International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
512 – 516

Data Retrieval and Sorting for Multidimensional Data Using Machine Learning in Big Data

Ms. Nipa D. Bhadja
M.E. Computer Engineering
Noble Group of Institutions
Junagadh, India
nipa.bhadja@ngivbt.edu.in

Prof. Ashutosh A. Abhangi
Computer Engineering
Noble Group of Institutions
Junagadh, India
Ashutosh.abhangi@ngivbt.edu.in

Abstract — How to retrieve accurately and quickely locate information from massive network data from the big data is focuse on data retrieve sorting optimization model.Based on traditional data retrieval sorting technology this work proposes multidimensional data retrieval and sorting considering the characteristic of data,users and application. This work use the financial microblog data retrieval base on real query intensions and financial tendency of the system. Finally this work shows the test results for multidimensional data retrieval and sorting using machine learning. In this thesis, proposed methodology has been presented. The proposed system is implemente using machine learning based on the multidimensional data characteristic. This proposed system is implemented in java . It shows soritng and retrieval result of the system based on the characteristic of multidimensional data.

Keywords- *Big Data; Information retrieval; Sorting optimization; Characteristic analysis*

*****

I. INTRODUCTION

21st century is century of information and technology where each and every entity produce the torrent of data so called big data, So just because of that it is difficult to manage this type of large data so that we have to manage the big data by using various technique so that we can easily, securely and fast access these data. Big Data is a term defining data that has three characteristics. First is the great volume of data, second the data cannot be structured into tables and third is velocity which means data is generated rapidly and thus is need to be processed and analyzed fast.

Big data is defamed as large amount of data which requires new technologies and architecture so that it becomes possible to extract value from it by capturing and analyzing process.

II. LITERATURE SURVEY

A. *"Design and Implementation of A Multidimensional Data Retrieval Sorting"*

How to accurately and quickly locate required information from the massive network data, especially from the current popular social network data, is the focus of data retrieval services. Based on the traditional data retrieval sorting technology, this work proposes a multi-dimensional data retrieval sorting optimization model, considering the characteristics of data, users and applications. This work implements this model in the system of financial micro blog data retrieval. It enables the retrieval system to sort the results according to the characteristics of them micro blog data, users' real query intentions and financial tendency of the system. This paper introduced a data retrieval sorting optimization model, which optimizes the Lucene search results sorting order from three dimensions: data features, user features and application features. Based on financial microblog data retrieval system, it also implemented this model and proved the model's optimization effect by practical testing.

B . *"Fast Scalable Selection Algorithms for Large Scale Data"* [2]

Selection finding, and its most common form median finding, are used as a measure of central tendency for problems in biology, databases, and graphics. These problems often require selection finding as a subcomponent where it can be called many times, and as such speed is important. The Map/Reduce framework has been shown to be an important tool for creating scalable applications. There are a number of valid implementations of the selection algorithms inside of a Map/Reduce framework, certain of which are compared in this paper. However, as the volume of data increases, subtle theoretical algorithmic implementation differences can lead to significant differences in practical application. Therefore, an efficient and scalable selection finding method has the potential to provide general benefit to a number of applications. This paper compares algorithms that have been redesigned or created for the Map/Reduce framework for the purpose of selection finding, or, finding the k-th ranked element in an unordered set. This paper takes the concepts used from two existing selection algorithms and translates them into a novel method using the Map/Reduce framework with two

 ISSN: 2454-4248
512 – 516

variations. Each approach uses a different methodology to reduce the total amount of workload needed for a selection. All the algorithms are compared together for scalability and efficiency in a computing cluster environment with up to 256 processing cores. The results show that the methods proposed in this paper outperform several common alternatives in identifying medians with Hadoop, including using sorting, Pig, and BinMedian methods. Our implementations are also available upon request.

B. *"Optimizing Sorting with Genetic Algorithms"* [3]

"Pure" sorting algorithm at the outset of the computation as a function of the standard deviation. The approach discussed in this paper uses genetic algorithms and a classifier system to build hierarchically-organized hybrid sorting algorithms capable of adapting to the input data. Our results show that such algorithms generated using the approach presented in this paper are quite effective at taking into account the complex interactions between architectural and input data characteristics and that the resulting code performs significantly better than conventional sorting implementations and the code generated by our earlier study. The best algorithm we have been able to generate is on the average 26% and 62% faster. In this paper we study machine learning techniques to extend empirical search to the generation of sorting routines, whose performance depends on the input characteristics and the architecture of the target machine learning.

C. *"Comparing Deep Learning And Support Vector Machines for Autonomous Waste Sorting"* [4]

Waste sorting is the process of separating waste into different types. The current trend is to efficiently separate the waste in order to appropriately deal with it. The separation must be done as early as possible in order to reduce the contamination of waste by other materials. The need to automate this process is a significant facilitator for waste companies. This research aims to automate waste sorting by applying machine learning techniques to recognize the type of waste from their images only. Two popular learning algorithms were used: deep learning with convolution neural networks (CNN) and support vector machines (SVM). Each algorithm creates a different classifier that separates waste into 3 main categories: plastic, paper and metal. The accuracies of the two classifiers are compared in order to choose the best one and implement it on a raspberry pi 3. The pi controls a mechanical system that guides the waste from its initial position into the corresponding container. However, in this paper we only compare the two machines learning techniques and implement the best model on the pi in order to measure its speed of classification. SVM achieved high classification accuracy 94.8% while CNN achieved only 83%. SVM also showed an exceptional adaptation to different types of wastes. The SVM model was finally implemented on a Raspberry pi 3 where it produced quick classification, taking on average 0.1s per image. In this research we presented a comparison of deep learning and SVM for waste sorting. Alex Net was able to achieve good classification accuracy of 83% however SVM bettered it with 94.8% on the test set. The SVM model was implemented on a raspberry pi and demonstrated the ability to sort waste by their images.

D. *"Multimodal Biometric System using Index Based Algorithm for Fast Search"* [5]

Establishing the identity of a person with the use of individual biometric features has become the need for the present technologically advancing world. Due to rise in data thefts and identity hijacking, there is a critical need for providing user security using biometric authentication techniques. A unimodal biometric system is known to have many disadvantages with regard to accuracy, reliability and security. Multimodal biometric systems combine more than one biometric trait to identify a person for enhanced security. The proposed MuItimodal biometric system combines three biometric traits for individual authentication namely Face, Fingerprint and Voice and implements indexing algorithm for faster searching and recognition of a person. The fuzzy nature of biometric data and the presence of varied degree of dimensionality hinder present search algorithms depending on sorting. Index based algorithm is used to reduce the search time and also to improve the speed and performance of multimodal biometric system. Map Reduce is used for analyzing and processing big data sets that cannot fit into memory.

III. PROPOSED WORK

We are trying to retrieval and sorting of multidimensional data using some machine learning algorithm so that I try to improve the result based on sort and retrieval of data.

 ISSN: 2454-4248
512 – 516

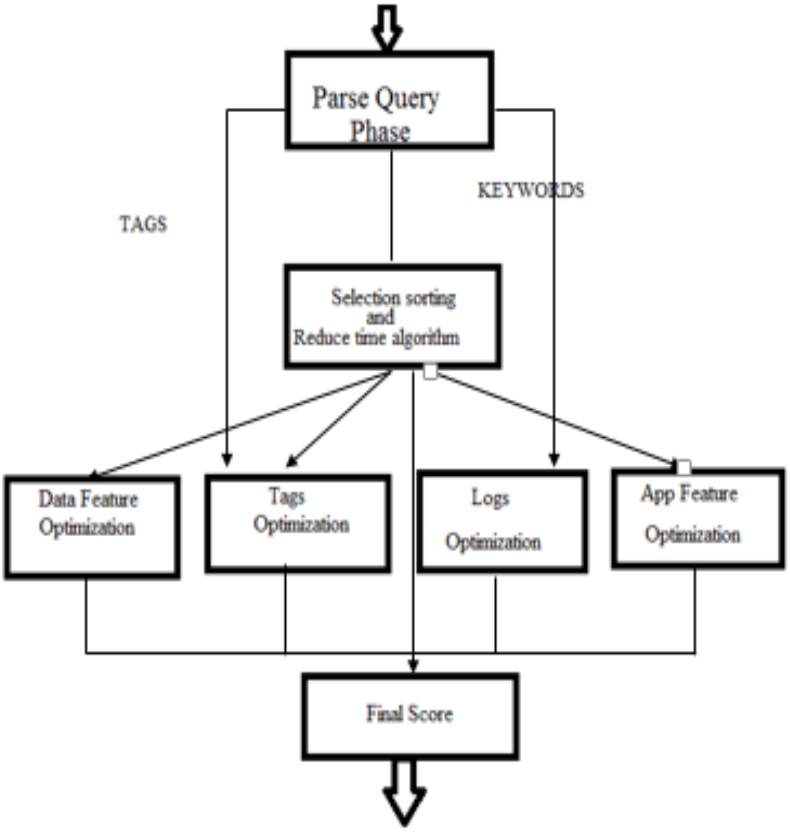


FIGURE 1 . PROPOSED SYSTEM DIAGRAM

A. MULTIDIMENSIONAL DATA CHARACTERISTIC

- Data characteristic
- Use characteristic
  1)tags optimization
  2)logs optimization
- Application characteristic

1. Based on data features to optimize:

- Text length
- Publishes time
- Comments
- Number of followers
- Author's published microblog

Attr(score)=(accurrence_times)/5

2. User features analysis.

a) Tag optimization

Weight(wtag)=count(wtag)/∑count(wtags)

b) Log optimization

Log =(avg i)/ ∑(avg j)

3. Application features parameters.

App(para)=∑count(result categorized finance)/query times

B. FINAL SCORE OPTIMIZATION

Optimize(score)=f(∑data(attr)+g(∑ user(attr))+h(∑app(attr))

C . ALGORIHTM STEPS.

 ISSN: 2454-4248
512 – 516

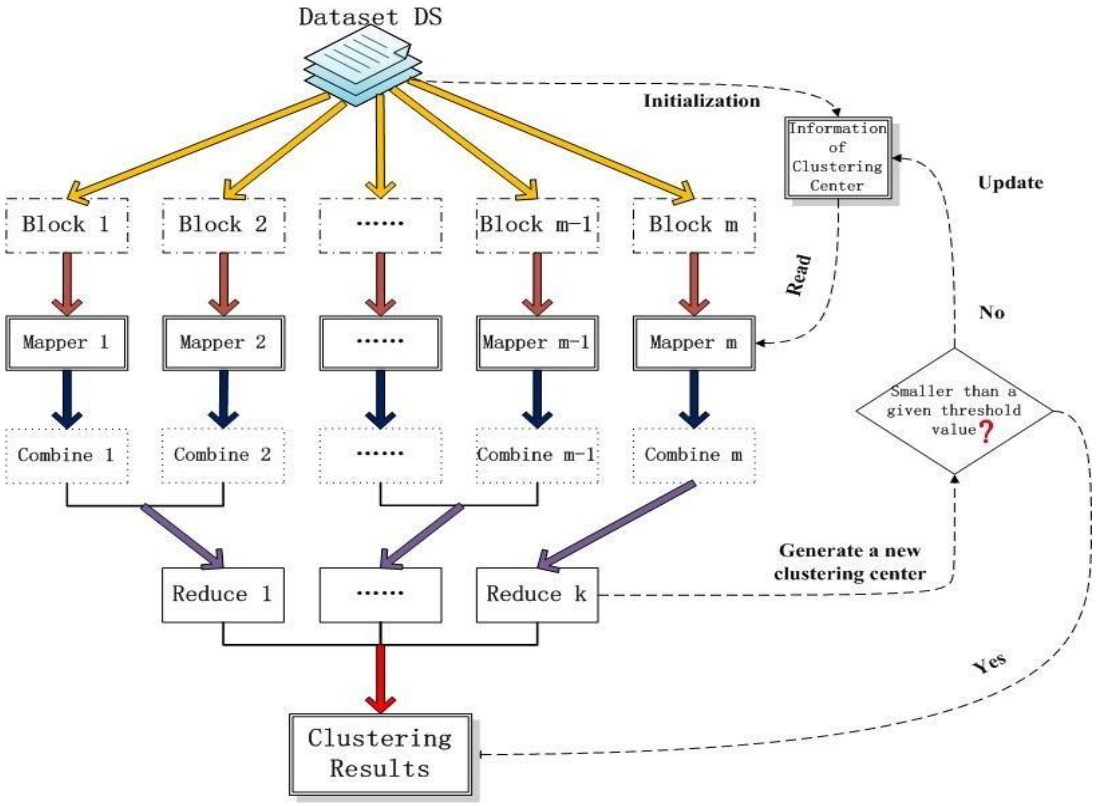
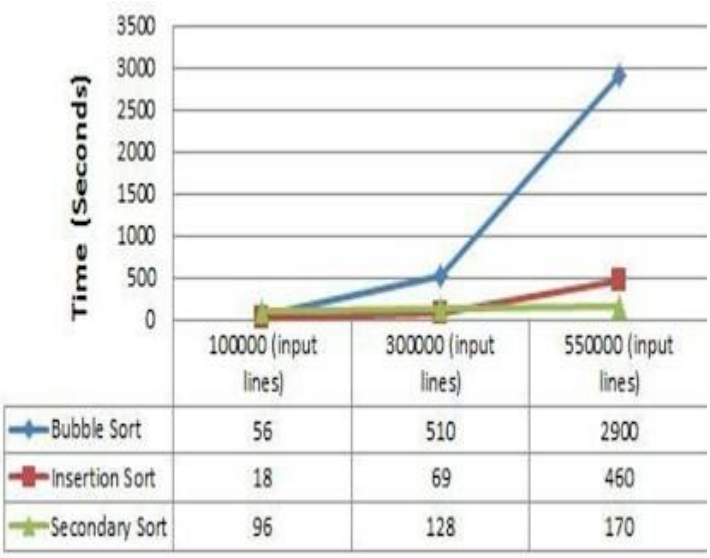


FIGURE 2. WORK FLOW OF ALGORITH

IV. RESULT



	100000 (input lines)	300000 (input lines)	550000 (input lines)
Bubble Sort	56	510	2900
Insertion Sort	18	69	460
Secondary Sort	96	128	170

In this result output the final score optimazation which obtaind with machine learning have take few time with different size of input lines .

V. CONCLUSION

Machine learning algorithm can be deployed to do some dynamic adjustments to obtain more accurate optimization scores that retrieve good result on the data query for the big data for sorting the query and retrieving the information on the query. And also take less time to retrieve the query.

VI. REFERENCES

[1]. Danfeng Yan, Liying Zhang, Xuan Zhao, Design and Implementation of A Multidimensional Data Retrieval Sorting Optimization Model, IEEE 2016.

 ISSN: 2454-4248
512 – 516

[2]. Xiaoming Li, Mar´ıa Jes ´us Garzar´an and David Padua, Optimizing Sorting with Genetic Algorithms,IEEE.
[3]. Arun Kumar, Anurag Pandey, Suman Kaushik, Machine Learning Methods for Solving Complex Ranking and Sorting Issues in Human Resourcing, IEEE 2017.
[4]. George E. Sakr, Maria Mokbel, Ahmad Darwich, Comparing Deep Learning And Support Vector Machines for Autonomous Waste Sorting,IEEE2016.
[5]. Meghana A Divakar, Multimodal Biometric System using Index Based Algorithm for Fast Search, IEEE 2016.
[6]. Somshubra Majumdar, Ishaan Jain, Kunal Kukreja, Professor Kiran Bhowmick, Adaptive Sorting Using Machine Learning, International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 7 (2) , 2016, 490-49.
[7]. Ella Peltonen, An approach to Machine Learning with Big Data, September 19, 2013
[8]. LU ZHI XIANG, Research and Improvement of PageRank Sort Algorithm based on Retrieval Results,IEEE 2014
[9]. Chr i st i naOr phani douandDav i dWong, Machi neLear ni ngModel sf orMul t i di mensional Cl i ni cal Dat a, SPRI NGER2017