An Overview of the MapReduce Model

S. Rajeswari, K. Suthendran(✉), K. Rajakumar, and S. Arumugam

Kalasalingam University,
Anand Nagar, Krishnankoil 626126, Tamil Nadu, India
s.rajeswari30@gmail.com, k.suthendran@klu.ac.in, rajakumarkk@gmail.com,
s.arumugam.klu@gmail.com

Abstract. Data is getting accumulated fast in various domains all over the world and the data size varies from terabytes to yottabytes. Such huge size data are known as Big Data. Extraction of meaningful information from raw data using special patterns are called Data Mining and sophisticated algorithms have been designed for this purpose. In this paper, the time complexity for MapReduce-based data mining algorithm is presented.

Keywords: Big Data · MapReduce · Hadoop · Datamining · Time complexity

1 Introduction

The size of data, storage capacity, processing power and availability of data are rising day by day. The traditional data warehousing and management systems tools are not capable of dealing with huge data. The analytic methods used to deal with huge data is referred to as Big Data analytics. Doug Laney was the first one to discuss about 5 V's in Big Data management [1], namely, volume, variety, velocity, variability and veracity.

Some of the technologies to process big data are Hadoop HDFS, MapReduce, Hive, HBase, Pig, Flume, Hadoop Mahout, Windows Azure etc. [1]. MapReduce was first developed by Google but now it is incorporated by the Apache. In Parallel, the large clusters of hardware are available to process huge amount of data. It has two functions Map() and Reduce(). Map() is used for filtering and sorting the data and Reduce() is used for summarizing the data [2] (Fig. 1).

The MapReduce process is depicted in Fig. 1. Here the input is application specific while the output is a group of $<$ key, value $>$ pairs produced by the Map function. In the mapping process, each single key-value input pair (key, value) is mapped into several key-value pairs: $[(l_1, x_1), \ldots, (l_1, x_r)]$ with same key, but different values. These key-value pairs are used for reducing the functions.

Data mining techniques are applied on raw data for extracting useful information. The process of finding a model is to describe the data classes by using a classification algorithm [3].

Before using the classification algorithm, preprocessing steps such as categorization and feature selection are included. This process is helpful for getting an

S. Arumugam et al. (Eds.): ICTCSDM 2016, LNCS 10398, pp. 312–317, 2017.
DOI: 10.1007/978-3-319-64419-6_40

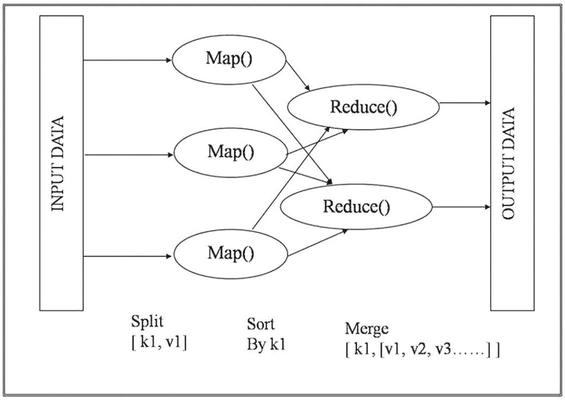

Fig. 1. MapReduce process

improved accuracy in the prediction. Categorization is the process of converting the data into a categorical format. Based on the condition, the data is categorized into a standard format. Feature selection is used to select the important features of the data and to remove the irrelevant attributes. MapReduce concept is used in the categorization and feature selection process.

Classification algorithm involves two steps, namely training set and testing set. The training set is used to build a model with the training data. Testing set is applied on the classification model and is used to check the accuracy [4].

A Decision tree is a classification model. It is mainly used to classify an object to a predetermined class. CHAID, CART, ID3, C4.5, C5.0 are decision tree algorithms and C5.0 is a widely used decision tree algorithm.

C5.0 algorithm handles continuous and categorical values. Feature selection is the basic step to construct a decision tree. The decision tree algorithm C5.0 is used to access the data and has higher speed when compared to ID3 and C4.5 [5]. MapReduce process is used to evaluate the data.

In this paper we discuss the time complexity for MapReduce – based C5.0 algorithm. MapReduce technique is a training model to be linked with performance for processing huge data sets. MapReduce concept runs on a large cluster of product machines and is highly scalable [6].

Matei Zaharia et al. proposed an improved scheduling algorithm that decreases the Hadoop reply time and the performance by using MapReduce [7].

2 Implementation Scrutiny

MapReduce with Datamining algorithm is used for tera bytes of data [8]. Using the categorization algorithm, attribute values can be grouped. Among categorized data, relevant attribute has to be picked up. For this purpose, feature selection algorithm can be used. These two algorithms are used for the prediction of class labels. We fix one example for the purpose of illustration of all the concepts presented here.

Example 1. Training data set and testing data set for prediction of mode of transport are given in Tables 1 and 2 respectively.

Table 1. Training dataset

Gender	Car owner	Travel cost	Income level	Transportation mode
Male	0	50	10000	Bus
Male	1	50	50000	Bus
Female	1	50	50000	Train
Female	0	50	10000	Bus
Male	1	50	50000	Bus
Male	0	500	50000	Train
Female	1	500	50000	Train
Female	1	1000	100000	Car
Male	2	1000	50000	Car
Female	2	1000	100000	Car

Table 2. Testing dataset

Gender	Car owner	Travel cost	Income level	Transportation mode
Male	1	Standard	High	?
Male	0	Cheap	Medium	?
Female	1	Cheap	High	?

2.1 Categorization

Categorization is the process of grouping objects into categories, usually for some specific purpose. Generally, categorization algorithm has two parts. First is Map part which is used to check the conditions of the attribute values and the second is reduce part which changes the numerical values to categorized values [9].

The attribute values are changed into categorized format based on the conditions. Finally the categorical values are stored in a new file.

For the case study given in Tables 1 and 2, for the categorization process, the travel cost and Income level attributes are used and the conditions are as follows.

Condition for travel cost	Condition for income level
if Travel cost == 500 then	if Income == 100000 then
Standard	high
else if Travel cost < 500 then	else if Income < 50000 then
cheap	low
else	else
Expensive	medium
end	end

2.2 Feature Selection

Feature selections is also called attribute selection or variable selection. After categorization, the feature selection process is used to select a subset of relevant attributes [10]. Chi-square (χ^2) feature selection based on MapReduce concept is used for finding the relevant attributes. It is one of the popular feature selection methods. Statistical test is also used to decide whether observed frequencies are much dissimilar from expected frequencies [11]. The χ^2-filter is defined by

$$\chi^2 = \frac{\Sigma(O - E)^2}{E} \tag{1}$$

where for each attribute, O is the Observed Frequency and E is the Expected Frequency.

The weight for each attribute is calculated by using (1). Attributes whose weight is greater than a chosen threshold value are taken for further processing.

2.3 C5.0 Classifier

Classification is one of the major components in data mining and C5.0 is decision tree based classification algorithm. It can handle continuous values, categorical values and numerical attributes. Let $\mathcal{C}$ be a categorization of a set S of objects into categories $C_1, C_2, \ldots, C_r$. Let p_i be the probability that an object in S is in category C_i. The entropy of S is defined as

$$\text{Entropy}(S) = -\sum_{i=1}^{r} p_i \log_2 p_i.$$

The entropy measures the homogeneity of objects.

To determine the best attribute to be chosen for a node in a decision tree, we use the concept of information gain. For any given attribute A, consider the set $V(A)$ of possible values of A. For any $v \in V(A)$, let S_v be the set of all elements of S having value v for the attribute A. The information gain of A with respect to S is given by

$$\text{Inf.Gain } (S, A) = \text{ Entropy } (S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \text{ Entropy } (S_v).$$

Thus the information gain measures the expected reduction in the entropy.

The attribute with highest information gain is taken as the root node in the decision tree and the values of the attribute are its children nodes. Then using the remaining attributes, the attribute with highest information gain is taken as a child node and the process is repeated. If $S(V_i) \neq \emptyset$, the tree constructed is added as a new branch at v_i. Each leaf node of the final decision tree gives a rule for predicting the class label.

For the example given in Table 1, the probability of an individual travelling by bus, car and train are respectively 0.4, 0.3 and 0.3. Hence the entropy value is $-(0.4 \log_2(0.4) + 0.3 \log_2(0.3) + 0.3 \log_2(0.3)) = 1.571$.

For each of the attributes Gender, Car-owner ship, travel cost and income the information gain in computed and the results are given in Table 3.

Table 3. Information gain result

Attribute	Information gain
Gender	0.125
Car owner	0.21
Travel cost	1.21
Income level	0.695

The attribute with highest gain value is travel cost which is taken as the root node and its branches are its values, namely, standard, expensive and cheap.

The next step computation of informations gain for the removing three attributes is carried out. Table 4 gives the results for the attribute cheap.

Table 4. Information gain for the node "cheap"

Attribute	Information gain
Gender	0.322
Car owner	0.171
Income level	0.171

In this way the process can be continued, giving the decision tree. From the decision tree teh transportation mode can be predicted in terms of the attributes. For example for the object with attributes female, car owner, cheap the predicted transportation mode is bus.

3 Conclusion

The input for the MapReduce model is a set S of objects with $|S| = n$, attributes $A_1, A_2, \ldots, A_k$ where the attribute A_i takes α_i values and rules for categorization. The complexity of categorization process is $O(n)$. The complexity of feature selection is $O(k)$. The complexity of C5.0 classifier is $O(n^2)$. Thus the overall complexity of the algorithm is $O(n^2)$.

Acknowledgment. The first author is thankful to the management of Kalasalingam University for providing fellowship.

References

1. Miranda Lakshmi, T., Martin, A., Mumtaj Begum, R., Prasanna Venkatesh, V.: An analysis on performance of decision tree algorithms using students qualitative data. Int. J. Mod. Educ. Comput. Sci. **5**, 18–27 (2013)
2. Vaitheeswaran, G., Arockiam, L.: Big data for education in students perspective. Int. J. Comput. Appl. Adv. Comput. Commun. Tech. High Perform. Appl. **4**, 11–17 (2014)
3. Laney, D.: 3-D Data management: controlling data volume, velocity and variety. META Group Research Note, pp. 20–25 (2011)
4. Ladha, L., Deepa, T.: Feature selection methods and algorithms. Int. J. Comput. Sci. Eng. **3**, 1787–1790 (2011)
5. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
6. Zaharia, M., Konwinski, A., Joseph, A.D., Katz, R., Stoica, I.: Improving MapReduce performance in heterogeneous environments. In: Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation, pp. 29–42 (2008)
7. Choura, R.: Use of data mining techniques for the evaluation of student performance: a case study. Int. J. Comput. Sci. Manag. Res. **1**(3), 425–433 (2012)
8. Naseriparsa, M., Bidgoli, A.M., Varaee, T.: A hybrid feature selection method to im-prove performance of a group of classification algorithms. Int. J. Comput. Appl. **69**(17), 28–35 (2014)
9. Azhagusundari, B., Thanamani, A.S.: Feature selection based on information gain. Int. J. Innov. Technol. Explor. Eng. **2**(2), 18–21 (2013)
10. Vanaja, S., Ramesh Kumar, K.: Analysis of feature selection algorithms on classification: a survey. Int. J. Comput. Appl. **96**(17), 29–35 (2014)
11. Yadav, S.K., Pal, S.: Data mining: a prediction for performance improvement of engineering students using classification. World Comput. Sci. Inf. Technol. J. **2**(2), 51–56 (2012)