

Data Mining and Business Process Management of Apriori Algorithm

Qiu Suzhen

NanNing University, GuangXi, China,530200

Abstract. The classical association rule data mining algorithm Apriori algorithm is widely used in various fields. Through the analysis and mining of data relevance, the information extracted has important reference value in the decision-making process. The continuous promotion and application of information technology, how to make full use of this data information to provide decision support for decision maker in various industries have become an urgent and thorny issue. In addition to using the existing relational database standard query statements to obtain general and intuitive information, It is necessary to mine the data relationships that are contained, unknown, and actually exist. The famous Apriori algorithm is an algorithm for mining association rules. In this paper, through the basic idea of Apriori algorithm, the embedded data relationship is mined and Apriori algorithm is implemented.

1. Introduction

Data Mining is the extraction of information that is hidden in the prior, but potentially useful, from a large number of incomplete, noisy, fuzzy, and random application data. Synonyms similar to data mining include data fusion, data analysis, and decision support. This definition includes several layers of meaning: the data source must be real, large, and noisy; the knowledge that is of interest to the user is discovered; the knowledge found is acceptable, understandable, and usable; it is not required to be discovered. A universal knowledge that only supports specific discovery problems.

2. Data mining definition

In a broad sense, data and information are also forms of knowledge, but people also regard concepts, rules, patterns, laws, and constraints as knowledge. People see data as a source of knowledge, as if mining or gold mining from ore. Raw data can be structured, such as data in a relational database; it can also be semi-structured, such as text, graphics, and image data; or even heterogeneous data distributed over the network. The method of discovering knowledge can be mathematical or non-mathematical; it can be deductive or inductive. The discovered knowledge can be used for information management, query optimization, decision support and process control, etc., and can also be used for data maintenance. Therefore, data mining is an interdisciplinary subject that promotes the application of data from low-level simple queries to mining knowledge from data and providing decision support. Under the demand of this kind of demand, researchers from different fields, especially database technology, artificial intelligence technology, mathematical statistics, visualization technology, parallel computing and other scholars and engineers, have joined the emerging research of data mining. The field forms a new technology hotspot.

Data mining is a new business information processing technology. Its main feature is to extract, transform, analyze and other model processing of a large amount of business data in a commercial database, and extract key data for assisting business decision-making. In short, data mining is actually



a deep class of data analysis methods. Data analysis itself has been around for many years, except that in the past, the purpose of data collection and analysis was for scientific research. In addition, due to the limitations of computing power at the time, complex data analysis methods for analyzing large data volumes were greatly limited. . Now, due to the realization of business automation in various industries, the business sector generates a large amount of business data, which is no longer collected for analytical purposes, but is generated by Opportunistic business operations. The analysis of these data is no longer just for research needs, but more importantly to provide real valuable information for business decisions, and then to obtain profits. But a common problem faced by all enterprises is that the amount of enterprise data is very large, and the information that is truly valuable is very small. Therefore, after deep analysis from a large amount of data, information that is beneficial to business operations and competitiveness is obtained.

Therefore, data mining can be described as: exploring and analyzing a large amount of enterprise data according to the established business objectives of the enterprise, revealing hidden, unknown or verifying the known regularity, and further modeling the advanced and effective methods.

DM cannot tell you the actual value of a model for your business. DM is a tool. It just helps business people analyze data more deeply and easily, but can't tell you the actual value of a model for your business. The resulting model must be validated in real life, and DM will not automatically discover the model without guidance. Data analysts must know how the DM tools you choose work, and what the principles of the algorithms are. DM will never replace the role of experienced business analysts or managers, it is just a powerful tool.

3. Introduction to association rules

If a transaction contains an X, then the transaction is likely to contain Y. The concrete form is $\{X\} \rightarrow \{Y\}$, which can usually be described as: When a customer buys something {pen} in a transaction (here X="pen"), it is likely that he also purchased {ink} (here Y= "ink"), this is the association rule. In the United States, there is a saying that "urine is not wet" and "beer" are often purchased together. This kind of statement has certain practical significance: 1) It may be that people who drink beer regularly in this age group just start raising children in the family; 2) Perhaps because beer is too much, it needs to be diaper. However, if there is no application of the association rules in data mining here, you can't imagine such a surprising "joke" anyway. The Business Process Management workflow is as follows.

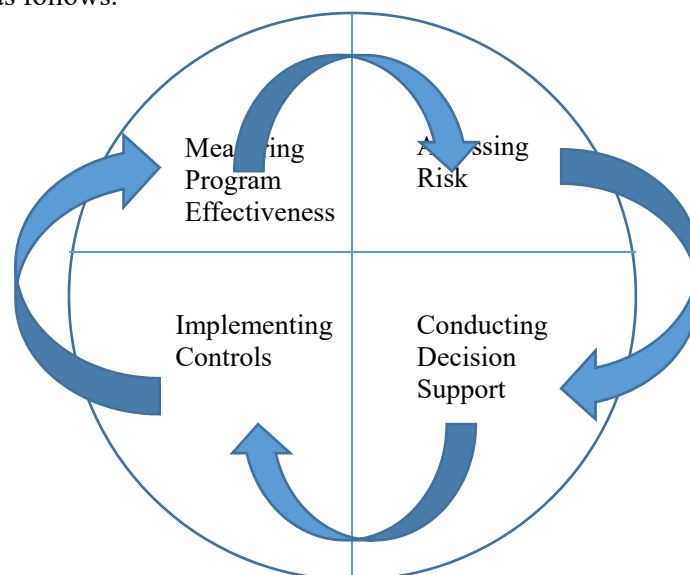


Fig.1 Business Process Management workflow

4. Apriori algorithm background introduction

A typical example of association rule mining is shopping basket analysis. Market analysts need to find the relationship between the different items that customers put into their shopping baskets from a large amount of data. If a customer buys milk, how likely is he to buy bread? What product groups or collection customers will buy at the same time in one purchase? For example, 80% of customers who buy milk also buy bread, or 70% of customers who buy hammers also buy nails, which is the association rule extracted from the basket data.

The problem of mining association rules is described as follows:

Let: $I = \{i_1, i_2, \dots, i_m\}$ is a collection of all items. D is a collection of all transactions (ie database), each transaction T is a collection of some items, T is contained in I , Each transaction can be identified by a unique identifier TID . Let X be a collection of certain items. If X is contained in T , then transaction T is said to contain X , and association rules are represented as follows (X is included in T) = The implication of $\Rightarrow$ (Y is contained in T), where X is contained in I , Y is contained in I , and $X \wedge Y = \Phi$. Its meaning is the appearance of certain items in a transaction, and other items can be derived Also appears in the same transaction (for simplification, denote (X is contained in T) $\Rightarrow$ (Y is contained in T) as $X \Rightarrow Y$, where ' $\Rightarrow$ ' is called 'association' operation, and X is called association rule Prerequisites, Y is called the result of the association rule).

The rule $X \Rightarrow Y$ in the transaction set D is constrained by the support s (support) and the confidence c (confidence), the confidence indicates the strength of the rule, and the support indicates the frequency of occurrence in the rule. The support degree $s(X)$ of the data item set X is the ratio of the number of transactions including X in D to the total number of transactions of D , but for convenience of description below, the support of the data item set X is the number of X s contained in the database D . Business Process Management element analysis as shown below.

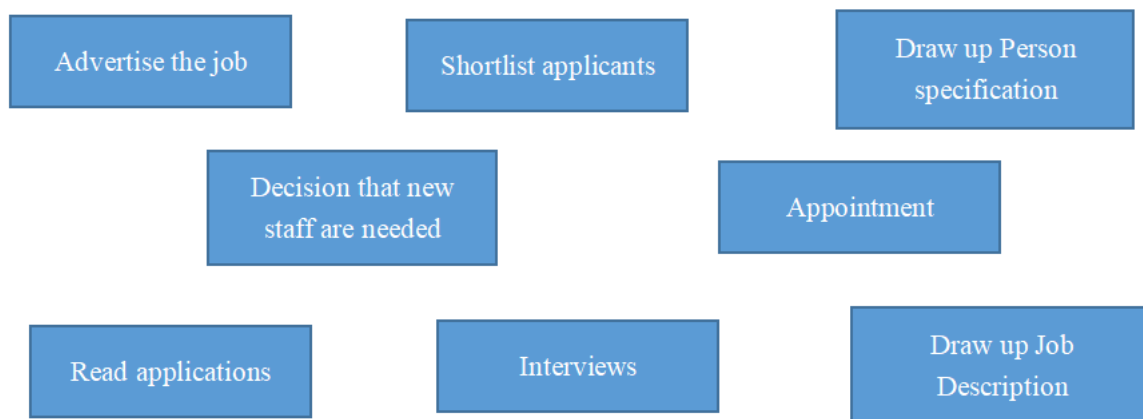


Fig.2 Business Process Management element analysis

The support s of the rule $X \Rightarrow Y$ is defined as: The proportion of the firm containing $X \cup Y$ in D is $s\%$, which represents the ratio of the number of transactions containing both X and Y to the total transaction volume of D ; Rule $X \Rightarrow Y$ The confidence c of Y is defined as: In D , $c\%$ of transactions contain X and also contain Y , indicating how likely it is to contain Y in the transaction containing X in D . Minimum support threshold minsupport indicates that the data item set is in statistics Minimum significance in the sense. The minimum confidence threshold minconfidence represents the minimum reliability of the rule. If the data item set X satisfies $X.\text{support} \geq \text{minsupport}$, then X is the set of big data items. The minimum confidence threshold is generally given by the user and The minimum support threshold. The rule with confidence and support greater than the corresponding threshold is called strong association rule, and vice versa is called weak association rule. The task of discovering the association rule is to find those confidences and support sizes equal to the given value from the database. Strong rules.

5. Description of the Apriori algorithm

5.1 Description of the Apriori algorithm

In the Apriori algorithm, the basic idea of finding the largest set of items is: The algorithm needs to perform multi-step processing on the data set. The first step is to simply count the frequency of occurrence of all items containing an element and find out those that are not less than the minimum support. The project set, that is, the one-dimensional largest project set. The loop process from the second step until no maximum project set is generated. The loop process is: In step k, the (k-1) dimension maximum project generated according to step k-1. The set generates a k-dimensional candidate item set, and then searches the database to obtain the item set support degree of the candidate item set, and compares it with the minimum support degree.

For the convenience of the following description, the agreement is as follows:

1. Items in the database transaction are arranged in alphabetical order. Each item is identified by $\langle \text{TID}, \text{item} \rangle$, where TID represents the identifier of the corresponding transaction, and item represents the item name.
2. The number of items in each item set is called the size of the item set. When the size of the item set is $\text{size}=k$, the item set is called k-itemset (k-dimensional item set).

5.2 Description of the Apriori algorithm

The first step in the Apriori algorithm is to simply count the frequency of occurrences of all item sets containing one element to determine the largest set of one-dimensional items. In the kth step, in two stages, the candidate item set C_k is generated by the maximum item set L_{k-1} generated in the (k-1) step. Then the search database calculates the support level of the candidate item set C_k .

6. Summary

The first step in the Apriori algorithm is to simply count the frequency of occurrences of all item sets containing one element to determine the largest set of one-dimensional items. In the kth step, in two stages, the candidate item set C_k is generated by the maximum item set L_{k-1} generated in the (k-1) step. Then the search database calculates the support level of the candidate item set C_k . In order to calculate the support of the project in C_k more quickly, the function `count_support` is used to calculate the support.

References

- [1] Elisa Giacosa, Alberto Mazzoleni, Antonio Usai. Business Process Management (BPM)[J]. Business Process Management Journal, 2018, 24(5).
- [2] Renata Gabryelczyk. Does Grade Level Matter for the Assessment of Business Process Management Maturity[J]. Naše gospodarstvo/Our economy, 2016, 62(2).
- [3] Vesna Bosilj Vukšić, Dalia Suša Vugec, Anita Lovrić. Social Business Process Management: Croatian IT Company Case Study[J]. Business Systems Research Journal, 2017, 8(1).
- [4] Aleksandra Stoiljković Radelović, Radenko Milojević, Marija Radosavljević. Identification and Analysis of Key Business Process Management Factors[J]. Economic Themes, 2018, 56(1).
- [5] Monika Malinova, Jan Mendling. Identifying do's and don'ts using the integrated business process management framework[J]. Business Process Management Journal, 2018, 24(4).
- [6] Renata Brajer-Marczak. Elements of knowledge management in the improvement of business processes[J]. Management, 2016, 20(2).