International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.19-22
ISSN: 2278-2419

Service Level Comparison for Online Shopping using Data Mining

K.Subhasree Chandini.[1], A.Roshini[2], A.Kokila [3], B.Aishwarya[4]
[1]Information Technology, Rajalakshmi Institute of Technology, subhashreechandini.k.2012.it@ritchennai.edu.in.
[2]Information Technology, Rajalakshmi Institute of Technology, roshini.a.2012.it@ritchennai.edu.in
[3]Information Technology, Rajalakshmi Institute of Technology, kokila.a.2012.it@ritchennai.edu.in

Abstract - The term knowledge discovery in databases (KDD) is the analysis step of data mining. The data mining goal is to extract the knowledge and patterns from large data sets, not the data extraction itself. Big-Data Computing is a critical challenge for the ICT industry. Engineers and researchers are dealing with the cloud computing paradigm of petabyte data sets. Thus the demand for building a service stack to distribute, manage and process massive data sets has risen drastically. We investigate the problem for a single source node to broadcast the big chunk of data sets to a set of nodes to minimize the maximum completion time. These nodes may locate in the same datacenter or across geo-distributed data centers. The Big-data broadcasting problem is modeled into a LockStep Broadcast Tree (LSBT) problem. And the main idea of the LSBT is defining a basic unit of upload bandwidth, r, a node with capacity c broadcasts data to a set of [c=r] children at the rate r. Note that r is a parameter to be optimized as part of the LSBT problem. The broadcast data are further divided into m chunks. In a pipeline manner, these m chunks can then be broadcast down the LSBT. In a homogeneous network environment in which each node has the same upload capacity c, the optimal uplink rate r, of LSBT is either c=2 or 3, whichever gives the smaller maximum completion time. For heterogeneous environments, an O(nlog2n) algorithm is presented to select an optimal uplink rate r, and to construct an optimal LSBT. With lower computational complexity and low maximum completion time, the numerical results shows better performance.The methodology includes Various Web applications Building and Broadcasting followed by the Gateway Application and Batch Processing over the TSV Data after which the Web Crawling for Resources and MapReduce process takes place and finally Picking Products from Recommendations and Purchasing it.

Key words - Classification, Clustering, and Data mining.

I. INTRODUCTION

The Big Data era has arrived. Every day, quintillion bytes of data are created and 90 percent of the data today were produced within the last two years. Ever since the invention of the information technology, the capability for data generation has never been so powerful and enormous. As another example, the first presidential debate on 4 October 2012, between President Barack Obama and Governor Mitt Romney hit more than 10 million tweets within 2 hours. Such discussions provide a means to sense the public interests and give feedback in real-time, and are mostly appealing compared to radio or TV broadcasting. Another one is Flicker, where pictures are shared in public, received an average of 1.8 million photos per day, from February to March 2012 [1]. Assuming the size of each photo is 2 megabytes (MB), and it requires 3.6 terabytes (TB) storage every day. The rise of Big Data where data sets has grown tremendously beyond the control of common software tools is demonstrated by the above examples. Exploring the large volumes of data and extracting the useful information only or knowledge for future actions is the most fundamental challenge for Big Data applications, which is very efficient and close to real time in many situations, because storing all the data sets is nearly infeasible. An example, the square kilometer array (SKA) in radio astronomy is of 1,000 to 1,500 15-meter dishes in a central 5-km area, which provides 100 times more sensitive vision than any other existing radio telescopes, that answers the fundamental questions about the Universe. Whereas, with a 40 gigabytes (GB)/second data volume, the data from the SKA are exceptionally large. Even the researchers have confirmed the interesting patterns, like transient radio anomalies, that can be discovered from the SKA data, the existing methods can only work in offline fashion. These are incapable of handling the Big Data scenario in real time. The result is that the unprecedented data volumes needs an effective data analysis platform and prediction platform to achieve faster response and real-time classification for such Big Data[2][4].

II. LITERATURE

A. Existing System

Existing Systems only provide users, with the products in their stocks and will render the Comparison within their products only. Thereby limiting the users to analyze before buying a product. Existing Service Recommender Systems suffers from big data problems like scalability and time consumption and thus lack of preciseness.

B. Disadvantage of existing system

Limited analysis before buying. Comparison within their products only and no deeply analyze on what product to choose and in which Application. More time consumption.

C. Proposed system

We propose a scalable, efficient and precise service comparison and recommender system which enables the shoppers to deeply analyze on what product to choose and in which application, ease and fair with our Gateway. The shoppers will be provided with clean indexes of various products with its specification, cost and also service ratings which are done in a statistical way. Our system uses the data from various web application and loads in its datasets collaboratively and process with batch jobs so as to categories classify and to index the data's in a distributed and parallel processing manner.Shoppers can analyze, get recommendations and can pick products and add to cart irrespective of the service provider. Hence our application

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.19-22
ISSN: 2278-2419

stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed whenever the shopper wants the product. All the information will be securely and precisely stored in the users' session. The purchase phase look up for the web services of the products service provider and can make the online payment with the banks from service provider. Once it got over process gets back to our gateway bringing out the track Id's from product service provider.

III. RESULTS AND DISCUSSIONS

A. Workflow

The Tab Separated Value (uncategorized) files are categorized and distributed to the web servers with offers and price quotes. Web crawling process takes place between the file system and service provider. The file system performs batch jobs. From the service provider the user chooses the products, purchase and complete the transaction as given in Fig 1.

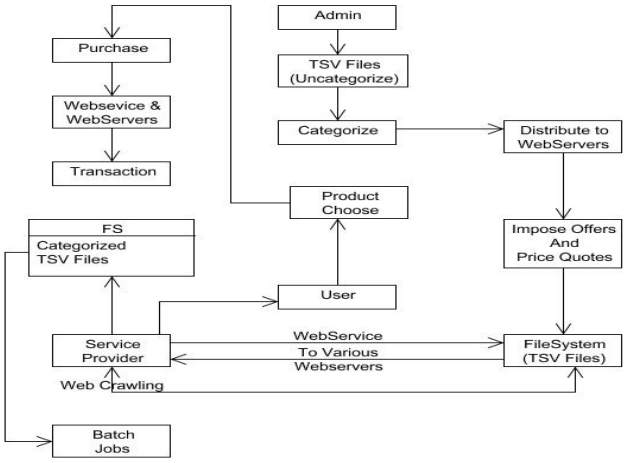

Fig. 1. Work Flow

B. Use case

It describes the actors involved and the roles played by them. It also describes a sequence of actions that provide something of measurable value to an action and is drawn as a horizontal ellipse. Following fig 2 is the use case diagram for the above proposal

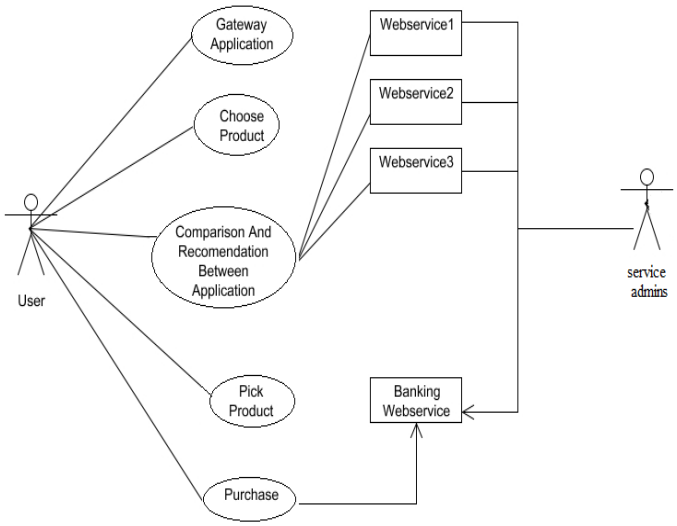

Fig 2 Use case Diagram

C. Class diagram

Figure 3 is the Class diagram which shows the classes, inter-relationships, operations and attributes of the classes.

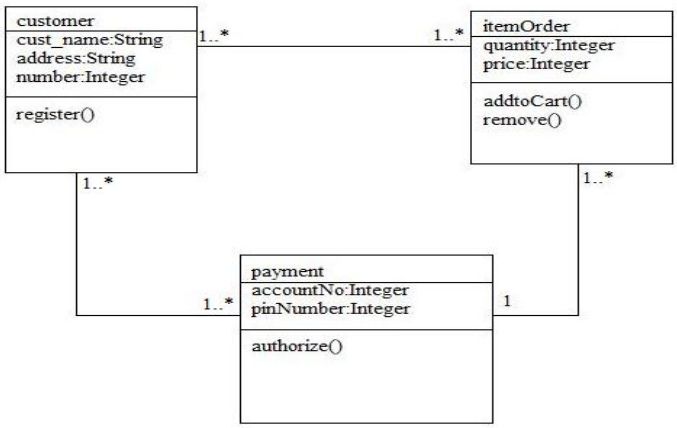

Fig 3 Class Diagram

D. Entity Relationship Diagram

ER diagram shows the entities, relationships and attributes saved in the database. As shown in Fig 4

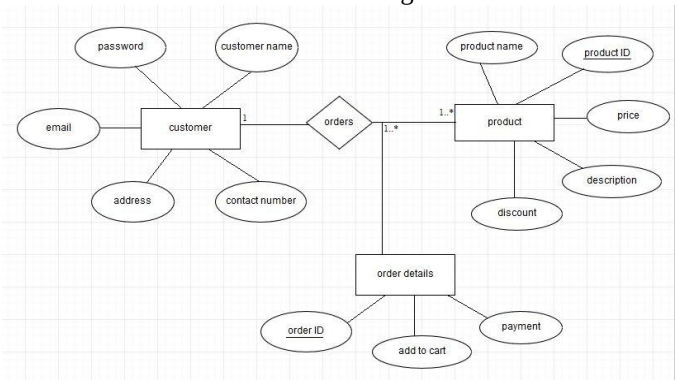

Fig 4 Entity Relationship Diagram

E. System architecture

System architecture is the formal descriptive model that shows the structure of a system, its behavior, and its views. For huge data sets, map reduce is done and the categorized tab separated values are taken as datasets. In the system architecture only the request process and web crawling process is given. Figure 5 explains the system architecture of the proposed.

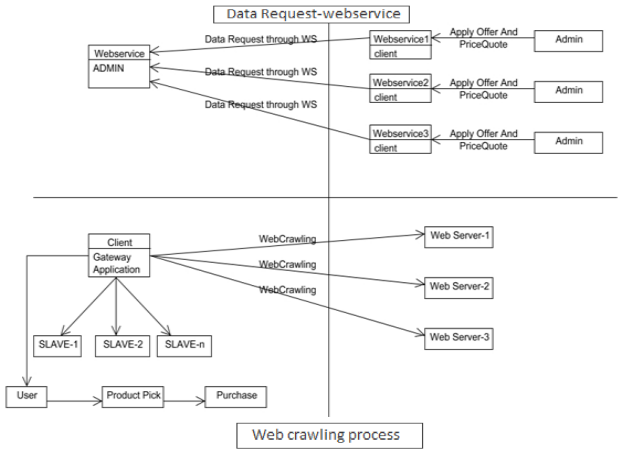

Fig 5 System Architecture

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.19-22
ISSN: 2278-2419

E. DFD

DFD show the flow of data from external entities into the system, showed how the data moved from one process to another, as well as the logical storage. Figure 6, 7 & 8 explains DFD level 0, 1 & 2.

Fig 6 DFD – Level 0

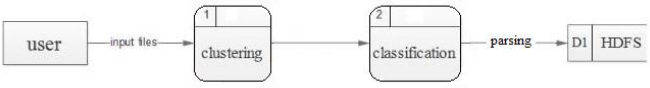

Fig 7 DFD – Level 1

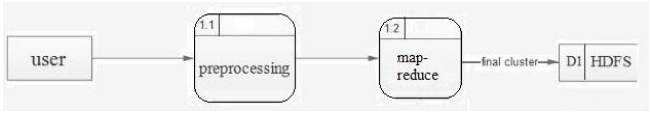

Fig 8 DFD – Level 2

IV. MATERIALS AND METHODS

The modules description is given here

A. Various Web applications Building and Broadcasting

Sample web applications were built so that the users can compare their products with different Service Providers. The application uses sample datasets that has been crawled in previously. Similar datasets were prepared for other applications too using the Meta model that has been crawled earlier. Each dataset was loaded independently in various web applications. Features and other specifications have been loaded differently for each application based on the service provider's requirement. These applications have been deployed in web servers so that the application is up and running. Web services have been written on each web application so that any third party can communicate with secure authentication.

- Pre-processing
- Clustering
- Classification
- Distribution

B. Gateway Application and Batch Processing over the TSV Data

Now our gateway application is built which gives users with recommendations and comparisons between the products in the market. Generally the resources provided by various web servers are in TSV (Tab Separated Values) format and should be batch processed before proceeding. For that we use our own API for TSV Manipulation. The TSV files were parsed for data. These data are used for further processing (i.e. For recommendation and comparison).

- Admin Login
- Conversion of Text to Object for Info
- Conversion of Text to Object for Features

C. Web Crawling for Resources and MapReduce

The users can register and can login to view various products available in market. This is done by writing a web service client process for each service provider. It can connect to the various web applications' web service and can pull all the needed data to our backend. A huge amount of data got accumulated now. Web crawling looks for web services provided by various web applications. The crawled resources are then reduced by MapReduce framework and converted into a single object. This reduced object contains all the necessary information for providing comparison and recommendations.

- User Registration
- User Login
- MapReduce [3]
- Product List

Map Reduce – It is a combination of a map task and multiple reduce tasks which is used for clubbing all the slave process outcomes.

D. Picking Products from Recommendations and Purchase

The recommendations were given based on the QOS, availability, delivery, offers, price and specifications of the particular product. The users can pick any product so that our application provides with a most genuine recommendation and a set of comparisons. The users are provided with neat and clean indexes so that he can pick a best provider for a particular product. The picked products were added in cart and can be purchased later. The user cart is equipped with Case-Based Reasoning (CBR) to identify and recommend the items that seem more suitable for completing a user's buying experience provided that he or she has already selected some items. The system models complete transactions as cases and recommended items come from the evaluation of those transactions. Because the cases aren't restricted to the user who purchased them, the developed system can generate accurate item recommendations for joint item selections, both for new and existing users. Having analyzed the previous transactions and identified the concepts within which concrete items appear, the given part of a new transaction is matched over the existing ones to find the more adequate solution. i.e. the best way to fill this basket. When the user initiates transaction our gateway will connect to the banking web services directly on behalf of the service provider and completes the transaction securely with help of OTP sent to their mail id given on user registration. The process will be back to our application as soon as the transaction is over and the purchased products will be reflected on the Bag List. i.e. purchased items list.

- Product Comparison [5]
- Add to Cart
- Case Based Recommendation
- Purchase
- User Bag

V. CONCLUSION

Thus the Broadcasted data are received using Web services through SOAP protocol and batch processed Hadoop. Map Reduce is done for batch job's output and the service level comparison is shown for the selected products. Recommendations are also given using Case based Recommendations and the transaction process is made. Hence we proposed a scalable efficient error prone system with Hadoop for big data from various providers through broadcasting technique.

Reference

[1] R. E. Bryant, R. H. Katz, and E. D. Lazowska, "Big-data computing: Creating revolutionary break throughs in

commerce, science, and society," In Computing Research Initiatives for the 21st Century., 2008.

[2] A. Szalay and J. Gray, "2020 computing: Science in an exponential world," Nature 440, 413-414, March, 2006.

[3] G. Brumfiel, "High-energy physics: Down the petabyte highway," Nature 469, 282-283 January, 2011.

[4] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Proc. of Operating Systems Design and Implementation (OSDI), 2004.

[5] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, , and R. E. Gruber, "Bigtable: A distributed storage system for structured data," Proc. of Operating Systems Design and Implementation (OSDI), 2006.

[6] W. D. Hillis and G. L. Steele, Jr., "Data parallel algorithms," Communications of the ACM, vol. 29, pp. 1170–1183, December 1986.

[7] U. Rencuzogullari and S. Dwarkadas, "Dynamic adaptation to available resources for parallel computing in an autonomous network of workstations," Proc. of ACM SIGPLAN PPoPP, 2001.

[8] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, "Managing data transfers in computer clusters with orchestra," Proc. of ACM SIGCOMM, pp. 98–109, 2011.

[9] D. Nukarapu, B. Tang, L. Wang, and S. Lu, "Data replication in data intensive scientific applications with performance guarantee," IEEE Transactions on Parallel and Distributed Systems, aug. 2011.

[10] C. Peng, M. Kim, Z. Zhang, and H. Lei, "Vdn: Virtual machine image distribution network for cloud data centers," Proc. of IEEE International Conference on Computer Communications (INFOCOM), 2012.