Research on the Decision-making Model of Value Information Selection in the Context of Big Data

Daqi Dai*
Xi’an Siyuan University, China, 710038

*Corresponding author e-mail: tiaquxwl172@163.com

Abstract. Data filtering is the first step in the utilization and development of big data, that is, obtaining value data from massive data resources, building system operation database, and realizing the mining and sorting of value data. However, restricted by the infinite big data environment and limited data acquisition ability, almost no technology can obtain all the value data. In this paper, a series of indicators are proposed to evaluate and enhance the data filtering algorithm.

Keywords: Decision-making Model, Value Information; Information Selection, Big Data

1. Introduction
With the rapid accumulation of data in big data environment, it is very important to screen out valuable data in order to analyze the value of massive data. Data filtering plays an important role in the whole process of data processing. The purpose of data filtering is to improve the availability of relevant data collected and stored before, which is more conducive to later data analysis. Data filtering includes data extraction, data cleaning and data loading. With the rapid accumulation of data in big data environment, it is very important to screen out valuable data in order to analyze the value of massive data [1]. Data filtering plays an important role in the whole process of data processing. With the increase of data volume and the complexity of structure, the data filtering for big data will inevitably consume more resources, so it is necessary to quickly and accurately filter out valuable data and select appropriate algorithm. Each algorithm has its own environment for use. The increase of data complexity in big data environment makes it difficult to select the appropriate algorithm [2]. Secondly, the huge amount of data in big data makes it more and more difficult to analyze valuable data through a single algorithm.

2. Basic model of data filtering
In the massive big data environment, no organization or individual can use any method to obtain all data, and no matter what technology is used, the amount of data obtained is a very small part of the big data environment [3]. However, the final data obtained through the selected technology of data
filtering does not necessarily coincide with the expected data set, so there are a series of algorithms to evaluate the efficiency of data filtering, as shown in Figure 1.

Especially under the technical conditions, the value data and the selectable data do not necessarily coincide. That is to say, under all the technical conditions of data filtering institutions or individuals, they do not necessarily get all the target data they need, and this state is very likely to occur.

![Figure 1. Data Screening Method and Result Evaluation Diagram](image)

For data selection results, the final data obtained by the data filter is called selected data, which is recorded as data set [A], the expected data is called value data, which is recorded as data set [O], the intersection of the two data sets is called hit data, which is recorded as data set [M], and the maximum data set that can be obtained under the technical conditions is called selectable data, which is recorded as data set [B], which is recorded as value data set other than the selected data set is called outside data and is recorded as data set [N]. Therefore, [M] / [A] is the filtering efficiency of data filtering, 1 - [N] / [O] is the filtering ability of data, [M] / [O] is the hit rate of data filtering, and [A] / [B] is the load rate of data filtering. The filtering efficiency directly reflects the accuracy of the filtering algorithm, the filtering ability reflects the channel strength of user's data set construction, the hit rate reflects the application value of the final filtering result, and the load rate reflects the filtering ability. The proportion of data channels used by the algorithm [4].

Therefore, the ultimate goal of all data filtering algorithms is to get the highest hit rate under the condition of minimum load rate.

3. Steps of Data Filtering

3.1. Data extraction

Data is finally put into the data warehouse after extraction, so the main task of data extraction is to transfer data from different data sources into the data warehouse according to the data format in the data warehouse, and its main task is to unify the data format. Different data sources use different database types, so data extraction can be roughly divided into two situations.

3.2. Data cleansing

Data cleaning includes four parts: missing data processing, repeated data processing, abnormal data processing and inconsistent data processing. This part is the first step of direct data processing, which directly affects the results of subsequent processing, so it is very important. Data missing is a common situation in database, but in order to get a complete information table for data mining, we must solve the problem of data missing. The duplicate data includes not only the real duplicate data, but also the attribute redundancy and the attribute data redundancy. For the real value or attribute value duplicate data processing is also relatively simple, can be directly deleted. But attribute redundancy and attribute data redundancy need to be analyzed and then deleted. Abnormal data is also a common situation, which means that some data in the data set is very different or inconsistent with other data. The difference does not mean that the data must be abnormal. These special data may also reflect the actual
situation.

3.3. Data loading
In the process of data loading into the database, there are two ways: full load and incremental load. Full load refers to the method of data loading after deleting the whole table; incremental load refers to that the target table only updates the data changed in the source table. From a technical point of view, full load is much simpler than incremental load. Generally, you only need to empty the target table and import the source table data in full before data loading. However, due to the requirement of data volume, system resources and real-time data, we need to use incremental loading mechanism in many cases.

4. Algorithm of Data Filtering

4.1. Analysis of classification algorithm
Classification data mining is to find out the same attributes of common things and the differences between different things. Use the same or different points to classify things. The advantage of decision tree is that its description is simple and it can quickly classify the data when the amount of data is large. The classification algorithm is usually based on decision tree. The set classification types are represented by leaf nodes, and the middle nodes are used to represent the attributes of things.

4.2. Clustering algorithm analysis
Clustering algorithm is used to group things with the same characteristics, also known as group analysis. Clustering algorithm can be used to roughly judge how many groups the objects are divided into and provide the eigenvalues of each group of data. In clustering analysis, given instances can be divided into different categories. The instances in the same category are related, but not to the categories are not related. The important part of clustering algorithm is the classification step. When classifying a given instance, you need to select any sample as the sample center, and then select the center distance. The instances less than the center distance are classified into one set, and the remaining ones greater than the center distance are classified into another set.

4.3. Analysis of association algorithm
Association algorithm is used to express the relationship or dependence between two things. There are two kinds of relations between things, one is called relevance, the other is called relevance. Both of them are used to represent the relevance between things, but the former is usually used to represent the relevance of Internet content and documents, and the latter is usually used to represent the relationship between the commodities of various websites between e-commerce, but there is no essential difference between them.

5. Summary
The biggest difference between data filtering in big data environment and traditional data filtering lies in huge data volume and complex data structure. In the traditional data filtering process, due to the limited amount of data, the data can be processed according to the needs by using high-performance computer or simple online analysis. But in the big data environment, traditional online analysis has encountered a series of problems. The massive data in big data has huge data value, but the data availability efficiency in big data is not high, which increases the difficulty of algorithm selection. Because of the huge amount of data, every filtering for big data will consume huge resources.

References
[1] Mehdi Toloo, Mahnaz Mirbolouki. A new project selection method using data envelopment analysis[J]. Computers & Industrial Engineering, 2019, 138:.
[2] C.S. Saravana Kumar, R. Santhosh. Effective information retrieval and feature minimization technique for semantic web data [J]. Computers and Electrical Engineering, 2020, 81:.

[3] Information Technology - Data Systems; Researchers from Urmia University of Medical Sciences Report Recent Findings in Data Systems (Landfill Site Selection Via Integrating Multi-criteria Decision Techniques With Geographic Information Systems: a Case Study In Naqadeh, Iran) [J]. Computers, Networks & Communications, 2019:.

[4] Yan Li. A Comparative Study of Tourism Business Models in the Internet Era. International Journal of Frontiers in Sociology (2019), Vol. 1, Issue 1: 99-108.