www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20350-20353
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.33

Data Mining With Big Data Using Clustering Based Collaborative Filtering

Monali V. Mhaske[1], *Priyanka Namdev Ubale*[2], *Dipika Nana Nagtilak*[3], *Punam Mahadev Jamdade*[4]

[1]B. E. Computer Science & Engineering,
Yashoda Technical Campus, Satara
monali.mhaske14@gmail.com
[1]B. E. Computer Science & Engineering,
Yashoda Technical Campus, Satara
Priyaubale25@gmail.com
[3]B. E. Computer Science & Engineering,
Yashoda Technical Campus, Satara
nagtilakdipika@gmail.com
[4]B. E. Computer Science & Engineering,
Yashoda Technical Campus, Satara
Punam.jamdade12@gmail.com

Abstract:

Today large amount of data is generated from various and heterogeneous sources in day to day life. There 300 million of users posts the images, messages and other type of data on Facebook per day. 3.5 billion Searches are processed by google per day. Traditional system that handles the data in megabytes and gigabytes cannot efficiently handle such a huge volume, distributed data that comes from heterogeneous sources.And such data which is huge in volume, complex, distributed, unmanageable and heterogeneous is called as Big Data. In 2004 google proposed MapReduce parallel processing model that provides the parallel processing.Whenever user hits the query, MapReduce model splits it and assigns to the parallel nodes to process that query parallely.The results evaluated by all the nodes are collected and delivered to the user. The Apache open source model called “Hadoop” adopted this MapReduce framework.

This paper represent system that uses collaborative filtering on the big data that have been clustered. This way of mining and managing big data is more efficient than using another traditional systems. The main challenges of big data are storing, searching, manipulating and security. The efficient system should be able to overcome all these challenges, and those what the parameters on which this paper focuses. The use of K-means algorithm for clustering has been recognized by many of big data handlers.Clustering divides the big data into clusters, with the data having same characteristics on one cluster. Clustering increases accuracy and takes less time to compute the results. Clustering are techniques that can reduce the data sizes by a large factor by grouping similar services together. This paper proposes of two stages: clustering and collaborative filtering. Clustering is a pre-processing step to separate Big Data into manageable parts.

Keywords: clustering, hadoop, Big data, collaborative filtering.

1. Introduction

Nowadays ‘Big Data’ is recognized all over in the world. Data that comes with complex, heterogeneous, tremendously large volume and from numerous, self-directed sources defines the Big Data. The demand of big data processing has grown immensely, which increases an inconveniency and heavy load on computation, storage and normal behaviour of traditional systems. And thus to reduce this load and to improve the efficiency of big data handling, the use of clustering with Hadoop framework [7] to mine and manage the Big Data has proven itself better.

Hadoop is an open source framework developed for distributed processing and distributed storage of huge volume data sets by Apache Foundation and was released in 2006. Apache Hadoop uses distributed storage system called “Hadoop Distributed File System (HDFS)” [7]. Hadoop processes big data using “MapReduce” [7][8] parallel processing model. Whenever user uploads data set on Hadoop, Hadoop divides that data sets into blocks and distributes those blocks among the nodes in cluster. And then the queries that users hit are divided according to the nodes in that particular cluster. “Hadoop Distributed File System (HDFS)” and MapReduce are the two core components of Hadoop.

This paper focuses on two stages: clustering and collaborative filtering. Clustering is first step used to separate ‘Big Data’ into controllable parts. Clustering partitions the big data into clusters, with the data having same characteristics on same cluster. And then it becomes easy to perform the mining operations on datasets in cluster.

After clustering the data, the technique collaborative filtering is used to abstract the useful data and to recommend it to users. Recommendation works on the basis of users having similar choice.

2. Literature Survey

Some researches have been made on clustering based collaborative filtering approach [1] for handling big data. This ClubCF approach have been used for service recommendation systems. Using clustering techniques in recommendation systems for Big Data have proven its best performance as compared with traditional system.K-means clustering algorithm [1] [4] [5] have been successfully used by some researchers for Big Data and also appreciated by them. K-means algorithm

The HACE theorem [2][3] was not used by some researchers, which results in more time taken by clustering mechanism. So we realized that, the HACE theorem should be used to minimize the time required for further clustering and collaborative filtering processes and to form the clusters with most relevant data in same cluster.The mechanism of collaborative filtering [6] uses user-rating data to compute similarity between users or items. There are two methods that are used for this mechanism: user-based and item based collaborative system.

The Hadoop framework [7] have proven its efficiency for accessing or mining Big Data as it uses MapReduce parallel computing model [8] with clustering. Clustering technique [9] is best at using it for Big Data processing, as Big Data comes from various resources and clustering make it possible to mine the data where it is generated. There are number of techniques used to gather necessary data from Big Data, among of them collaborative filtering [10] is the most efficient for commercial use.

Clustering based collaborative filtering [11] is the most popular and effective technique used for recommendation purpose such as e-commerce websites.So, today lots of researches have been made on efficient ways to mine the Big Data [12] in effective manner.

3. Proposed Work

As Big Data is growing daily with huge volume and is stored at different places, the efficient system should able to operate on the heterogeneous types of data and to extract useful data among the Big data. To make this possible, first using clustering that handles the data parallely and works almost same as MapReduce and then applying collaborative filtering to access only necessary data is the best way. There number of clustering techniques, among them the K-means partitional clustering has proven its efficiency and best usage for mining Big Data.

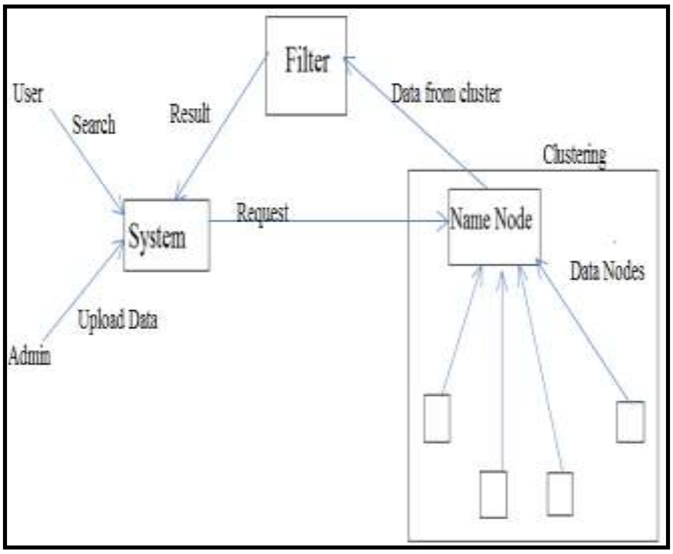

Fig 1 : System Architecture of clustering based collaborative filtering

3.1 K-means algorithm

K-means algorithm simply intends to partition the ‘n’ number data objects into ‘k’ number of clusters (k ≤ n), where each object must belongs from one and only cluster. Data is clustered on the basis of similarity. Each object in cluster is having nearest mean with its cluster center. K-means works in following steps:

1. Partition the data objects {X_1, X_2, X_3, ------, X_n} into ‘k’ clusters.

2. Select 'k' points randomly as cluster center { O_1, O_2, O_3, ------, O_k } for each cluster.

3. Using following "Euclidean Distance Function", find out the distance between each object and cluster center and assign each object to its closest cluster center.

Euclidean Distance Function:

$$d = \| X_i - X_j \|$$

4. Re-evaluate the new cluster center using :

$$O_i = \left(\frac{1}{Ci}\right) \sum_{j=1}^{Ci} Xi$$

Where C_i = total number of objects in i^{th} cluster.

5. Repeat steps 3 and 4 until no object is reassigned..

3.2 Collaborative Filtering

There are two most popular techniques of collaborative filtering: Item based collaborative filtering and user based collaborative filtering.

On the basis of your need and requirements the best one can be chosen. We have decided to use user based collaborative filtering technique as we are going to use this project for shopping mall.

The user – based Collaborative filtering simply finds the users having similar choices for services and then suggests them each others' services other than common ones.

4. Conclusion

In this paper, we represent the system that handles Big Data efficiently by applying collaborative filtering to the clustered data. Computation time, efforts and speed are the factors that are affected by using the clustering based collaborative filtering approach.

Number of requests on processing Big Data in cluster is considerably less than that of in traditional central architectural system, which reduces the difficulty of collaborative filtering algorithm. And hence system becomes able to provide most accurate results in minimum of time.

5. Acknowledgement

This research was supported by Professor Mahesh A. Shinde. We are thankful to our colleagues who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper.

We are also grateful to our institute Yashoda Technical Campus, Satara and staff for assistance who moderated this paper and in that line improved the manuscript significantly.We have to express out appreciation to our family members for sharing their pearls of wisdom with us during the course of this research.

References

[1] Rong Hu, Wanchun Dou, Jianxun Liu, "ClubCF: Clustering Based Collaborative Filtering Approach For Big Data Application", IEEE Transaction on Emerging Topics in Computing, pp 10 March 2014.

[2] PremaGadling, MahipBartere, "Implementing HACE Theorem for Big Data Processing", IJSR, Vol. 5, Issue 6, June 2016.

[3] Deepak S. Tamhane, Sultana N. Sayyad, "Big Data Analysis On Hace Theorem", IJARCET, vol. 4, issue 1, pp Jan. 2015

[4] Sachin Shinde, Bharat Tidke, "Improved K-means Algorithm for searching Research papers", International Journal of Computer Science & Communication Networks, vol 4(6).

[5] Jyoti Yadav, Monika Sharma, "A Review of K-mean Algorithm", IJETT, Vol. 4, Issue 7, pp July 2013.

[6] AtishaSachan, VineetRichariya, "A Survey On Recommender Systems based on Collaborative Filtering Technique", IJIET, Vol. 2, Issue 2, pp April 2013.

[7] A.Pradeepa, Dr. Antony SelvadossThanamani, "Hadoop File System and Fundamental Concept of MapReduce Interior and Closure Rough Set Approximations", Vol. 2, Issue 10, October 2013.

[8] AbdelrahmanElsayed, Osama Ismail, and Mohamed E. El-Sharkawi, "MapReduce: State-of-the-Art and Research Directions", Vol. 6, No. 1, Febuary 2014.

[9] N. Kamalraj, A. Malathi, "Hadoop Operations Management for Big Data Clusters in Telecommunication Industry", vol. 105, No. 12, November 2014.

[10] Punam B. Thorat, R. M. Goudar, SunitaBarve, "Survey on Collaborative Filtering, Content Based Filtering and Hybrid Recommendation System", International Journal of Computer Application, Vol. 110, No. 4, January 2015.

[11] Rohit C. Joshi, Ratnamala S. Paswan, "A Survey Paper on Clustering Based Collaborative Approach to generate Recommendations", IJSR, Vol. 4, Issue 1, January 2015.

[12] Shilpa, Manjit Kaur, "Big Data and Methodology : A Review", IJARCSSE, Vol. 3, Issue 10, October 2013.

DOI: 10.18535/ijecs/v6i2.33

Author Profile

Monali V. Mhaske
Student of last year B. E. CSE.
Yashoda Technical Campus, Satara

Priyanka N. Ubale
Student of last year B. E. CSE.
Yashoda Technical Campus, Satara

Dipika Nana Nagtilak
Student of last year B. E. CSE.
Yashoda Technical Campus, Satara

Punam Mahadev Jamdade
Student of last year B. E. CSE.
Yashoda Technical Campus, Satara