Application and Exploration of Computing Architecture Based on Big Data Platform

Yan Hou

Qilu Normal University, Jinan, Shandong, 250014, China.

122365871@qq.com

Abstract. To discuss how to provide a reliable platform for processing big data, by studying some sub projects in core technology and its ecological environment, an extensible big data application system architecture used for processing large scale data sets is implemented. In addition, the architecture is organized into four levels in the form of hierarchical model: platform layer, data layer, business logic layer, and application layer. From the angle of hierarchy, the optimization of Hadoop and HBase, MapReduce programming mode, Mahout technology details and clustering algorithm are discussed. In combination with the customer value hierarchy model and neural network classification model, a distributed brand value model is proposed and successfully deployed on the framework application layer. As a result, it improves the research and implementation of the whole system architecture. Finally, the research contents of each layer are summarized, and the future research directions of the corresponding parts are discussed.

Keywords: Hadoop; UGC; system architecture; hierarchical model; neural network; customer value.

1. Introduction

The rise of SINA micro-blog has brought about the emergence of a large number of UGC (User Generated Content). On this basis, there is an emerging UGC processing system. The government examines the effectiveness and public opinion of the policy according to the users' messages and comments; the public institutions track the effect of services and optimize their own services according to the users' response; online retailers try to mine the related information in the user's interest map, so as to push the products. The emergence and development of SINA micro-blog greatly stimulated the sensitivity of the academic and industrial circles to the UGC, and also made data mining and machine learning technology paid more attention to by each industry again. However, in this upsurge of UGC value mining, a name is even more frequent than UGC and data mining, that is, big data. Many people predict that the next 10 years or even longer will be "the age of big data". Google has always been a leader in information processing technology, and it is also one of the first companies to face big data impact. Google proposed a MapReduce processing model at the Operating System Design and Implementation (OSDI) conference in 2004. This model together with the Google File System (GFS) proposed in the previous year constitute a distributed processing system for big data of Google. With the openness of MapReduce and GFS technology, Hadoop came into being. Hadoop is an open source implementation of MapReduce and GFS, which builds a complete ecological environment with its multiple subprojects to provide a distributed big data solution [1].

The content of micro-blog is of great commercial value. Micro-blog is different from the previous network UGC. Micro-blog content is generally organized by topic-centered. Its content structure is more compact and its topic is more relevant. This is why more and more organizations and enterprises are concerned about micro-blog. In the context of such a large amount of data in micro-blog, the traditional data analysis model has not been able to undertake such a scale of data analysis task. It is of great significance how to build a reliable and high-performance big data application system architecture under the micro-blog.

2. State of the Art

The MapReduce framework in Hadoop includes a Jobtracker and a certain number of rTasktracke, and Jobtracker is generally running on the host the same as the name node. With the development of Hadoop, its ecological environment is growing and maturing. Kong et al. [2] used an item-based collaborative filtering algorithm to creatively propose a strong LOD (Link Open Dataset) method to link the user's interest and social relations, so as to improve the accuracy of the recommended results. Guo et al. [3] proposed a multi-level privacy protection method to solve the privacy problems that may arise in collaborative filtering systems and protect users' privacy. In addition to improving the practicability of recommender system by improving privacy, it is also effective by avoiding the single nature of recommendation results and improving the diversity of recommendation results [4]. Ding and Wang [5] used content-based recommendation algorithm to enable users to choose more different items and improve the coverage rate of recommendation system according to the content properties of the project. In the field of data mining, noise reduction is always an essential data cleaning process. Ma et al. [6] proposed a reliable collaborative filtering algorithm, which can reduce false scores and redundant information in e-commerce and improve the accuracy of recommendation results. Darman and Fasihi [7] broke the limitations of traditional collaborative filtering algorithms and took into account the changing interest of users. A similarity method based on dynamic trust is proposed to improve the performance of the proposed algorithm.

3. Customer Value Hierarchy Model

A distributed data analysis system - model brand value model using the neural network and customer value is put forward. By using the UGC of SINA micro-blog, the customer value is modelled using neural network, and distributed big data processing is realized through Hadoop platform technology. The model is realized and successfully deployed on the platform application layer.

The customer value hierarchy model can indicate the psychological picture of the customers' value judgment in the process of buying and using the goods. However, the traditional customer value or the content of the customer satisfaction research is mostly limited to the study of the specific property of the product. As a result, the information obtained is limited to the customer's judgment of the product attribute layer, and thus it is unable to get customers' deep buying motivation. Therefore, the emergence of the customer value hierarchy model is a great progress in the study of traditional customer value and customer satisfaction degree. It reveals that the understanding of the customer's value of the commodity is not limited to the cognition of the goods itself, but is the pursuit of the recognition of the objectives to be realized. That is the reason why traditional customer value cannot be deeply studied, but focuses only on the attribute level of goods. Only after the in-depth study of the outcome level and target layer of customer value can the true value of this commodity to consumers be obtained.

One of the difficulties in the study of customer value model is the acquisition of sample data. The model used here uses sample data from SINA micro-blog. Through the certain pre-processing of user reviews, the data similar to questionnaires can be obtained. Because of the huge amount of information, micro-blog has completely solved the situation that a large number of samples cannot be obtained through questionnaires. The normalized method is used to transform statistical information into values between 0-1. 1 suggests completely positive; 0 represents completely negative; 0.5 indicates neutral. The normalization method can be expressed as follows:

$$AttVal(x) = 0.5 + \frac{AttNum_p(x) - AttNum_n(x)}{TotalNum(x)} \tag{1}$$

4. Experimental Results and Analysis

4.1 Model Convergence Speed and Training Precision Test

The experiment uses the iPhone as the model brand and obtains data from SINA micro-blog. The test data is 25M, a total of 5736 records. The former 4000 records are used as training sets used to train the neural network, and the latter 1736 data are used as test sets to test the accuracy of the neural network. The error calculation method is as follows:

$$ErrSum = E_k = \frac{1}{2}\sum_{i=1}^{m}\left(T_{ik} - Y_{ik}\right)^2 \quad (2)$$

The convergence rate and the correct rate reflected by the error results are shown in Figures 1 and 2.

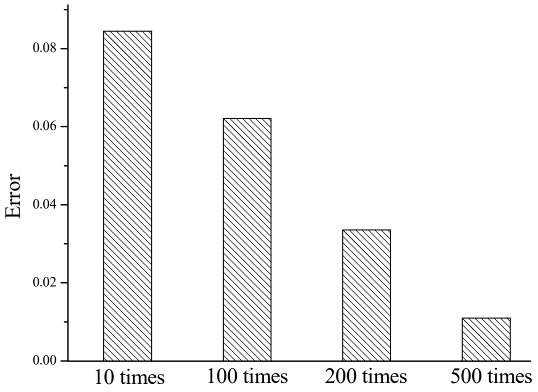

Figure 1. Model Convergence Rate Graph

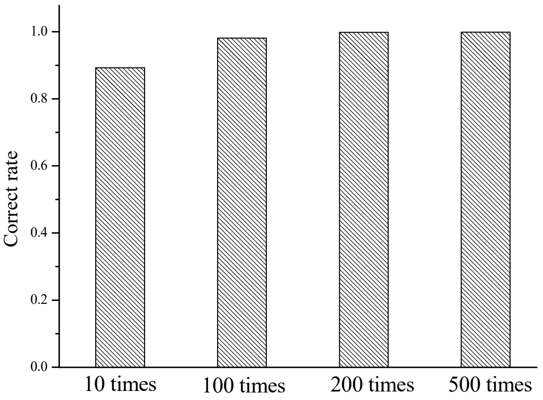

Figure 2. Model Correct Rate

From the experimental results above, it can be seen that increasing the number of training times can appropriately reduce the error and increase the correct rate. When the number of iterations is up to 100 times, the accuracy has reached a certain standard. Because the BP neural network has the effect of overtraining a sample and reducing the accuracy rate by fitting, the same sample should not be over-trained. The number of specific training should be weighed according to the actual situation.

4.2 Large Amount of Data Training Test

Through the iterative training of the same data for several times, it is found that, in the same system environment, the training time is basically proportional to the number of iterations, so as to reduce the interference of other factors such as the network. The training time is proportional to the number of iterations, as shown in Figure 3.

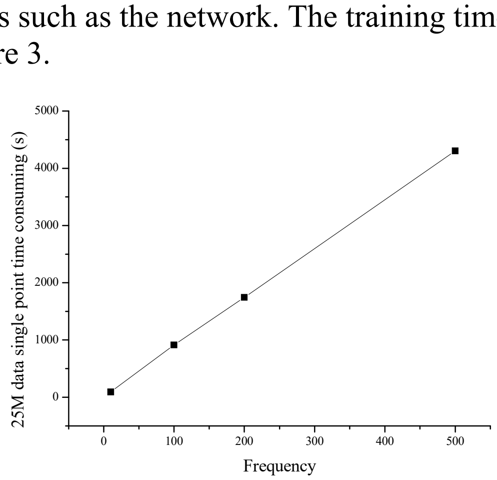

Figure 3. The Proportional Relationship between the Training Time and the Number of Iterations

After getting the relationship between the training time and the number of iterations, greater data are used to test the speed ratio of parallel algorithms, and test whether the algorithm has a high scalability. The experiment uses the data of 100G for the test. According to the above experimental results, the training time is linearly related to the number of iterations, so the data in Table 1 only describe the time of iteration. Speedup ratio is shown in Figure 4.

Table 1. Time used for one iteration of single machine and cluster

Data set	The number of cluster nodes	Running time T	Single machine reference running time t	Speedup ratio t/T
iPhone7: 100G	3	13549s	37395s	2.76
	5	8328s		4.49
	7	7177s		5.21
	9	6714s		5.57

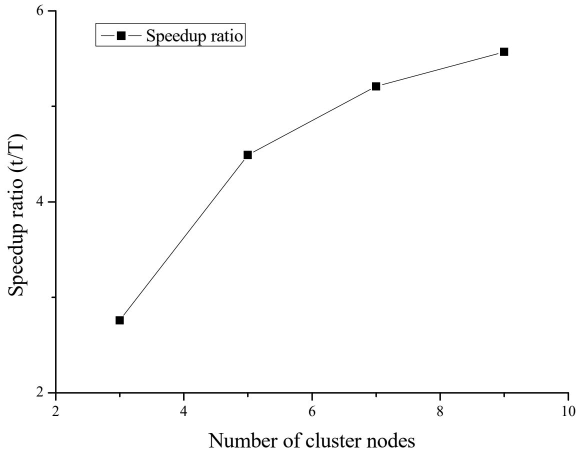

Figure 4. Model Cluster Training Speedup Ratio

The experimental results show that the speedup ratio is also increasing as the number of nodes in the cluster increases. In the case of 5 cluster nodes, the speedup ratio has reached 4.49, close to the number of parallel nodes. However, with the increase of cluster data, the trend of speedup ratio has been slowed down. When the number of nodes increases to 9, the speedup ratio is only 5.57. Therefore, it is not the better the more number of cluster nodes. The advantage of subsequent number of nodes increase is offset by the communication cost between nodes. The increasing number of nodes may even outweigh the losses. There are many factors that affect the number of cluster nodes, such as the size of the sample data, the number of tasks, system hardware, and algorithm optimization. How to determine the number of clusters should also be decided by many experiments and implementation details of specific algorithm.

4.3 Prediction Results

The experiment selected the network after 500 times training of 10G sample as the prediction model, collated and counted the DGC content of 100G, and finally calculated the attribute value of the target layer of the brand: iPhone. After statistical normalization, the statistical samples of the data samples are as follows:

(0.829, 0.824, 0.853, 0.811, 0.84, 0.772, 0.5, 0.5, 0.738, 0.372, 0.687, 0.857, 0.827, 0.797, 0.886, 0.632, 0.835, 0.771, 0.867, 0.68, 0.793, 0.875).

The target level attribute values calculated from statistical samples are shown in Figure 5. 0 indicates completely negative, 0.5 suggest neutral, and 1 refers to completely positive.

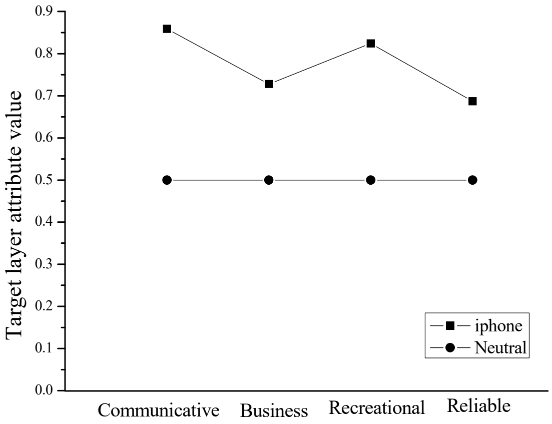

Figure 5. IPhone Brand Target Value

As shown in Figure 5, the value of iPhone's brand target layer is positive, and the value of Communicative is 0.8588. Compared to the high value of Communicative and Recreational, the value of iPhone is slightly lower, which is 0.6869. Through these data, where the satisfaction of customers for iPhone brand is known, where the dissatisfaction is, and where is worth being improved.

5. Conclusion

The design and implementation of brand value model is mainly introduced. The theoretical basis and implementation process of brand value model are summarized, and large amount of data, training accuracy, cluster acceleration ratio and other experiments are carried out relying on the platform. Based on the hierarchical model and neural network idea, the brand value model is proposed by applying the Hadoop distributed platform. The target value of the brand is calculated by the UGC comment content of the brand, thus helping the decision-makers to formulate the long-term planning and research direction of the brand.

A customer value model of iPhone is defined. Because iPhone is a special individual, its universality is not strong. In subsequent studies, some more general customer value models can be defined. For example, a customer value model of a "cell phone" is defined, and the general function of a general cell phone is modelled to make its "brand value model" in combination with the neural network. After having a new mobile phone in the market, as long as the content of the mobile phone is acquired, the brand value model of the mobile phone can be used to calculate the brand value of the mobile phone, without the need of modelling, training and other tasks. As a result, the commercial value and scalability of the model will be greatly increased.

References

[1]. Duffau H. A two-level model of interindividual anatomo-functional variability of the brain and its implications for neurosurgery. Cortex, 2016, 86, pp. 303-313.

[2]. Kong D, Zhang K, Schubert G. A Fully Self-consistent Multi-layered Model of Jupiter. Astrophysical Journal, 2016, 826(2), pp. 127.

[3]. Guo Y, Tang Q, Gong D Y, et al. Estimating ground-level PM 2.5, concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. Remote Sensing of Environment, 2017, 198, pp. 140-149.

[4]. Hamamoto I, Osnes E. Simple two-level model of the hole-vibration quadruplet in 39 K. Physics Letters B, 2016, 53(2), pp. 129-132.

[5]. Ding W, Wang Y. Confinement loss in hollow-core negative curvature fiber: A multi-layered model. Optics Express, 2017, 25(26).

[6]. Ma Y, Fan Z, Zhang B, et al. Theoretical study of optical pump process in solid gain medium based on four-energy-level model. Journal of Optics, 2018, pp. 20.

[7]. Darman M, Fasihi K. A new compact circuit-level model of semiconductor lasers: investigation of relative intensity noise and frequency noise spectra. Journal of Modern Optics, 2017, 64(3), pp. 1-7.