Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 631-635
doi:10.4028/www.scientific.net/AEF.6-7.631

The Application of Fuzzy Association Rule Mining in E-Commerce Information System Mining

Qing Duan[1, a], Jian Li[1] ,Yu Wang[1]

[1]Office of Academic Affairs, The Central Institute For Correctional Police, Baoding 071000,Hebei,China

[a]duanqing2011@sina.com

Keywords: E-Commerce; data mining; fuzzy correlation; association rule mining

Abstract. Data mining in e-commerce application is information into business knowledge in the process. First of all, the object of clear data mining to determine the theme of business applications; around the commercial main data collection source, and clean up the data conversion, integration processing technology, and selects the appropriate data mining algorithms to build data mining models. This paper presents the application of fuzzy association rule mining in E-commerce information system mining. Experimental data sets prove that the proposed algorithm is effective and reasonable.

Introduction

E-commerce is the use of digital electronic means business on the Internet for data exchange and business operational activities carried out. With the wide application of the rapid development of database technology and the Internet, e-commerce is more powerful vitality, e-commerce website every day there may be millions of online transactions, accumulated more and more business data on the server the current database system can achieve efficient data entry, query, statistics and other functions, but can not find the rules of the relationship that exists in the database can not predict the future development trend based on available data, using data mining technology can be effectively found a large number of regularity of the data behind the data hidden inside knowledge and means to eliminate the data explosion but knowledge poverty "of the phenomenon.

Association rule mining process consists of two phases: the first phase of data collection must first find all the frequent group (Frequent Itemsets), the second stage to generate association rules from these frequent group (Association, the Rules). Association rules originated in the field of data mining, people use it to find large amounts of data between itemsets (interesting / useful) of the association. It is an important research topic in the field of data mining in recent years is due to the industry widely used and highly valued [1]. Apriori algorithm is one of the most influential frequent itemsets mining Boolean association rules algorithms. Its core is a set of ideas based on a two-stage frequency recursive algorithm. The association rules belong in the classification of one-dimensional, single Boolean association rules. Here, all support is greater than the minimum support itemsets called frequent itemsets, referred to as the frequency set.

The purpose of the association rule is a hidden order to tap the mutual relationship between the data to find the association rules of the customers between the various documents on the site. Associated analysis techniques are the confidence and support in the statistical analysis. In general, the only degree of confidence and support high of association rules may be of interest to users, useful connection rules. Fuzzy sets to represent and manipulate uncertainty data, fuzzy set attached to the representative of the function concept, it can not only deal with incomplete data, noise or imprecise data, can also be used for the development of uncertainty in the data model, can provide smarter, smoother performance than traditional methods. This paper presents the application of fuzzy association rule mining in E-commerce information system mining.

Association rule mining in e-commerce information system

Data mining is the extraction from large data or "mining" knowledge, and then the data mining must first consider what kind of data mining knowledge, that is data mining the data source.

FP-Growth algorithm is the core of the FP-tree (Frequent Pattern Tree, frequent pattern tree) is constructed, this particular data structure, FP-Growth algorithm compared with Apriori algorithm, the performance was significantly improved [2]. However, a closer look at the FP-Tree implementation, you can find it with the string processing algorithm Prefix-Tree algorithm, has the same purpose. FP-Tree by merging a number of repeated paths to achieve data compression, which makes frequent item set, is loaded into memory become possible. After the operation of the tree traversal, instead of Apriori algorithm in the most time-consuming transaction log traversal, thus greatly improving the efficiency of operations is shown by equation 1.

$$a_i(k) = \beta_i(k)s(k) + \alpha_i(k) + \varepsilon_i(k) \tag{1}$$

The server's log file: The user visits a page; the Web server's log will add a record, record Cookies and CGI query parameters to describe the behavior of different users. For example, the domain name of the customer's purchase of a product, a large number of buyers from which countries or regions, based on this information to adjust the e-commerce online marketing strategy to increase business activities in which regions or countries.

The original records from the e-commerce data source, not only a huge amount of data, but there may be a lot of noise data, redundant data, sparse data or incomplete data, etc., directly on excavation very difficult. In fact, data mining is the final success or not, whether there are economic benefits, the data ready to play a crucial role in data preprocessing including data cleaning, integration, selection and transformation. This result is in good order, first of all in accordance with the frequency from reaching the sort, and then sorted in alphabetical order. Should be noted that the sorting here is very important, and after each transaction item should be ranked in that order, this is a prerequisite for the effective integration of repeated path, is shown by equation 2.

$$F(a,b) = \sum_{i=0}^{n} \varepsilon_i^{\,2} = \sum_{i=0}^{n} (y_i - ax_i - b)^2 \tag{2}$$

Data mining requires data integration, data sources include data from multiple data sources is combined processing, to solve the semantic ambiguity and stored in a unified data store (such as data warehouses, databases, etc.), e-commerce page images, graphics, multimedia, the URL path and the relevant log files, etc., involves three aspects: entity recognition schema integration to remove data redundancy and detecting and processing the data value conflicts, it is also known as the base material, as are shown by equation3.

$$\sum_{i-1}^{N} e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) \tag{3}$$

The data selection is based on the tasks and data content understanding, looking for depends on finding the target expression data useful features, to reduce the data size, thus streamlining the amount of data under the premise of maintaining as much data the original, by the data selection can make more apparent regularity and potential features of the data. At the same time reduce the data size, data selection is complete, the need to cover the relevant data involved in the business objectives. To search for all internal and external data related to the business object, and choose the data for data mining applications, as is shown by figure1.

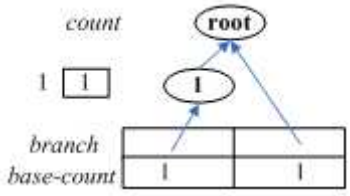

Fig. 1 The Association rule mining in e-commerce information system

The association rule has two parts, provided that (if) and (then). The premise that the items are found in the data is in it. The result is an item found in combination with the premise. Association rules (association rule) and applies the standard of support and trust created by the frequency of the model assumptions of the analysis of data to conclude that the most important relationship. Support the frequency in the database. Trust shows verified as the true number of assumptions statement.

Convert the data into an analytical model, this analysis model is established for the mining algorithm to create a truly analytical model suitable for mining algorithms is the key to the success of data mining. Include: discrete data, new variables, transition variables, split the data and format conversion. Data the actual mining process, data cleaning, data integration, data selection and data transformation may not necessarily be used. In addition, their use is not the order of a certain kind of pretreatment may have to be repeated.

Using fuzzy association rule mining to build E-commerce information system mining

Fuzzy sets to represent and manipulate uncertainty data, fuzzy set attached to the representative of the function concept, it can not only deal with incomplete data, noise or imprecise data, can also be used for the development of uncertainty in the data model, can provide smarter, smoother performance than traditional methods [3]. The traditional association rule mining and fuzzy set that combines a key method, the first elements of fuzzy sets of each property treated as a database property, then the pruning step in association rule mining will have the same properties set to delete.

Fuzzy association rules mining can be divided into two steps: (1) According to the support generated frequent sets from the fuzzy database; (2) from the frequent focus to generate the candidate rule set, and every one of candidate rules to calculate the degree of implication, and then strong fuzzy association rules based on ĐMin .... imp. This approach is intuitive and easy to understand, but there are problems: the calculation of each candidate rule "implies a degree need to scan the entire database, while the candidate rules number of very large (if frequent set of m items, the resulting r-2 rule), which brought a large number of database scan operation.

Suppose that D is a collection of a typical transaction, database, recorded in D for D = t, t, t, ..., t, which t1 ≤ i ≤ n the i-th transaction, the database contains properties for the set R =r, r, r, ..., r, r1 ≤ j ≤ m in the database field, d is the data item. Boolean association rules and quantitative association rule mining, transaction support attribute count (vote) is based on the transaction to calculate the number of occurrences in all affairs. Rules of XY in the transaction database D support (support) S-affairs to focus on the transaction number and the ratio of the total number of transactions containing X and Y, denoted by S (XY) is the nature of equation 4.

$$P\{X_1 = x_1, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i; \theta_1, \cdots, \theta_k) \tag{4}$$

Rule XY focus on the credibility (confidence) in the transaction C is the number of transaction that contains the X and Y and the ratio of the number of transactions that contain X, denoted by C (XY), C (XY) = | {T: X-∪YT, T ∈ D} | / | {T: XT, T-∈ D} |.Before carrying out excavation, and the traditional association rule mining, you must define the minimum support S and the minimum confidence C Mining fuzzy association rules, support count is calculated with fuzzy attribute data of the database D membership degree of each attribute to a real number between 0 and 1.

Mining fuzzy similar association rules algorithm is to determine the minimum similarity α sets and equivalence classes of goods, use of existing algorithms to determine the membership function, and convert the value into a fuzzy set expressed in the language, and statistical items each membership value of fuzzy attributes and fuzzy attribute of the maximum value of the items; equivalence class of fuzzy attribute take the maximum value of the membership of the equivalence class of goods as the equivalence class of fuzzy attribute; to determine the candidate see the membership value is to take the intersection of the membership values of the candidate set of items in each transaction, that is to take the candidate to focus on the items attached to the minimum, and add them together, the candidate set of membership values;membership values of the candidate set of items is greater than or equal to the minimum support the candidate set of large items on the equivalence class of sets L r 1. Mining fuzzy association rules is to find the rules to meet the greater than a user-specified minimum support, minimum confidence and minimum similarity.

Attribute r, j represents the jth fuzzy set of attributes, namely the j-th column; p is the concentration of the p-th attribute of the property. For any one property, the support of all the affairs of the property to count the sum divided by the total transaction number n, ie, all the affairs of the column corresponding to attribute the degree of support: vote R = dn, such as vote.r =(d + d + d + ... + d) / n is the degree of support of all transactions on the property r.After all of the properties of support, will support less than S attributes to remove, you get the frequent 1 - itemsets L, as is shown by equation5.

$$\Sigma_{\varepsilon(k)} = \mathrm{diag}[\sigma^2_{\varepsilon 1(k)}, \sigma^2_{\varepsilon 2(k)}, \cdots, \sigma^2_{\varepsilon q(k)}] \tag{5}$$

Get the support of all itemsets, pruning of C. Pruning include three parts: (1) delete the C support is less than S is set; ② Remove containing non-frequent set of itemsets in C; ③ delete belong to the same fuzzy set properties set C contains, so no practical significance of the items set for the final association rules can simplify the algorithm, but also reduces the amount of computation. Repeat the above steps until L = Φ, contains the greatest attributes of frequent j-itemsets L (to meet the maximum, and the support of the set is greater than S), generated by L, frequent itemsets L, by Lgenerate association rules [4].

Fuzzy set of fuzzy expression of uncertainty most directly reflect a human brain of objective things. This is exacerbated by the complexity and diversity of the fuzzy set membership determined by the membership function. Difficult to use a unified model is in order to determine the membership function.

Assumptions <X,A> behalf of a "set - fuzzy set X is a collection of attributes x, x ∈ X, A is the fuzzy set a set a ∈ A. Support count of each transaction is calculated by the membership function of x to the i-th transaction value of x, tx, Atx t the degree of membership, the support of a transaction count is greater than 0, that is, to meet <X, A>.After all x's membership to get a transaction, you can get the record t <X,A> total support count.

In order to reduce the number of scans, it can first generate all the candidate rule sets, and then scan the database once to get the implication of all candidate rules. Algorithm 1 to reduce the number of database scanning to improve the computational efficiency, to a certain extent, but when candidate rules to achieve a certain number, the computer capacity may not be sufficient, and thus unable to complete the excavation task analysis of candidate fuzzy rules can be found in the previous section, the candidate rule set redundancy of fuzzy association rules is a strong fuzzy association rules need to calculate the implication of the degree. Take advantage of this nature, before scanning the database, has been dug out of the strong fuzzy association rules from the candidate rule set to delete the redundant fuzzy association rules, thereby enhancing efficiency.

In this paper, the fuzzy association rules between itemsets mining customer transaction data, the core is a recursive algorithm for frequent set based on two-stage thinking. The association rules in a one-dimensional, it is single and Boolean association rules of FP-Growth, classification Fuzzy algorithm Aprior algorithm, as is shown by figure2. Used in this study the machine configuration Intel Pentium 4 processor, 2G memories, 500G hard disk, the operating system to Windows XP, use the tools of MATLAB 9.

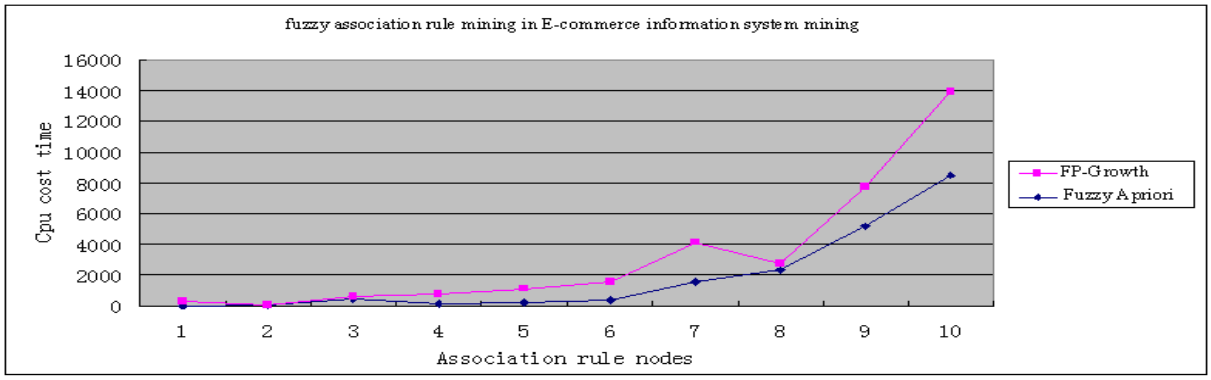

Fig. 2 The compare of building materials science and information system based on Fuzzy Apriori with FP-Growth algorithms

Summary

Data mining is a new information processing technology, through the analysis of business data processing; you can find the business knowledge hidden in the data mining data intrinsically linked to the rules and patterns, auxiliary business decisions. Data mining in e-commerce environment, customer access to information mining, it is article data sources, a brief introduction of the process of data preprocessing. Finally, the fuzzy sets and the traditional association rule mining algorithm combined.

References

[1] Yuan Wang, Lan Zheng, "Endocrine Hormones Association Rules Mining Based on Improved Apriori Algorithm", JCIT, Vol. 7, No. 7, pp. 72 ~ 82, 2012

[2] 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.

[3] He Yueshun , Du Ping, "The Research of Landslide Monitring and Pre-warning Based on Association Rules Mining", JCIT, Vol. 6, No. 9, pp. 89 ~ 95, 2011.

[4] Xiaopeng Jian, Yanfang Li, zhengqiang Jiang, Lite Li, "The Research on Technologic Economics Comprehensive Appraisal of Wastewater Treatment in the Paper Making Factory with Intuitionistic Fuzzy Information", AISS, Vol. 4, No. 4, pp. 293 ~ 299, 2012.