International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.4., 2015
(July-Aug)

RESEARCH ARTICLE

ISSN: 2321-7758

COLLABORATIVE FILTERING APPROACH FOR BIG DATA USING DATA MINING

S.S.ASOLE[1], AYUSHI V. RATHOD[2]
[1]Assistant professor, [2]ME Student,
Department of CSE,BNCOE Pusad

AYUSHI V. RATHOD

ABSTRACT

This paper presents Big Data Problems result using Data Mining. There is broad recognition of the value of data as well as products obtained through analyzing it. About big data "Size is the only thing that matters." Various Popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist the New York Times and National Public Radio Industry is abuzz with the promise of Big Data Government agencies have recently announced significant programs towards addressing challenge of Big Data. The data are big or small depend on size of data and there are challenges with Big Data.

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. so for mananaging and processing this data we must have some techniques. In this work first different existing technique will be studied and analyzed about Big Data. Here we will include the concepts related to big data like analysis and prediction and result of problems.

The existing approaches perform better but having some drawbacks. So, they cannot be applied to the various situations. To overcome some drawbacks we are going to propose Clustering Approach for Collaborative Filtering which will gives us better result. So,we can apply it to big data, depending upon these techniques, we will try to implement concepts related to Big Data processing. and finally Clustering Approach for Collaborative Filtering performs better.

Keywords: Data mining challenges with big data,.*Hadoop's distributed file system,* map-reduce framework, clustering approach for collaborative filtering

INTRODUCTION

Big data is nothing the large amount of data. There is broad recognition of the value of data as well as products obtained through analyzing it About big data "Size is the only thing that matters." Various Popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist the New York Times and National Public Radio Industry is abuzz with the promise of Big Data Government agencies have recently announced significant programs towards addressing challenges of Big Data. But yet, many have a very narrow interpretation of what that means, and we lose track of the fact that there are multiple steps to the data analysis, whether the data are big or small depend on size of data.

There is work to be done at each step and there are challenges with Big Data.

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a"tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions .In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. The key challenges for such as Data accessing and computing, Data privacy and domain knowledge.

Big Data as always telling us the truth, but this is actually far from reality about big data. We have to deal with erroneous data: some news reports are inaccurate A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

In short, there is collaborative filtering is required to extract value from data. Heterogeneity, incompleteness, scale, timeliness, privacy and process complexity give rise to challenges at all phases.

I.PROBLEM STATEMENT OF BIG DATA:-

1)Data accessing and computing (Tier I)

2)Data privacy and domain knowledge(Tier II)

3)Big Data mining algorithms (Tier III)

The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.

The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process.

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.

II.HADOOP'S DISTRIBUTED FILE SYSTEM

Hadoop's Distributed File System is designed to reliably store very large files across machines in a large cluster.

It is inspired by the Google File System. Hadoop DFS stores each file as a sequence of blocks, all blocks in a file except the last block are the same size. Blocks belonging to a file are replicated for fault tolerance. The block size and replication factor are configurable per file. Files in HDFS are "write once" and have strictly one writer at any time.

III.MAP-REDUCE FRAMEWORK

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster Users specify a *map* function that processes akey/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key.

Map Reduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

A Map Reduce program is composed of a Map() procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name)

Reduce() procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). Even though the previous pseudo-code is written in terms of string inputs and outputs, conceptually the map and reduce functions supplied by the user have associated types:

map (k1,v1) ! list(k2,v2)

reduce (k2,list(v2)) ! list(v2)

I.e., the input keys and values are drawn from a different domain than the output keys and values. Furthermore, the intermediate keys and values are from the same domain as the output keys and values.

IV.CLUSTERING APPROACH FOR COLLABORATIVE:-

To overcome some drawbacks we were use Clustering Approach for Collaborative Filtering which will gives us better result and also we are going use some framework to overcome drawback.

Collaborative filtering is required to extract value from data Why "collaborative"? Basically, someone else (in fact many someones) have gone to the effort of viewing/filtering things, and chosen the best few. You get a recommendation of the best few, without having to spend the effort.

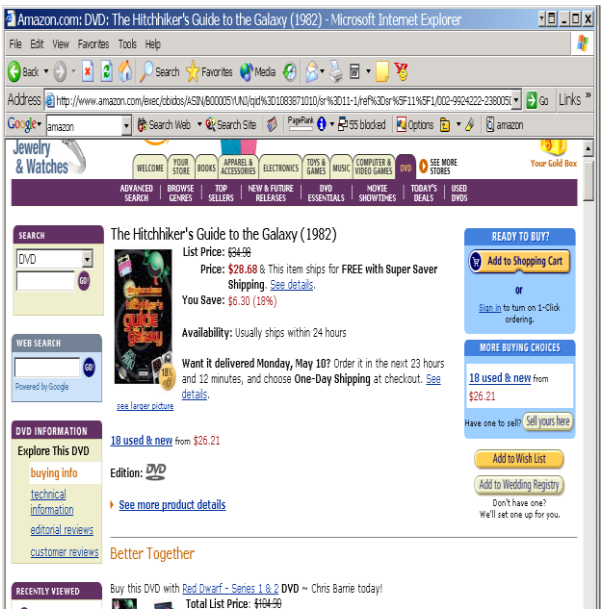


Fig 1.Everyday Examples of Collaborative Filtering

Fig 2. show the detail design of the collaborative filtering.

In this above example it is clear that three different user have rated three different movies.Mr.richerd rated all three movies whereas mary rated only batman and ironman and steve also rates all three movies. in collaborative filtering preferences of users are clustered .In This Example sorting and filtering is done by using the map reduce framework.and movies similarities and movies filtering is done.And final output is calculated by using the co-relation and pearson formula.

Fig 2.-Rating example

1.DETAILED DESIGN:

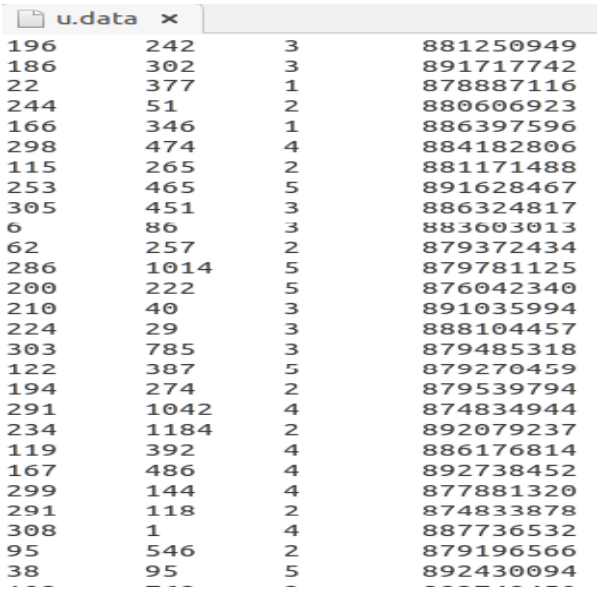


Fig 3-u data

In Fig(3) Udata is a data took from imdb site for the purpose of solving the problem of bigdata analysis and processing. in the below figure column first show user id where as second column shows movie id and third column shows rating of the movies

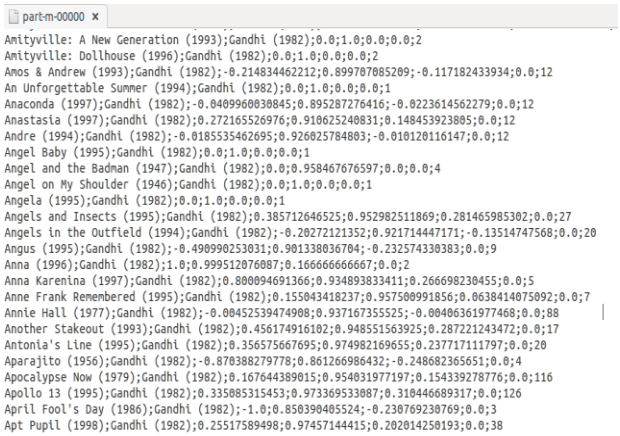


Fig 4-values comparison

In Fig 4. shows the comparison of movies on the basis of parameters. each parameter is compared with each other and nearby values are sorted to find top rated movies.

For example if Gandhi movies is having value0.5;07;0;09;0 so in this case we consider the nearby values of movies to the Gandhi movie.In this way we can calculate the top movies.

2.SYSTEM EXECUTION DETAILS

As per the proposed design i.e. collaborative filtering. First we will look at the publicly available modules screenshots with the home page for the application.

Fig 5.-shows five top most movies

Figure(5) shows the movies search in a data set and gives the output of related top five movies .calculation done in fig(4)

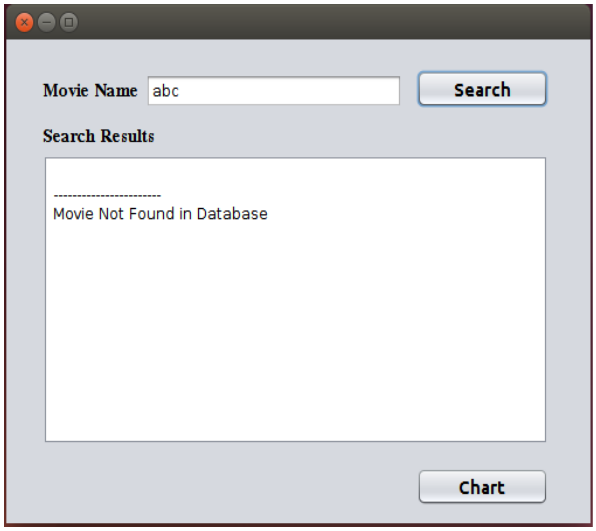
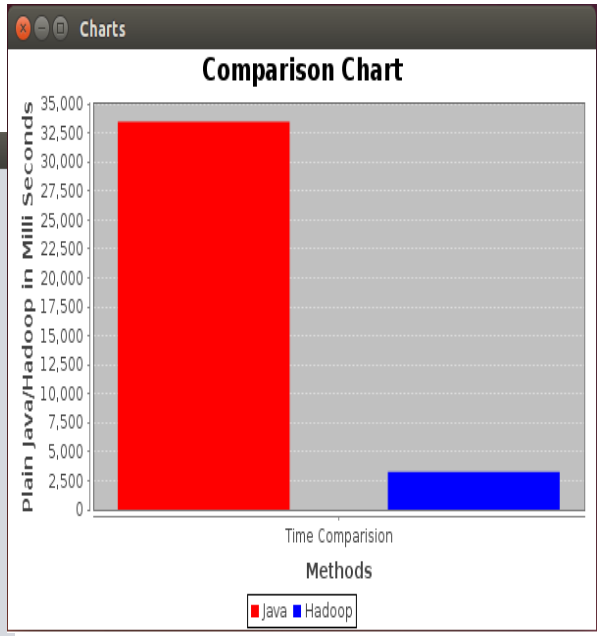


Fig 6.-shows result if movie not found in data set

Figure(6) shows the result of movies search and gives the output if move is not present in a data set. search is made in data set and if movie is not available in data set then this message box appear.

3.COMPARISON GRAPH OF TIME REQUIRED TO SEARCH A PARTICULAR MOVIE USING JAVA AND HADOOP.



CONCLUSION

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, search, processing and computation of big data. This system include study of existing techniques for computing and accessing of big data problem.

Using Map reduce frame work, this system implemented Clustering Approach for Collaborative Filtering. It observe that from above study, the time required to process big data using hadoop is better than other existing tool.

REFERENCES

[1]. R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol.

[2]. A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033,2012.

[3]. S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[4]. J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 20

[5]. G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data,pp. 1015-1018, 2009.

[6]. S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to-End Mining," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08),pp. 512-521, 2008.

[7]. G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179,2007.

[8]. C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K.Olukotun, "Map-Reduce for Machine Learning on Multicore,"Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS'06), pp. 281-288, 2006.

[9]. D. Gillick, A. Faria, and J. DeNero, MapReduce: DistributedComputing for Machine Learning, Berkley, Dec. 2006.

[10]. Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining,"J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[11]. E. Birney, "The Making of ENCODE: Lessons for Big-DataProjects," Nature, vol. 489, pp. 49-51, 2012.

[12]. J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[13]. S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[14]. J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[15]. D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[16]. E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for MiningLarge-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia,(MM '09,) pp. 917-918, 2009

[17]. http://en.wikipedia.org/wiki