INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SURVEY ON BIG DATA USING DATA MINING

AYUSHI V. RATHOD, PROF. S. S. ASOLE

BNCOE, Pusad Department of CSE

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: In this paper presents Big Data Problems 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. 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 pipeline, whether 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 managing 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 etc. 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 various situations, depending upon these techniques, we will try to implement concepts related to Big Data processing.

Keywords: Data Mining Challenges With Big Data*, Hadoop's Distributed File System,* Map-Reduce Framework, Clustering Approach for Collaborative

Corresponding Author: MS. AYUSHI V. RATHOD

Access Online On:

www.ijpret.com

How to Cite This Article:

Ayushi V. Rathod, IJPRET, 2015; Volume 3 (9): 334-339

PAPER-QR CODE

INTRODUCTION

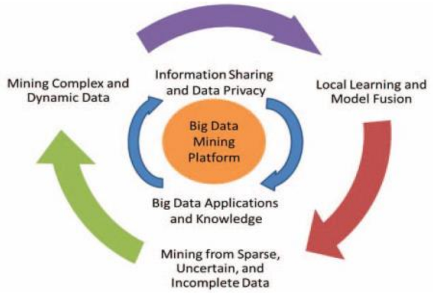
Big data is nothing the large amount of data. There is broad recognition of the value of data as well as products obtained through analyzing it [2]. About big data "Size is the only thing that matters." Various Popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist the New York Times [5], 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.

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 [1]. 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.

We are used to thinking of 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.DATA MINING CHALLENGES WITH BIG DATA:-


Mining Complex and Dynamic Data
Information Sharing and Data Privacy
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

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 clusterUsers 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.

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

A MapReduce 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

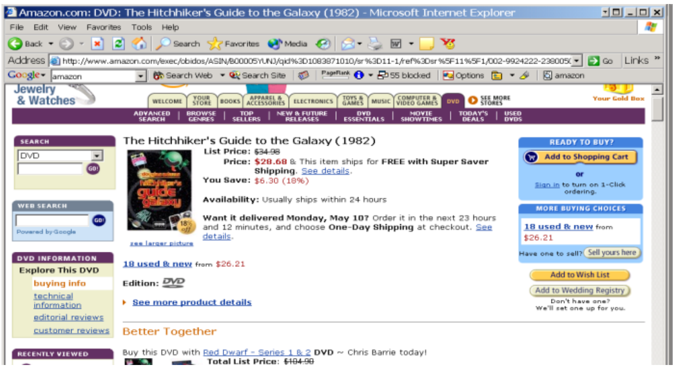
domain than the output keys and values. Furthermore,the intermediate keys and values are from the same domainas the output keys and values.

IV.CLUSTERING APPROACH FOR COLLABORATIVE:-

To overcome some drawbacks we are going to 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.


Amazon.com: DVD: The Hitchhiker's Guide to the Galaxy (1982) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.amazon.com/exec/obidos/ASIN/B00005YUNJ/qid%3D1083871010/sr%3D11-1/ref%3Dsr%5F11%5F1/002-9924222-238005 Go Links
Google amazon Search Web Search Site PageRank 55 blocked Options amazon
Jewelry & Watches
WELCOME YOUR STORE BOOKS APPAREL & ACCESSORIES ELECTRONICS TOYS & GAMES MUSIC COMPUTER & VIDEO GAMES DVD SEE MORE STORES
Your Gold Box
ADVANCED SEARCH BROWSE GENRES TOP SELLERS NEW & FUTURE RELEASES DVD ESSENTIALS MOVIE SHOWTIMES TODAY'S DEALS USED DVDS
SEARCH
DVD
GO!
WEB SEARCH
GO!
Powered by Google
DVD INFORMATION
Explore This DVD
buying info
technical information
editorial reviews
customer reviews
RECENTLY VIEWED
see larger picture
READY TO BUY?
Add to Shopping Cart
or
Sign in to turn on 1-Click ordering.
MORE BUYING CHOICES
18 used & new from $26.21
Have one to sell? Sell yours here
Add to Wish List
Add to Wedding Registry
Don't have one? We'll set one up for you.

The Hitchhiker's Guide to the Galaxy (1982)

List Price: ~~$34.98~~
Price: $28.68 & This item ships for FREE with Super Saver Shipping. See details.
You Save: $6.30 (18%)

Availability: Usually ships within 24 hours

Want it delivered Monday, May 10? Order it in the next 23 hours and 12 minutes, and choose One-Day Shipping at checkout. See details.

18 used & new from $26.21

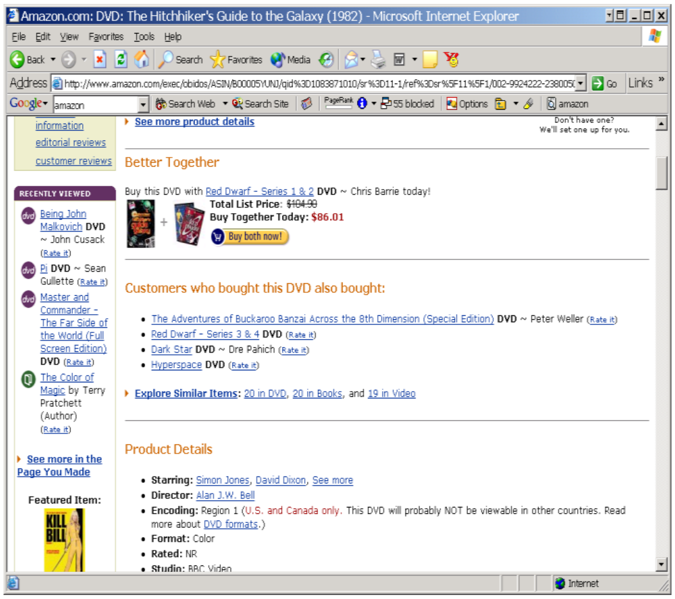
Edition: DVD

See more product details

Better Together

Buy this DVD with Red Dwarf - Series 1 & 2 DVD ~ Chris Barrie today!
Total List Price: ~~$104.90~~

Fig 1. Everyday Examples of Collaborative Filtering


Amazon.com: DVD: The Hitchhiker's Guide to the Galaxy (1982) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.amazon.com/exec/obidos/ASIN/B00005YUNJ/qid%3D1083871010/sr%3D11-1/ref%3Dsr%5F11%5F1/002-9924222-238005(Go Links
Google amazon Search Web Search Site PageRank 55 blocked Options amazon
Internet

information
editorial reviews
customer reviews

See more product details

Don't have one?
We'll set one up for you.

RECENTLY VIEWED

Being John Malkovich DVD ~ John Cusack (Rate it)
Pi DVD ~ Sean Gullette (Rate it)
Master and Commander - The Far Side of the World (Full Screen Edition) DVD (Rate it)
The Color of Magic by Terry Pratchett (Author) (Rate it)

See more in the Page You Made

Featured Item:

Better Together

Buy this DVD with Red Dwarf - Series 1 & 2 DVD ~ Chris Barrie today!
Total List Price: ~~$104.98~~
Buy Together Today: $86.01
Buy both now!

Customers who bought this DVD also bought:

- The Adventures of Buckaroo Banzai Across the 8th Dimension (Special Edition) DVD ~ Peter Weller (Rate it)
- Red Dwarf - Series 3 & 4 DVD (Rate it)
- Dark Star DVD ~ Dre Pahich (Rate it)
- Hyperspace DVD (Rate it)

Explore Similar Items: 20 in DVD, 20 in Books, and 19 in Video

Product Details

- Starring: Simon Jones, David Dixon, See more
- Director: Alan J.W. Bell
- Encoding: Region 1 (U.S. and Canada only. This DVD will probably NOT be viewable in other countries. Read more about DVD formats.)
- Format: Color
- Rated: NR
- Studio: BBC Video

Fig 1.1.Everyday Examples of Collaborative Filtering

CONCLUSION:-

In this paper, some of the most important methods for Big Data are described. Thus we studied existing techniques for various problem.Depending on these techniques like Map Reduce, we are going to proposed Clustering Approach for Collaborative Filtering.

REFERENCES:-

1. R. Ahmed and G. Karypis, “Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

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, 2011.

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, 200.