ATLANTIS PRESS

Advances in Computer Science Research, volume 83

8th International Conference on Social Network, Communication and Education (SNCE 2018)

Analysis of Web Mining Method Based on Intelligent E-Commerce Data

Dahai Wang [1,a]

[1]College of Humanities and Sciences of Northeast Normal University, Changchun, Jilin Province 130117, China

[a]11268222@qq.com

Keywords: Intelligent; E-commerce; Web; Mining

Abstract. This paper mainly contains the collection and preparation, the pre-processing, and the transformation of intelligent e-commerce data. The collection and preparation of Web mining for intelligent e-commerce data is the biggest obstacle to carry out intelligent e-commerce data mining. According to needs, it can integrate all kinds of intelligent e-commerce data from multiple intelligent e-commerce data sources, or derive needed indicators from existing intelligent e-commerce data. Then, the collected intelligent e-commerce data will be preprocessed, including noise removal, filling in missing values, intelligent e-commerce data transformation and so on. The main purpose of intelligent e-commerce data transformation is to reduce the dimension of intelligent e-commerce data (Reducing dimension), and reduce the burden of Web mining for intelligent e-commerce data.

Foreword

Determining a theme is the first thing to be determined in the intelligent e-commerce data mining. Only to determine what kind of knowledge to discover can plan the whole intelligent e-commerce data mining process. In the process of determining a theme, several problems should be solved: where to start, what intelligent e-commerce data to check, how much intelligent e-commerce data to be used, what degree to mine intelligent e-commerce data, and so on. After defining the theme, formulate the task of intelligent e-commerce data mining.[1]First of all, define task-related intelligent e-commerce data, and then determine the conditions of intelligent e-commerce data selection, the conditions of intelligent e-commerce data grouping, the related attributes or dimensions and so on. Secondly, identify task-related types of knowledge, namely, finding out one or more types from the aspects of characterization and discriminant, association, classification, prediction, Web mining and evolution analysis. Then obtain certain background knowledge, which is the basis for correctly making the concept stratification and maintaining the user's connection to the intelligent e-commerce data. Finally, define the measurement and task-related mode interestingness: interestingness measurement includes the conciseness (such as rule length), certainty (confidence coefficient), practicality (support degree) and novelty of evaluation mode.

Intelligent E-commerce Data Mining

The intelligent e-commerce data mining phase is based on the knowledge type of intelligent e-commerce data mining task defined in the first stage, such as classification, Web mining, association rule discovery or sequential pattern discovery, and decide what algorithm to use. This part is also the core of the whole process. According to the different characteristics of intelligent e-commerce data and the requirements of users or the actual operation system, the corresponding intelligent e-commerce data mining algorithm will be chosen. For example, some users want the descriptive type, and some users want to get the predictive type. A detailed intelligent e-commerce data mining method will be given below.

Pattern evaluation is based on the measurement and mode interestingness defined in the intelligent e-commerce data mining task, and identification shows the truly interesting mode of representing knowledge. Knowledge representation refers to what technology is used to express the knowledge of mining to the user.[2] The knowledge found through intelligent e-commerce data

mining does not all meet the requirements of the user. It is necessary to propose redundant or unrelated knowledge modes. Sometimes it is necessary to return to one of the previous stages and redo it with another method, and even start again, by step by step selecting the intelligent e-commerce data, adopting a new intelligent electronic commerce data transformation method, changing an algorithm and so on. The mined knowledge mode will be reassessed until the user's requirements are met. The ultimate goal of intelligent e-commerce data mining is to serve the user, and the mined mode needs to be expressed in the way the user's easy to understand.

Intelligent e-commerce data mining is a repeatable process, until the user is satisfied. But the quality of mined knowledge is not only related to the technology of intelligent e-commerce data mining, but also closely related to the quality and quantity of intelligent e-commerce data. Either of them is not satisfied, and the result may not satisfy the user.

The Technical Method of Intelligent E-commerce Data Mining

The technology scope of intelligent e-commerce data mining is very extensive, which integrates statistics, AI (artificial intelligence) and machine learning. Various methods are listed below:

Statistical Analysis Method. Statistical analysis method is to carry out statistical analysis of the attributes of intelligent e-commerce database based on the principles of probability theory and statistics, to find out the relation and rule between attributes. Statistical analysis method is one of the main technologies of intelligent e-commerce data mining. The relation between intelligent e-commerce data items in intelligent e-commerce database can be roughly divided into two types, namely function relation and correlation. The correlation refers to the relation between the intelligent e-commerce data items of the intelligent e-commerce database that cannot be represented in the form of function.

Rough Set Method. The rough set theory was proposed by Professor Z.Pawlak of Poland in 1982. It is an intelligent e-commerce data analysis theory, which can effectively analyze incomplete information such as inaccuracy, inconsistency and incompleteness.

Genetic Method. The genetic algorithm is an optimization technique, which was first proposed by Professor J.Holland of Micchigan University in 1975. It is the model of natural selection and genetic mechanism in the theory of biological evolution, in order to search the optimal solution.

Genetic algorithm can play a role in producing excellent offspring. After several generations of heredity, the offspring which satisfy the requirements will be obtained (the solution to the problem). This algorithm has the advantages of simple calculation and good optimization effect, and it has certain advantages in dealing with combinatorial problems.

Decision Tree Method. The decision tree method is to use a test function in the training set, building branches of a tree, a lower level node and branch according to different values, so that a decision tree can be generated. Finally, the decision tree is transformed into a rule, which can be used to classify things. The most influential and earliest decision tree algorithms in the world are the ID3 algorithm of Bayes and Quinlan. The ID3 algorithm can only handle discrete attributes. Many decision tree methods later are developed on the basis of the ID3 algorithm, such as the classic C4.5 algorithm.

Neural Network Method. The principle of neural network method is to simulate the neuron structure of human brain, which uses MP model and HEBB learning rule to establish 3 major kinds of neural network models: feedforward network, feedback network and self-organizing network.

Fuzzy Theory. The fuzzy mathematics focuses on the fuzzification of both this and that, which is a new breakthrough in the mathematics development history. Fuzziness is an objective existence. Zadeh's reciprocal principle says: to a problem, the higher the complexity is, the lower the significative accuracy is, which means stronger fuzzification.

Rule Induction. Rule induction is one of the common methods of intelligent e-commerce data mining. It is mainly used to discover the effects of some attributes on other attributes in the intelligent e-commerce database, represented with the probability. The classical algorithm, such

as the Apriori algorithm, will explain the algorithm in detail later.

Visualization Technology. Visualization technology is an auxiliary method. It is a more intuitive graphical way to display the mode of mining, and users understand the intelligent e-commerce data more clearly. The visualization of intelligent e-commerce data is becoming more and more important, which is a powerful method to reveal the status, inherent nature and regularity of intelligent e-commerce data.

The Common Types of Knowledge Found in Intelligent E-business Data Mining

Generalized Knowledge. Generalized knowledge refers to the generally described knowledge of the category characteristics. According to the microcosmic characteristics of intelligent e-commerce data, find the representational, universal, high-level concept, mesoscopic and macroscopic knowledge, reflecting the common nature of similar things, which is the generalization, refinement and abstraction of intelligent e-commerce data.

This kind of knowledge can be used to refine and summarize the overall concept and feature information of intelligent e-commerce data. The extracted knowledge can be presented directly to the user through the visualization technology, and generalization can provide the basic knowledge for other applications.

Association Knowledge. Association knowledge reflects the knowledge of dependence or association between an event and other events. If there is a dependency between intelligent e-commerce data items in an intelligent e-commerce database, one item can be applied to predict another item. The most classic one of the association rule method is the Apriori algorithm first proposed by R. Agawal.

Classification Knowledge. The purpose of classification knowledge is to classify intelligent e-commerce data into one of a series of known classes, reflecting the differential knowledge between different transaction s.

The most classic one in the classification is the classification method based on the decision tree. Its approach is through a supervised learning and training model. There is a supervision here to indicate that the model takes the category target as the output result.

Predictive Knowledge (Prediction). Predictive knowledge refers to the estimation of the missing values or attribute values existing in the intelligent e-commerce data set. In the intelligent e-commerce data of time series, the future intelligent e-commerce data is predicted through learning historical intelligent e-commerce data. It is actually the association knowledge of the intelligent e-commerce data in time series.

Deviant Knowledge (Deviation). Deviant knowledge is used to reveal abnormal phenomena that deviate from the routine. Sometimes the interestingness will discover information which deviates a lot from the object standard, for example, the detection of abnormal behavior in a credit card. There are many technologies for mining intelligent electronic commerce data mining, such as outliers of Web mining, deviations of predicted values, and so on.

Sequential Patterns. Sequential pattern refers to the common behavior pattern found in multiple intelligent e-commerce data sequences. For sequential intelligent electronic commerce database D, the discovering problem of sequential pattern is to find all frequent sequences or all the longest frequent sequences in the intelligent e-commerce database. R. Agrawal says the longest frequent sequence that satisfies a user's specified minimum support exists in customer - sequence intelligent e - commerce database.

Evolutionary Knowledge (Evolution). Evolutionary knowledge refers to the detection or evaluation of the evolution law of the intelligent electronic commerce data of a certain object whose behavior changes with time, which involves the characteristics, classification, association and Web mining of time-related intelligent e-commerce data.

Research on Association Optimization of Intelligent E-Commerce Web Mining Overview of Association Rules

Association rule is a very important technology in the field of intelligent e-commerce data mining, and is also an important field in the research of intelligent e-commerce data mining. It is a rule that represents a certain relation between a set of objects in an intelligent e-commerce database. The most important process of association rules is the generation of frequent sets.[3]

Association rules were put forward by Agrawal, Imieliski and Swamiu in 1993, which are used to discover the relation among different commodities in the intelligent e-commerce database.

The Optimization Study of Association Rule Algorithm. Set $I = \{i_1, i_2, i_3, ..., i_n\}$ as the item set of the intelligent e-commerce database (item set: called K set because of containing k items), the elements of which are the items in the intelligent e-commerce database. D is a collection of transaction s, where each transaction T is a set of items, $T \subseteq I$, and transaction T contains X when and only when $X \subseteq T$.

Association rules are the implication form of $X \Rightarrow Y$, of which $X \subseteq I, Y \subseteq I$, and $X \cap Y = \varnothing$. X becomes the first component of the rule (condition), and Y is the second component (result) of the rule. The interestingness of association rules can be measured with two important parameters, support and confidence

Definition 1. The degree of support for item set X contains the ratio of the transaction quantity of item set X to the total transaction quantity in D. Support is generally simplified as sup.

$$sup(X) = \left|\{T \in D, and X \in T\}\right| / \left|\{T \in D\}\right| \tag{1}$$

The support degree of rule $X \Rightarrow Y$ in transaction D is the ratio of the transaction number of the transaction set D to the transaction set of X and Y and all the transaction sets.

$$sup(X \Rightarrow Y) = sup(X \cup Y) = \left|\{T : X \cup Y \subseteq T, and T \in D\}\right| / \left|D\right| \tag{2}$$

Definition 2 The confidence degree rule $X \Rightarrow Y$ in the transaction set D refers to the ratio of the transaction quantity containing X and Y to the transaction quantity containing X. Confidence is generally simplified as conf

$$conf(X \Rightarrow Y) = sup(X \cup Y) / sup(X) \tag{3}$$

Support and confidence are two important concepts in describing association rules. Support is used to measure the occurrence frequency of item set of the association rules in the whole transaction intelligent e-commerce database, and confidence is used to measure the credibility of association rules. Generally speaking, users are most interested in association rules with high occurrence rate (high support) and high credibility (confidence).[4] When a transaction set D is given, the user sets a minimum support and minimum confidence, which are respectively called the minimum support threshold (minsup) and the minimum confidence threshold (minconf). The association rule satisfying the minimum support and minimum confidence is called the strong association rule and vice versa. The task mined by association rules is the strong association rule, which is presented to the user.

Study on the Category Optimization of Association Rules. Next, the association rules will be classified in different cases:

According to the different types of intelligent e-commerce data, association rule can be divided into Boolean association rule and numerical association rules. The values dealt by Boolean association rules are all discrete, while the data fields of intelligent e-commerce data mined by numerical association rule can not only handle discrete ones, but also deal with numerical types.

For example, the most classic rule of association rules: Diaper $\Rightarrow$ beer is Boolean association rule, while age (40~45) $\wedge$ year income（≥50,000）$\Rightarrow$ loan（150,000～200,000） is numerical association rule.

According to the different level of abstraction involved in the rules, it can be divided into single layer association rules and multi-layer association rules.[5]

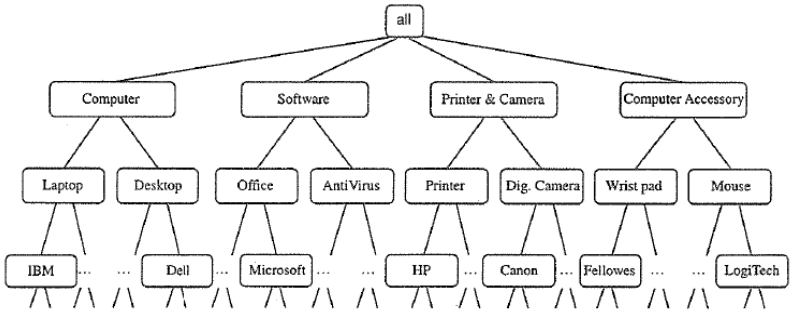

Figure 1. Concept hierarchy optimization of computers commodities

Single-layer association rules do not consider the hierarchy of intelligent e-commerce data, while multi-layer association rules take full account of the hierarchy of intelligent e-commerce data. [6] The above shows the concept of a computer commodity. The first layer includes computer, software, print camera, and computer accessory, and the second layer includes laptop computer, office software, antivirus software and so on. The association rules between the same layers are single-layer association rules, for example, IBM computer⇒HP printer is a single-layer association rule, while IBM computer⇒Printer is a multi-layer association rule, which involves a multilevel association rule between a high level and a detail level. [7]The higher the level of the association rules is, the easier it is to dig, and the lower it is sometimes hard to dig. This is because the lower the level is, the lower the support of the intelligent e-commerce data items is.

According to the different dimensions of the intelligent e-commerce data involved in the rules, it can be divided into single-dimensional association rule and multidimensional association rule. [8]

Using the terms used in a multidimensional intelligent e-commerce database, each different predicate is called dimension. For example, buys (X, “IBM computer”)⇒buys (X, “HP printer”) is a single-dimensional association rule with only one single predicate. Sometimes different kinds of intelligent e-commerce data may be involved, such as intelligent e-commerce database contains not only the quantity, price, address of the sales shop, but also other information such as customer's age, occupation and so on. [9]Association rules of multiple predicates can be mined, such as: age(X,“20～29”)∧occupation(X, “student”) ⇒ buys(X, “laptop”),this type of association rule is multidimensional association rule.

Summary

Web mining analysis is mainly to carry out Web mining or classification based on the characteristics of things. Web mining methods include traditional multivariate statistical analysis Web mining method, the fuzzy Web mining and neural network Web mining mentioned in the previous article. [10] The commonly used CLARAWeb mining method in statistics is developed on the basis of CLARA and PAM. This method has a fixed sample at each search stage. Some scholars have also made improvements to this method and randomly select samples at each search stage. The objectivity of the search is enhanced, which improves the quality of Web mining to some extent. Web mining technology and pattern recognition are also one of the most important technologies in intelligent e-commerce data mining. [11]

Reference

[1] Huang Weijian, Sang Zhichao, Du Wei. Architecture Design of Web Data Mining System under the Environment of E-commerce [J]. Journal of Hebei University of Engineering, 2014(6): 83-85.

[2] Wu Siyuan. The Application of Web Data Mining Technology in E-commerce [J]. Intelligent

Technology. 2016(12): 96
[3] Zhang Xiaobing. The Effective Combination of Database and Web Mining Technology in E-commerce [J]. Network Security Technology and Application.2015 (4): 30
[4] Hou Dongxiu .Design and implementation of personalized learning platform based on Web log mining, [D], Shandong Normal University, 2017-06
[5] Wang Yan. Analysis of Knowledge Mining Methods in Multimedia Database [J]. Library Science Research; 2013(12)
[6] Zhou Qifeng, Zhang Lichen. The Application Data Mining in the Experimental System [J]. Technology Information; 2013(14)
[7] Li Chunya. Research on Web log analysis based on data mining -- Taking YF website as an example, [D], Wuhan University,2016-11
[8] Liu Guoxiong. Research on the Key Technology of Web Mining Based on E-commerce [J]. Hubei Agricultural Mechanization; 2017(5): 69
[9] Zhu Beifang. Analysis of the Application of Web Mining Technology in E-commerce [J]. Jiangsu Scientific and Technological Information; 2016(8): 57-58
[10] Ma Zong ya, Zhang Huiyan.Web data mining technology in the application of electronic commerce research [J]. modern economic information, 2014(6):23-24.
[11] Zhang Suzhi, Qu Xukai, Zhang Lin. Research on Web data mining based on e-commerce research [J]. modern computer,2015(6):12-18