ANALYSIS OF KNN ALGORITHM WITH MAPREDUCE TECHNIQUE ON BIG DATA

Tatikonda.Bhavana¹, J.padmavathy², R.Sethuraman³, J.K Jeevitha⁴

¹,²UG Scholar, ³Assistant Professor Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, ⁴Assistant Professor, PSNA College of Engineering and Technology

bhavanachowdary424@gmail.com, keerjey@gmail.com, srssethuraman@gmail.com, selvajeeva31@gmail.com

Abstract: Due the fast growth of new technology application like social media analysis, web data analysis and medical information network analysis, here the various types of data are processed frequently. The large amount of effective data management and analysis is very vital goal. To reduce the data processing complexity, time complexity, and space complexity in Big Data, the paper going to propose the k-nearest neighbor join (KNN) operation. KNN is used to find the K nearest points in S. It is a computational task that will handle the large range of applications such as knowledge discovery or data mining. When the volume and the dimension of data increases, then only distributed approaches can perform the big operations in a given time. Recent works have done on implementing the efficient solutions using the map reduce programming model because it is used for distributing the large scale data processing. Although these works provide different solutions for the same problem, each one has particular constraints and properties. This paper compares the existing of different computation of KNN on MapReduce. First the paper compares the solutions in to three steps for KNN computation on MapReduce: 1) Data processing, 2) Data partitioning and 3) Computation. The Experiment in this paper explains the variety of different data sets, and analyzes the data volume, data dimension and the value of k from many perspectives like time and space complexity, and accuracy.

1. Introduction:

As we have to get the relevant information or details from large amount of data we use KNN algorithm and Map Reduce algorithm [1][2]. We will merge both the algorithms. KNN algorithm means it will find or search the data nearest to it. MapReduce means it will process large amount of data. Suppose if we take a query like (1990, 2002) employee name. It will give all the employee details in single form or code. From the huge amount of data we can easily get the information what we required. Here we will use Hadoop frame work that provides a distributed platform for executing Map Reduce jobs.

Hadoop frame work is an open source network that will provide the large amount of data sets. It consists of K nearest neighbor that helps in finding the nearest points in a dimensional space. MapReduce is a
parallel programming model and this consists is based on three concepts 1) it will represent the data as key-value pairs 2) it will define a map function 3) it will define reduce function. The map function will take the key value pairs as input, and it produce zero or more key value pairs. Outputs with the same key are gathered together, and that key pairs are given to reducers. Data preprocessing will transform the original data that will benefit the particular properties.

This step is done before the partitioning of data in to two different goals. 1) either to reduce the dimension of data. 2) to select central points of data clusters. Data partitioning and selection process is also used. MapReduce is a shared–platform in order to process data. We have to divide the data sets into independent pieces called partitions. It can be applied to a large number of fields such as medical imagery, bio-information. The primary application of a KNN join is k-nearest neighbor classification. Some data points are given for training and some new unlabeled data is given for testing. The aim is to find the class label for the new points. For each unlabeled data; a KNN query on the training set is to be performed to estimate the class membership. This process will be considered as a KNN join for testing set with training set[3][4][5].

2. Related work:

The author proposed the MapReduce method in cloud computing for improving the performance of data processing and information retrieval; additionally they concentrated the KNN algorithm in Hadoop platform for the analysis of time consumption [6].

Pradesh [7] implemented the data mining concepts in KNN on distributed environment with apache server. The apache server ran in the distributed computing environment, which used Hadoop platform and MapReduce technique for huge amount of data processing.

The author implemented the MapReduce with KNN algorithm in large scale data set up to 30 dimensions for improving the efficiency in data processing [8].

3. Proposed Work:

![Figure-1: Work flow diagram for MapReduce and KNN.](image-url)
Figure 1 summarizes the work flow of MapReduce and KNN computation. The work flow of MapReduce and KNN computation architecture processed in three steps.

Step1: First step had done the data processing; moreover it performs the selections of data projection (dominating points) from maximum to minimum dimension.

Step2: This step is used to splitting and segregating the processed data. This step makes simplifies the KNN computation process.

Step3: Third step did one or two function of MapReduce, according to the distance calculation and sorting.

KNN algorithm is very useful and uncomplicated algorithm. Here KNN used the whole dataset in the phase of preparation. While the prediction is demanded the unknown data inputs, that time it seeks out the prepared data set for K-nearly related input and the data with most suitable input. At the final stage the most suitable input is considered as prediction.

4. Result and Discussion:

In this we will use both KNN algorithm and Map Reduce algorithm. KNN algorithm is used to find the nearest point’s. Map Reduce is used for process the large amount of data. KNN can be measured by using distance function formulas. The distance functions are implemented in mathematical way [9][10].

![Distance Function](image)

Figure 2: Distance Function.

It will bring the issues of using the variables that starts with 0 and 1.

Figure 2 shows the graph between loan and age. This points can be represented in default and non-default. By using Euclidean distance formula, this paper measured the accurate value.

In the table 1 suppose if we consider k=5 then Y=4 default and N=1 default. If we want to find the default value we have to find distance for all the persons in the graph and then we have to take the smallest values from the given table and then from that take the smallest value and we have to find the value.
MapReduce can be done by the following steps:
Step1) The input for the map has to be prepared
Step2) Run the Map code provided
Step3) Shuffle the outputs of the map to the Reducer Preprocessors
Step4) Run the code
Step5) Product the final output
In this MapReduce Data Flow can be happened in following steps:

1) **Input Reader:** The input reader will divide the input in to different sizes and read the data that is provided by the user

2) **Map Function:** It will take the input of key value pairs and provides output of zero or more key value pairs

3) **Partition Function:** Provide the key and number of reducers and return the index of the reducer

4) **Comparison Function:** Compare the inputs of the map and reduce functions

5) **Reduce Function:** It will take the input values and provide the single output

6) **Output Writer:** It will give the output for the reduce function

5. **Conclusion:**

In this paper had analysed the performance of KNN algorithm with the process of MapReduce. This approach measured the solution in to three different steps for computing the accuracy of the space and time complexity in Big Data with the combination of the Map Reduce function and KNN computation. And this process had done in three steps data partitioning, pre-processing and computation. In our future work, we decided to improve the speed of the data processing.

**References:**

[1] D. Li, Q. Chen, and C.-K. Tang, “Motion-aware knn laplacian for video matting,” in ICCV’13, 2013.

[2] K. Inthajak, C. Duanggate, B. Uyyanonvara, S. Makhanov, and S. Barman, “Medical image blob detection with feature stability and knn classification,” in Computer Science and Software Engineering, 2011.

[3] C. Yu, R. Zhang, Y. Huang, and H. Xiong, “High-dimensional knn joins with incremental updates,” Geoinformatica, 2010.

[4] C. Ji, T. Dong, Y. Li, Y. Shen, K. Li, W. Qiu, W. Qu, and M. Guo, “Inverted grid-based knn query processing withmapreduce,” in Proceedings of the 2012 Seventh Grid Annual Conference, ser. CHINAGRID ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/ChinaGrid.2012.19.

[5] Apexa B. Kamdar, Ishan K. Rajani, “Improved Adaptive K Nearest Neighborhood algorithm using MapReduce”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 6, June 2015.

[6] Prajesh P Anchalia, Kaushik Roy,” The k-Nearest Neighbor Algorithm Using MapReduce Paradigm”, 2014 Fifth International Conference on Intelligent Systems, Modeling and Simulation.

[7] Chi Zhang, Feifei Li, Jeffrey Jestes,” Efficient Parallel knn Joins for Large Data in MapReduce”, March 26-30, 2012, Berlin, Germany, ACM.

[8] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Commun. ACM, 2008.

[9] G. Song, J. Rochas, F. Huet, and F. Magoules, “Solutions for Processing K Nearest Neighbor Joins for Massive Data on MapReduce,” in 23rd Euromicro International Conference on Parallel, Distributed and Network-based Processing, Turku, Finland, Mar. 2015.

[10] M. I. Andreica and N. T. Pus, “Sequential and mapreduce-based algorithms for constructing an in-place multidimensional quadtree index for answering fixed-radius nearest neighbor queries,” 2013.