International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 55-62

A COMPREHENSIVE REVIEW OF RECOMMENDATION SYSTEMS ON MAP REDUCE IN BIG DATA ENVIRONMENT

Manish Kumar Kakhani[1] and Anil Kumar[2]

[1]*Department of Computer Science and Engineering Mody University of Science and Technology Lakshmangarh, Sikar, Rajasthan, India. Email: manishkakhani@gmail.com*

[2]*Senior Member, IEEE, Department of Computer Science and Engineering Mody University of Science and Technology Lakshmangarh, Sikar, Rajasthan, India. Email: dahiyaanil@yahoo.com*

Abstract: Recommendation system is a very valuable tool for providing appropriate suggestions to the users from the large space of possible options. Recommendation system has redefined our lives in completely different manner. With the explosion of data online and offline in last one decade creates the huge possibility for better and effective recommendation system in terms of scalability and accuracy. Presently, the majority of existing recommendation systems work well with structured data but with Big Data, the implementation of recommendation system generates new opportunities of research. The existing recommendation systems have scalability and inefficiency issue when processing or analyzing Big Data due to distributed processing of data in cluster of thousand nodes. Map Reduce is a popular programming framework to process and analyze Big Data that made it possible to implement various recommendation systems in Big Data environment. The paper presents various recommendation systems on Map Reduce in Big data environment and gives details about challenges and possibilities of implementing recommendation systems in Big Data environment.

Keywords: Recommendation System, Big Data, Hadoop, Map Reduce, Mahout.

1. INTRODUCTION

We are living in an information era where data is very valuable asset to any organization in today's digital world. We are generating data at an explosive rate. Today data is coming from variety of sources like social media sites, sensor devices, mobile phones etc. This data is beyond the processing capability of current data processing system is known as Big Data [1, 4]. Big Data is best defined by three V's-Volumes, Velocity and Variety. Volume deals with the size of data in Terabytes, petabytes or beyond. Velocity deals with the rate at which data is generating around the globe. Variety deals with all format of data – structured, semi structured and unstructured [2, 3, 5]. The biggest challenge among the decision makers is to analyze Big Data for better insight and improved decision making for more profitability and it is a key to competitive advantage [6, 7].

Recommendation system is a system that guides the user in a very special way to give fascinating, helpful and constructive services from a large space of possible options. A Recommendation system filters our world by removing all the unnecessary options available and showing only the most useful and interesting options to us [8]. The field of recommendation system is a very beautiful example of predictive analytics. Recommendation system are used in variety of areas including online shopping, movies, music, news, books, research articles, search queries and social media. The two main industries benefitting from the recommendation system is retail industry and media industry.

Recommendation System has gained the special attention in the Big Data research community and it is becoming extremely popular everyday in the world of Big Data. Recommendation systems for big data applications is promising research area with the rapid growth of number of internet users, online services and online information [9]. The accuracy of

recommendation increases with the increase in the data. At the same time, recommendation system has to deal with scalability issue to handle the growing data in big data environment. Big Data creates new challenges and opportunities for recommendation system too [10]. With Big Data, recommendation services will improve for the end users as data space increases for analysis. The accuracy and scalability of recommendation system become important research issue with Big Data.

The merging of Recommendation system with Big Data technologies give birth to new horizon of research problems and opportunities. The implementation of many existing recommendation system on Map Reduce framework can handle Big Data in effective manner. With the development of tools like Apache Hadoop, Apache Mahout and Apache Spark, it becomes possible to design and implement scalable and accurate recommendation system in Big Data environment.

The remainder of the paper is organized as follows: Section II presents various types of recommendation systems. Then components of recommendation system are discussed in Section III. Section IV described performance metrics of a recommendation system. In Section V, the related work of recommendation system in the Big Data environment is discussed. The tools for implementation of recommendation system in Big Data environment are explained in Section VI. In Section VII, the research scope of recommendation system in Big Data environment is explored. Finally, Section VIII concludes the paper and provides prospect of new work.

2. TYPES OF RECOMMENDATION SYSTEM

Recommendation Systems has evolved as separate research area in mid 1990s. Over last two decades various recommendation systems have been evolved. The Recommendation Systems are mainly categorized into three categories [10].

1. *Content based Recommendation System:* It recommends services or products to user that is based on services or products like by user in the past. A set of keywords are used to describe an item and a user profile is built from these keywords based to choices and likes of the user in the past. User's browsing history, likes, purchases and ratings are taken into account before providing recommendations. For Example, A user U likes product X, Y in the past and product Z is of similar attributes as X and Y. It is likely that User U will like the product Z and hence it will be recommended to user U [11]. The Content based recommendation method is illustrated in Figure 1 below.

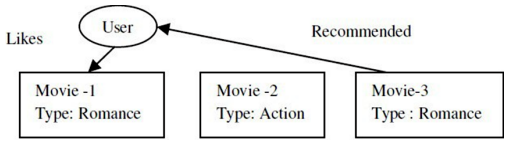

Figure 1: Example of Content based Recommendation System

2. *Collaborative filtering Recommendation System:* It is the most successful and commercially accepted recommendation system and implemented by many organizations. It recommends services to the user that other similar user like in the past. All the similar users choice is compared and a user gets a recommendation. For example, user U1 likes product A, B, C, D and User U2 likes the product A, B, C, D and E. Then it is likely that user U1 will also like product E and E is recommended to user U1 [8]. The collaborative filtering recommendation method is illustrated in Figure 2 below:

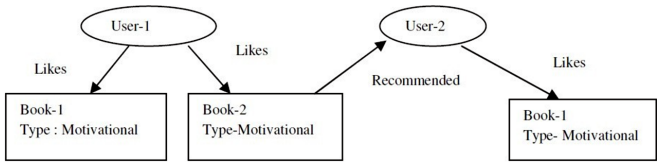

Figure 2: Example of Colloborative Filtering Recommendation System

3. Hybrid Recommendation System: The two techniques discussed above have their strengths and weaknesses. The hybrid model combines content-based and collaborative filtering recommendation approach in several different

ways to get more effective and accurate recommendation results by overcoming the shortcomings.

Other than above three major categories, there are various recommendation system developed by the academia and industry with different flavors. Context-aware recommendation system considers contextual information, such as time, place and the company of other people. Risk-aware recommendation system takes into account the risk of disturbing the user in specific situation like during a professional meeting, early morning, late-night. Personality-based recommendation system analyzes social media data in order to predict a user's personality and to subsequently derive its personality-based product preferences.

3. COMPONENTS OF RECOMMENDATION SYSTEM

A Recommendation system is composed mainly of four components [12] and it is illustrated in figure 3.

1. Data collection and processing: This is the first step in the recommendation system. It mainly deals with collection of raw data from various data sources like log files, data bases etc.
2. Recommendation Model: It is the core component of the recommendation system. It contains mainly the recommendation algorithm that will recommend the services or products to the user from large set of items.
3. Business Logic and Analytics: It is used to apply certain business logic and analytics to restrict the output recommendation. It is used not to recommend certain types of items to particular users and increase the accuracy of recommendation to users by applying some filtering operations.
4. User interfaces: This is the final step and it is about the presentation of recommendation. It is used to show the output according to need of the user. The information required by different users vary according to their need. The same user in different context have different information need. It explains the users why they are being recommended an item, for example you may like product X because you liked product Y, Z. This step makes the recommender's decisions more transparent.

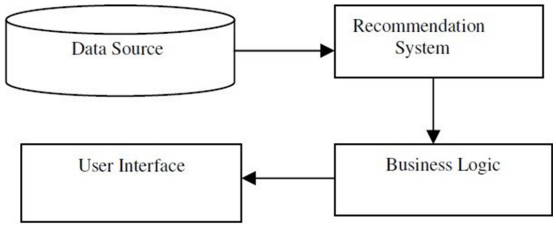

Figure 1: Example of Content based Recommendation System

4. PERFORMANCE METRICS OF RECOMMENDATION SYSTEM

A good recommendation system has following characteristics [13, 14].

1. *Accuracy:* The purpose of any recommendation system is to predict the most accurate recommendation for its end user according to his requirements. For example, a travel recommender system gives you places you have already traveled to. Then such recommendation system would be a poor recommendation system. A good recommendation system should predict most accurate and appropriate recommendation to the users. Recommendations are based on prediction and approximation, the scope of improvement in accuracy is significant.
2. *Speed:* Speed is one of the top performance bottleneck of any recommendation system. Speed means quickness of any system. It means system should give correct result in stipulated time. If the recommendation system is taking too long time to respond to user query, then that recommendation system is not a good recommendation system.
3. *Scalability:* A recommendation system should deal with large amount of data and increasing data effectively. Scalability means system should adapt easily to more volume of data with same

performance. As data is continuously increasing in big data environment, the recommendation system should give accurate result even with increase volume of data.

4. *Cold start ability:* The cold start problem is defined as the incapacity of recommendation system to provide recommendations for new users as enough data is not available. The recommendation system should be efficient enough to deal with cold start problem and it has ability to start making good prediction for the new user.
5. *Sparse data handling:* When large numbers of users do not rate the large number of products and there are products that haven't been rated by large number of users then it causes the problem of spare data in User-product matrix. The recommendation system should still be good enough to give better prediction in case of large amount of sparse data.

5. RELATED WORK

The development of recommendation system as separate research area started way back in mid 1990s.

There have been many recommendation systems developed in last two decades. In [10], authors proposed a keyword-aware service recommendation system named KASR. In KASR, Keyword are used to indicate user's preferences and a user based collaborative filtering is adopted to generate appropriate recommendation. A more recent research from Zhiwei Yu et. al., [15] proposed three new algorithms for context aware service recommendations based on role-mining. The algorithms were implemented on Map Reduce style take advantage of popular distributed computing platform. Evan Casey [16] studied an algorithmic framework built on top of Apache Spark for parallel computation of the neighborhood-based collaborative filtering problem, which allows the algorithm to scale linearly with a growing number of users. In [17], Zhi- Dan Zhao addressed the problem of representing Collaborative filtering on Map Reduce framework in parallel manner by dividing the calculation process by user ID, calculating the recommendation process for each user. In [26], Hao Wang et. al., applied the hierarchical Bayesian model deep learning technique in collaborative filtering based recommendation algorithm to improve the efficiency and accuracy of recommendations. The more detailed discussion on memorization and generalization of recommendation system was presented on wide and deep learning for recommendation system by jointly training wide linear models and deep neural networks in [27]. In [28], authors addresses the problem of cold start in collaborative filtering recommendation and proposed a new deep- content based recommendation algorithm that uses a predictive latent factor model for recommendation and applied the model on large scale music data files. Jan Zahalka et. al., proposed interactive and multimodal content-based recommendation algorithm for venue recommendation with the help of social media platforms in [29]. In [30], authors addresses the issue of learning representation features of images over large social networks and proposed a novel deep model that learns the combined feature representations for both users and images. This is done by changing the heterogeneous user-image networks into homogeneous low-dimensional representations. Shuiguang Deng et. al., [31] proposed a two phase trust-aware recommendation system using deep learning techniques to address the problem of varying recommendation on social networks by own characteristics and friends recommendations. Paulo Chiliguano and Gyorgy Fazekas [32] proposed a hybrid deep learning based recommendation algorithms for music recommendation considering real world information and high level representation of audio files. In [33], authors addresses the problem of automatic understanding and discrimination of users liking for images and presented a deep bi-modal knowledge representation of images based on their visual content and associated tags.

6. IMPLEMENTATION OF RECOMMENDATION SYSTEM IN BIG DATA ENVIRONMENT

Hadoop is a very popular tool to process big data in parallel and distributed manner on thousands of nodes in a cluster [20]. Hadoop is open source software

framework, java based implementation of Google technologies-GFS and Map Reduce for distributed storage and distributed processing of very large data sets [1, 2]. It is part of the Apache project sponsored by the Apache Software Foundation.

It emerged as de-facto standard for Big Data Processing. Hadoop is reliable, Scalable and fault tolerant. Hadoop follows a Master/Slave architecture. There is one master node and several slave nodes. The two core components of Hadoop are HDFS and Map Reduce. Hadoop distributed file system (HDFS) is the storage component of Hadoop. Map Reduce is the processing component of Hadoop. In HDFS, Namenode is master node and Data nodes are slave nodes. In Map Reduce, Job tracker is Master node and Task trackers are slave nodes. Name node maintains the namespace and stores the metadata about all the data nodes. Data Node stores the replica of block of data. By default, 3 replicas are created for each block of data which can be further configurable. Job Tracker is responsible for splitting the input data, scheduling of tasks and monitoring of execution process. Task Tracker: does the actual processing of data and periodically sends the heartbeat message to jobtracker for progress of task completion [1, 2].

Map Reduce is a programming model and a framework for processing large data sets in distributed manner [21]. It is very simple, powerful and easy to use framework. It was developed by Google and is very popular due to its open source implementation Hadoop. It is widely used today throughout the world to process big data. Map Reduce follows a functional programming approach derived from LISP programming language. Every job is divided into two tasks-Map and Reduce. Map task reads one data chunk and processes it produce intermediate results (key-value pairs) and Reduce task fetch the intermediate results and carry out further computations to produce the final results. The Map Reduce framework provides great abstraction to the user. The programmer will only write Map Reduce programs and Map Reduce framework takes care of the every internal processing including partitioning the input data, scheduling of tasks across various nodes of cluster, handling failures, and communication between nodes. Map Reduce programming model can be used to develop recommendation system for big data applications. To deal with large data sets, it should be divided into smaller chunks of data and stored in distributed manner and processing to be applied on it. This task is well performed by Map Reduce framework. Map Reduce framework is very useful to deal with large data sets in recommendation system. A recommendation system can be implemented on Map Reduce framework using Map Reduce programming and able to handle big data in proper manner. If we want to take advantage of Map Reduce for recommendation system than we have to write our recommendation system in map reducible manner. Map Reduce framework can be useful to address the problem of pairwise similarity of users and similarity and correlation of data items. It provides fault tolerance, reliability and scalability which are good for recommendation system.

Apache mahout is a Java based open source machine learning library runs on the top of Hadoop framework to develop scalable machine learning algorithms in distributed environment using the Map Reduce model [23, 24]. The project is developed and supported by Apache Software foundation. Apache mahout supports mainly four categories of machine learning algorithms-Recommendation algorithms, Clustering algorithms, Classifications algorithms, Frequent item set mining. The Hadoop cluster stored Big data in its storage layer Hadoop distributed File System (HDFS) and Mahout provides effective mechanism to apply various machine learning algorithm over the data stored in Hadoop cluster and find meaningful patterns in those data sets. Mahout provides quicker and efficient ways to find information from the large data sets stored in Hadoop cluster in distributed manner. Mahout provides flexibility to the programmer to apply a ready-to-use framework for various machine learning tasks on large data sets.

There are rich set of functions available in Mahout Library to implement recommendation system on Map Reduce in Big Data environment. Mahout provides functionality to implement Collaborative Filtering recommendation algorithms which is one

of the most commercially accepted Recommendation algorithm by the industry. There are two broad categories of recommendation algorithms under collaborative filtering approach-user-based and item-based recommendation algorithms, both are well supported in Mahout.

Apache Hadoop, Apache mahout, Apache spark and Neo4j are key tools to implement scalable and efficient recommendation systems in big data environment [25]. Apache Mahout provides a rich set of functions from which one can build a better recommendation system in big data environment. Hadoop is a batch processing system and it is unable to deal with real time streaming data sets. Apache spark is the solution to this problem. Apache spark provides flexible library to implement rich set of machine learning algorithms including recommendation system over real time streaming data sets. Neo4j is a very popular java-based open- source graph database and stores data in graph structure rather than tables. Neo4j gives appropriate recommendation by using richer set of graph information.

7. RESEARCH SCOPE OF RECOMMENDATION SYSTEM IN BIG DATA ENVIRONMENT

Big Data refers to large and complex data sets that are beyond the processing capabilities of current tools and technology within acceptable time period. Big Data has three major characteristics-Volume, Velocity and Variety [1, 2, 3]. The traditional tools for building recommendation system on big data face some serious challenges. The new recommendation system should work well with big data in a distributed environment as big data is processed in distributed environment. The big data is high in volume so our recommendation system should be scalable enough to deal with high volume of data and able to deal with increasing volume of data with time. The big data is generating at very fast rate in real time so the recommendation system should handle the fast generating real time data in proper manner and generate relevant recommendation. The big data is of all variety-Structured, semi-structured and unstructured. The recommendation system on Big data framework should be able to analyze all categorized of data whether they are text, video, audio files and predict the consumer needs and recommend proper product and services. The effectiveness of recommendation system to deal with big data and provide fast recommendation to end users proved to be key differentiator for business and would improve the sales of the organization. The selection of appropriate recommendation system will not only reduce time to market of their offerings, but also increase customer satisfaction leading to better business outcomes.

Organization like Google, Amazon, twitter, facebook already using recommendation system over big data. Amazon tracks and stores data on all customers' behavior and activity on the servers. For every click, the user makes, the record of the event is logged into the database. Events are stored for all kind of action like user liking a product, adding product to cart and purchasing a product. Ratings are important because they reveal what the user thinks about the product. Recommendation system takes into account the rating and feedback the users provide Filtering product based on rating and other user data.

To handle big data problem for recommendation system, deep learning techniques are playing significant role in recommendation system. Deep learning is a branch of machine learning that performs learning over large volume of data using multi layer model and gives better results.

8. CONCLUSION

In this research study, we have reviewed various recommendation systems and suggested the solutions on Map Reduce framework in big data environment. The objective of the study is to provide better insight and valuable knowledge to research community interested in recommendation system on Map Reduce framework in Big Data environment. This research study is to review, analyze and examine the scope of research of recommendation system in big data environment. Through our study, we are able to conclude that recommendation system over Big Data generates new challenges and opportunities to research communities both in industry and academia.

Acknowledgment

We are thankful to Mody University of Science and Technology, Lakshmangarh, Rajasthan, India for the valuable support in our research. We are also grateful to reviewers to improve the quality of paper.

References

[1] Jean-François Weets, Manish Kumar Kakhani, Anil Kumar, "Limitations and Challenges of HDFS and Map Reduce", Proceeding of International Conference on Green Computing & Internet of Things, Greater Noida, India, pp 345-349, IEEE, 2015

[2] Manish Kumar Kakhani, Anil Kumar, Jean-François Weets," Research Scope of HDFS and Map Reduce", International Journal of Engineering Technology, Management and Applied Sciences,Volume 3, pp 380-382, September 2015.

[3] Manish Kumar Kakhani, Sweeti Kakhani, S.R. Biradar, "Research Issues in Big Data Analytics", International Journal of Application orInnovation in Engineering and Management, Volume 2, Issue 8, August 2013.

[4] Nada Elgendy and Ahmed elragal, "Big Data Analytics: A literature Review Paper", ICDM 2014, Springer International publishing, pp. 214- 227, 2014.

[5] Amir Gandomi, Murtaza Haider, beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Volume 35, Issue 2, pp. 137–144, April 2015.

[6] Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya, Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, Vol. 79–80, pp 3-15, 2014.

[7] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, Ananth Grama, Trends in big data analytics, Journal of Parallel and Distributed Computing, Vol. 74, Issue 7, pp. 2561-2573, 2014.

[8] Nachiket Sadashiv Bhosale, Sachin S. Pande, "A Survey on Recommendation System for Big Data Applications", Journal Data Mining Knowledge Engineering, Vol. 7, No 1, 2015.

[9] Jinhong Kim1 and Sung-Tae Hwang, "Big Data Platform of a System Recommendation in Cloud Environment", International Journal of Software Engineering and Its Applications, Vol. 9, No. 12 pp. 133-142, 2015.

[10] Shunmei meng, Wanchun dou, Xuyun zhang, and Jinjun chen, "KASR: A Keyword-Aware Service Recommendation Method on Map Reduce for Big Data Applications", IEEE transactions on parallel and distributed systems, Vol. 25, No. 12, December 2014.

[11] Recommender Systems, http://recommender-systems.org/, 2012.

[12] "The Components of a Recommender System, https://buildingrecommenders.wordpress.com/2015/11/10/the- components-of-a-recommender- system/"

[13] "Recommender systems, http://www.cs.carleton.edu/cs_comps/0607/recommend/recommen der/algorithms.html".

[14] "Recommender Systems—It's Not All About the Accuracy, https://gab41.lab41.org/recommender-systems-its-not-all-about-the-accuracy-562c7dceeaff#.ye1gbaszo"

[15] Zhiwei Yu, Raymond K. Wong, Chi-Hung Chi, "Efficient Role Mining for Context-Aware Service Recommendation Using a High-Performance Cluster", IEEE Transactions on Services Computing, 2015.

[16] Evan Casey, Scalable Collaborative Filtering Recommendation Algorithms on Apache Spark, Thesis report, Claremont Mckenna College California, United states, 2014.

[17] Z.D. Zhao and M.S. Shang, "User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop", Proc. Third International Workshop Knowledge Discovery and Data Mining, pp. 478-481, 2010.

[18] Xiwang Yang, Yang Guo, Yong Liu1" Bayesian-inference Based Recommendation in Online Social Networks", IEEE INFOCOM, 2011.

[19] Jeffrey D. Ullman, "Designing good Map Reduce algorithms, XRDS: Crossroads, The ACM Magazine for Students, Vol. 19, No. 1, 2012.

[20] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System", IEEE, 2010.

[21] J. Dean and S. Ghemawat, "Map Reduce: simplified data processing on large clusters," OSDI, 2004.

[22] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System", Proceedings of the nineteenth ACM symposium on Operating systems principles, pp 29-43, 2003.

[23] "ApacheMahout,http://hortonworks.com/apache/mahout/#secti on_1"

[24] "What is Apache Mahout?, http://mahout.apache.org/"

[25] The Right Recommendation System for Big Data (White Paper), Fractal Analytics Limited, 2015.

[26] Hao Wang, Naiyan Wang, Dit-Yan Yeung, "Collaborative Deep Learning for Recommender Systems", ACM, 2015.

[27] Heng-Tze Cheng et. al., "Wide & Deep Learning for Recommender Systems", arXIV, 2016.

[28] Aaron vanden Oord et. al.,, "Deep content-based music recommendation", Proceeding NIPS'13,Pages 2643-2651, 2013.

[29] Jan Zahálka, Stevan Rudinac, Marcel Worring, "Interactive Multimodal Learning for Venue Recommendation", IEEE Transactions on multimedia, Vol. 17, No. 12, December 2015.

[30] Xue Geng, Hanwang Zhang, Jingwen Bian, Tat-Seng Chua, "Learning Image and User Features for Recommendation in Social Networks", IEEE International Conference on Computer Vision, 2015.

[31] Shuiguang Deng, Longtao Huang, Guandong Xu, Xindong Wu, Zhaohui Wu, "On Deep Learning for Trust-Aware Recommendations in Social Networks", IEEE Transactions on neural networks and learning systems, 2016.

[32] Paulo Chiliguano, Gyorgy Fazekas, "Hybrid music recommender using content-based and social Information", ICASSP, 2016.

[33] Sharath Chandra Guntuku, Joey Tianyi Zhou, Sujoy Roy, Weisi Lin, Ivor W. Tsang, "Understanding Deep Representations Learned in Modeling Users Likes", IEEE transactions on image processing, Vol. 25, No. 8, 2016.