriate contraction factor, making them more closer to the cluster center. Its time complexity is 0 (n). Its advantage is to select more than one representative of the algorithm can be adapted to non spherical geometry, cluster shrinkage or condensation can contribute to the effect of noise control, the method uses a combination of random sampling and segmentation to improve the efficiency, has good shrinkability of large database [8].

Web usage mining mainly refers to the mining of Log log records on Web. The Log log on Web records access information, including URL requests, and it is IP addresses, and time [9]. The analysis and discovery of the rules in the Log log can help us identify potential customers, track the quality of Web services, and detect the hidden dangers of illegal access, as is shown by equation (4).

$$\|Y_m\| = \left(\sum_{n=1}^{N} |Y_m(n)|^2 \right)^{1/2} \tag{4}$$

Where: PSiSt+1 is the I time system calls for Si, t+1 system calls for St+1 probability,.Wi transfer probability weights state Si to St+1 by t+1-i step transtion-probablity matrix, is the embodiment of the effects of t+1-i and St+1 from Si to St+1, the smaller the distance effect is bigger, so with I increase the weight becomes large, the Wi= I2 (1 I t). If the probability is less than a given threshold, the sequence is an abnormal sequence, otherwise normal.

System Experiments and Analysis

The rule of classification is to classify the new data according to the rules. The application domain is one of the most widely used technologies in the field of data mining. Many classification algorithms are included in the package of statistical analysis tools. Classification has been widely used in the fields of business, banking, medical diagnosis, biology, text mining, Internet filtering and so on. The paper presents research on novel model of data mining based on improved association rules and clustering algorithm.

KDD is a kind of important means of mining knowledge from a large number of historical data, and it is also an important method of knowledge acquisition based on knowledge. Through the experiment, understanding between database, knowledge base and model base, and the "three base" relationship with decision support system, understand the importance of data mining in decision support, grasp the basic principle of Apriori algorithm [10]. Experiment content: the basic algorithm for Apriori algorithm in data mining, by layer iteration, find frequent itemsets programming with high-level programming language, basic knowledge of database mining, as is shown by equation (5).

$$P_{j-1} f = P_j f + Q_j f = \sum_k c_k^j \phi_{jk} + \sum_k d_k^j \psi_{jk} \tag{5}$$

If the large database, reduce the time overhead of the data mining efficiency is obviously, the Apriori method is not involved in the new generation; second, the improved algorithm to scan the database after the 'Delete' (some can not support the frequent sets of records, here the so-called delete is actually not with scan again to compare the condition of the records through the exchange record the contents of the shift to the database at the end of the end of the new record in the record position. At the same time reduce the number of records in the database logically.

Therefore, the associated factors of a data mining can be associated with retention factors between the complete analyses based on the relationship between the factors, so as to provide services. PSC inspectors according to figure 3 is located in the heart of priority check project in the inspection for PSC related inspection. If the defects can be inspection related to larger, and the adjacent is to the other items.

Apriori algorithm is one of the most effective algorithms for mining frequent itemsets of Boolean Association rules. The name of the algorithm is based on the fact that the algorithm uses a priori knowledge of the properties of frequent itemsets. It uses an iterative method called layer by layer search, the k- item set is used to explore (k+1) - itemsets. First, we find the set of frequent 1- itemsets. Then we

use the former to find the set of 2- item sets, so that we can't find the frequent k- itemsets. Finally, association rules are generated by frequent itemsets.

Summary

FP-growth (Frequent Pattern-growth) uses a compact data structure to store all the information needed to find frequent itemsets. The paper presents research on novel model of data mining based on improved association rules and clustering algorithm. The algorithm will provide compressed frequent itemsets database to a FP-tree to keep items related information, and then the compressed database into a set of conditions for the database (a special type of projection database), a relational database for each condition of frequent itemsets. SOFM is a kind of unsupervised clustering method, which is based on repeated learning to cluster the data. This method is based on unsupervised learning, visualization, maintains the topology and probability keep its characteristics, widely used in clustering analysis, image processing, speech recognition and other information processing field, but it also has many deficiencies.

References

[1] Han J et al. FreeSpan: Frequent parttern-projected sequential pattern mining. In Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining. Boston, USA. Aug. 2000: 355~359.

[2] Xiaoyan Wan, "Research on Data Mining Technology of Association Rule", JCIT, Vol. 8, No. 6, pp. 628 ~ 635, 2013.

[3] Yan Hai, HongLing Han, Guiming Lu, "Data Mining based on Rough Set and Decision Tree Optimization", JDCTA, Vol. 6, No. 12, pp. 480 ~ 489, 2012.

[4] Xiang-Rong Jiang, Le Gruenwald . Microarray gene expression data association rules mining based on BSC-tree and FIS-tree Data & Knowledge Engineering, Volume 53, Issue 1, April 2005, Pages 3-29.

[5] Srivastava J, Cooley R, Deshpande M, et al. Web usage mining:discovery and application of usage patterns from web data. SIGKDD Explorations, 2000,1(2):12-23.

[6] YiJie Chen, "The Development Of The Commodity Flow Analysis System Based On Association Rule Mining", IJACT, Vol. 4, No. 13, pp. 430 ~ 436, 2012.

[7] Yuh-Jiuan Tsay, Jiunn-Yann Chiang. CBAR: an efficient method for mining association rules Knowledge-Based Systems, Volume 18, Issues 2-3, April 2005, Pages 99-105.

[8] Somboon Anekritmongkol, Kulthon Kasamsan, "The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model", IJACT, Vol. 3, No. 1, pp. 58 ~ 67, 2011.

[9] Fernando Berzal, Juan-Carlos Cubero, Nicolás Marín, José-María Serrano .TBAR: An efficient method for association rule mining in relational databases Data & Knowledge Engineering, Volume 37, Issue 1, April 2001, Pages 47-64.

[10] Shelly Salim, Sangman Moh, "Energy-Efficient Clustering Based on Game Theory for Apriori Rule Techniques ", IJEI, Vol. 5, No. 3, pp. 31 ~ 37, 2014.