Real Time Financial Analysis Using Big Data Technologies

Pradeep Kumar M. Kanaujia
School of Computer Engineering,
KIIT University,
Bhubaneswar, Orissa, India
pkanaujia26@gmail.com

Manjusha Pandey
School of Computer Engineering,
KIIT University,
Bhubaneswar, Orissa, India
manjushafcs@kiit.ac.in,

Siddharth Swarup Rautaray
School of Computer Engineering,
KIIT University,
Bhubaneswar, Orissa, India
siddharthfcs@kiit.ac.in

***Abstract*— Due to advancement in Science and Technologies there are enormous amount of data available on internet. A large volume of structured, semi-structured and unstructured data is being created at a very rapid speed every day from heterogeneous sources like reviews, ratings, feedbacks, shopping details, etc., it is termed as Big Data. This data generated from different users share many common patterns which can be filtered and analysed to give some recommendation regarding the product, goods or services in which a user is interested. Recommendation systems are the software tools used to give suggestions to users on the basis of their requirements. Many people are not so much aware of different profitable and economical alternatives before using their money for goods or services. They are not so intelligent that they can quickly compare and judge that which product or service is better. The presented paper proposed a recommended system for management and utilisation of three components of salary i.e. saving, investment and expenditure. Many savings and investment consulting systems are available but no system provides effective and efficient recommendation regarding management and beneficial utilisation of salary. The advantage of proposed recommended system is that it provides better suggestion to a person for saving, expenditure and investment of their salary which in turns maximises their wealth. Due to enormous amount of data involved, Apache Hadoop framework is used for distributed processing. Apache Mahout is used for analysing the data and implementation of the recommender system.**

Keywords*— *Big Data, Recommender system, Apache Hadoop, Apache Mahout

I. INTRODUCTION

Data is essentially the facts and statistics collected from the operations and services related to a business, government, environment, public, health, transportation, etc. They can be used to evaluate or store information related to a broad area of activities. This raw data can be processed and interpreted to give some meaningful and accurate information. The obtained information can be used for service or process optimization, monitoring systems, developing automated and intelligent systems, research and analytic, etc. An enormous volume of structured, semi-structured and unstructured data is being created at a very rapid speed every day from heterogeneous sources, it is termed as Big Data [1]. These data can be processed to obtain meaningful patterns and useful information for the benefit of business, government units and social enterprises. The term Big Data refers to the data that exceeds the processing or analysing capacity of existing database management systems. The inability of existing DBMS to handle Big Data is due to its large volume, high velocity, pertaining veracity, heterogeneous variety and non-atomic values [2]. To process such a huge and continuously increasing amount of data, there is a need of advanced, fast and efficient technology and mechanism that can acquire, store, manage, analyse and visualise the data. This processing of data is required to extract useful information or value from the voluminous data. A lot of research is being done and is still going on to obtain enhanced tools, techniques and platform for Big Data Analytics [3]. A large amount of data gets generated from different users from heterogeneous sources in the form of reviews, ratings, feedbacks, shopping details, etc. These data share many common patterns which can be filtered and analysed to give some recommendation regarding the product, goods or services in which a user is interested. Recommender systems are the software tools used to give suggestions to users on the basis of their requirements.

The application of big data on which the presented work focusses is Financial Analytics. Financial analysis is done on the salary of persons who needs to make better utilization of their money. People utilises their salary for different purposes according to their requirements. Today no system is available for suggesting a person on how to use their money for saving, where to invest and how to manage expenditures. Few consulting systems are available which provide investment and saving tips but they are not much effective and are much complex. Many people use suggestions from their neighbourhood, friends and colleagues as the basic method to plan their activities with the salary. Many people get help through social media platforms such as Facebook, WhatsApp, LinkedIn, etc. to get idea regarding the savings and investments plans. Overall all these methods are not effective and are of less helpful to people seeking suggestion on planning salary usage. Using the data available from different platforms and users, the proposed recommender system aims to provide simple, effective and efficient suggestion to utilise and manage the salary. The salary usage is divided into three main components that is Expenditures, Investments and Savings. The data related to different person's salary and their utilisation are huge which comes under Big Data and needs advanced data analytics and filtering to get better recommendations. The presented paper proposed a recommender system for management and utilisation of three

components of salary i.e. saving, investment and expenditure. Many savings and investment consulting systems are available but no system provides effective and efficient recommendation regarding management and beneficial utilisation of salary. The advantage of proposed recommender system is that it provides better suggestion to a person for saving, expenditure and investment of their salary which in turns maximises their wealth.

The presented paper is organised in following manner. Section 1 is the introduction of recommender systems which also briefs the characteristics of big data. Section 2 discusses the literature about the recommended system, its categories and gives overview of Hadoop framework and Mahout. Section 3 details about the proposed framework for Real time financial analysis using big data technologies. Section 4 describes the implementation work followed by the conclusion in Section 5.

II. STATE OF ART

Recommendation systems have proved to be very useful and their popularity is increasing day by day. Recommender systems are the software applications that attempt to predict the "preference" or "rating" of an item that a user has not purchased or rated yet and based on that it suggest item to that user. It is used in variety of application area such as suggesting movies, music, recommending books, research articles, news, product, plans and tariffs, etc. The general term "item" used to represent the thing which the system recommends to the interested user [4]. Recommendation systems are the software tools used to give suggestions to users on the basis of their past experiences and requirements. Suggestions includes different decision-making processes [5] such as 'where to invest money', 'how to plan savings', 'where to minimise and maximise expenditures', 'what product to buy', 'which place to visit'. Recommendation techniques can be categorized into following three types: collaborative filtering, content-based filtering and hybrid approach.

A. Collaborative filtering

Collaborative filtering plays significant role in recommendation process and that is why collaborative filtering is the most extensively used method for designing recommendation systems [6]. In this recommendations are made using the past evaluation of a large group of user's data. Basically collaborative filtering techniques are based on gathering and analysis of a huge volume of user's information related to interests, preferences, activities and behaviour. With these information and tastes of a particular user, prediction is made using the similarity with other users [7] [8]. Collaborative filtering works by assuming that people who have an opinion about some item or thing previously will have same opinion in the later occasions. So, it means that they will prefer almost identical types of items as they preferred previously [9]. One popular example used in collaborative filtering is item-to-item filtering i.e. a user who purchased a product 'x' will also purchase product 'y'. This technique is used by Amazon's recommender system [10]. Following Table I show an example of collaborative filtering using three schemes X, Y and Z of a bank. It also contains preferences of four users for these schemes.

TABLE I. A BANK'S THREE INVESTMENT SCHEMES AND PREFERENCES BY USERS

Schemes	Users			
	Ram	**Jay**	**Tina**	**Jill**
X	5	4	3	5
Y	2	5	4	1
Z	5	1	2	**?**

User's preferences are given in range of 5 (high) to 1 (low). High preference means user is likely to invest and low preference means user may not invest in that particular scheme. The task is to find whether Jill will invest in scheme Z or not. Using collaborative filtering, one could be able to find that there is similar pattern in preferences of Ram and Jill. Ram gave high preference to scheme Z so there is more possibility that Jill may also give high preference to scheme Z. Hence, Jill may invest in scheme Z. Some collaborative filtering methods use matrix factorization, a low-rank matrix approximation method [11] [12] [13].Two main classification of collaborative filtering techniques are: memory based and model based methods. Memory based collaborative filtering techniques uses all or small subset of database of the user items to give prediction. The k Nearest Neighbors algorithm is the most extensively used algorithm for collaborative filtering. Dimensionality reduction techniques can also be used for memory based collaborative filtering [14]. A user-based nearest neighbour algorithm is a good example that describes memory based collaborative filtering approach [15].

Model based collaborative filtering techniques first develops a model based on dataset of user ratings and then provides recommendations. It can be assumed as a system that extracts information from a given dataset and then the information is used as a model to obtain recommendation without using the complete dataset again and again. This methods is beneficial in terms of both scalability and speed. It also improves prediction accuracy of algorithm. An example of model based approach is Kernel Mapping Recommender [16].

Advantages of Collaborative filtering is that it does not require any content information regarding user and item as a result it doesn't depend on a machine recognizable content. Therefore it is able to accurately recommend complex item without requiring to have "understanding" of that item. It is easier to implement recommendation systems using memory based collaborative filtering technique [17]. New data can be added easily and incrementally when using memory approach. Model based approach improves prediction performance.

Collaborative filtering has following three disadvantages. Cold Start: Collaborative filtering technique requires large data related to users to make accurate and efficient recommendations [18]. Scalability: Since enormous number of users and items are involved, a large computation ability is required to give recommendation [19]. Sparsity: The amount of products or items sold on ecommerce websites is enormous. A large number of active customers may only have given rating to a small size of the whole database. Thus, the very popular product may have very less and few rating [20].

B. Content-based filtering

A content based filtering system chooses items based on the correlation between the content or feature of the items and the user's preferences [21]. It is based on a profile or a description record containing preferences of a user and a description of the items under consideration [22]. Recommendations given by content based filtering is based on users past experiences. A profile of user's orientation and items description is required in content based filtering technique. By discovering different important features of an item, a weighted vector is created of that item. Based on it a content-based profile of different users and their preferences is created. The weights are used to denote how important a feature of an item is to the user. It can be computed from ratings and reviews from individual users by employing a variety of techniques. Simple technique like average values obtained from rating and reviews of an item; to sophisticated techniques like cluster analysis, Bayesian classifiers, decision tree and artificial networks can be used. Content-based filtering algorithms try to recommend items based on similarity count [23].

Advantage of this techniques is that Content-based recommender system provide user independence through exclusive ratings which are used by the active user to build their own profile. Content-based recommender system provide transparency to their active user by giving explanation how recommender system works. Content-based recommenders system are adequate to recommend items not yet placed by any user. This will be advantageous for new user.

Content-based filtering also have some disadvantages. Sometimes it is tough to generate attributes for items of certain range. It recommends the same types of items because of that it suffers from an over specialization problem. It is harder to get feedback from users because they do not rank most of the items therefore, it is not possible to determine whether the recommendation is correct.

C. Hybrid filtering

Hybrid approach is an approach in which collaborative filtering and content-based filtering are combined to improve accuracy and effectiveness of recommendations.

Hybrid filtering technique can be developed and used in following different ways: content based and collaborative based systems can be made as separate individual part and then they are combined as one single system; a single independent collaborative based system can be made and then some functionality of content based technique is added in it; similarly, a single independent content based system can be made and then some functionality of collaborative based technique is added in it; or one model can be created in which content based and collaborative based techniques are merged.

An example of hybrid recommendation system is Netflix [24]. It recommends by making comparison between the observations and exploring patterns of users having similar interests (collaborative filtering). It also provides or suggest movies having same features with the films that the same viewer has given high rating (content based filtering). Following Table II lists few algorithms that are used for collaborative and content-based filtering. Advantage of Hybrid filtering technique is that it solves the cold start and the sparsity problems in recommender systems.

TABLE II. FEW ALGORITHMS USED FOR COLLABORATIVE AND CONTENT-BASED FILTERING

Collaborative filtering	**Content based filtering**
K-nearest neighbor	Vector-space representation (TF–IDF representation)
Pearson correlation	Relevance Feedback
Mean-Squared Difference (MSD)	Rocchio's Algorithm
Vector cosine	Linear Classifiers
Matrix factorization	Probabilistic Methods
Bayesian classifiers	Naïve Bayes
Regression based methods	Cosine Similarity Function
Dimensionality reduction techniques	Decision Trees

III. PROPOSED FRAMEWORK

The proposed framework works by first dividing Salary into three main components i.e. Saving, Expenditure and Investment. The data of several persons will be analysed and filtered to get their preferences and choices. Filtering will find what method people have used for saving, where they have invested and how they spend their money for purchase and transaction purposes. After that recommendations will be made for them so that they can take correct decision. This will lead to better utilisation of money and profit maximisation. Saving component can be improved by suggesting people about how and where the money should be kept so that it can give higher outcome. Expenditure component will be suggested in such a way that it would lead to purchase of better and economical products than others. Investment component can be made profitable by suggesting people about safe and popular schemes. Expenditures, Savings and Investments dataset will be analysed using item-based collaborative filtering techniques with the help of Apache Mahout and Hadoop. Different independent datasets related to expenditures, savings and investments of people from various areas, community and fields can be used.

Item-to-item collaborative filtering finds matching product or item that was purchased by a user earlier. The product or item with similar feature as that of earlier rated or purchased product by the same user is obtained. Those similar products or items are then kept in a list which would be suggested or

recommended to that user in future. Similarity between the products or items can be computed using various ways such ratings given by the user and description of the product or item. To obtain better similar pairs of products or items best approach is to first find all those products which all the customer bought. Then find all other items bought by those customers. Now finally perform the similarity computation for only those set of products or items. Algorithm like Cosine-based similarity, Pearson correlation and Matrix factorization will be used to obtain similarity and prediction. The advantage of selecting item-to-item collaborative filtering is that it provides better predictions in comparison to user-based method. Figure 1 shows the framework for proposed Real time financial analysis using big data technologies.

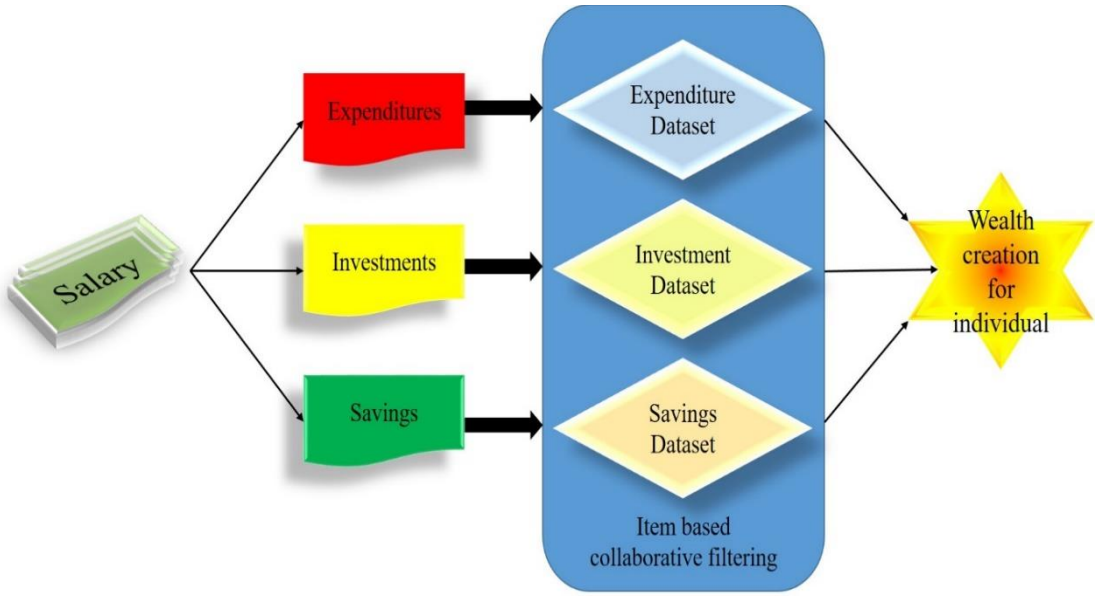

Fig. 1. Proposed framework

IV. IMPLEMENTATION

The proposed system for Real time financial analysis using big data technologies will be implemented using Apache Hadoop Framework and Apache Mahout. The Item-based collaborative filtering techniques will be used for filtering and analysing the dataset. A snapshot of Hadoop running on single node system is shown in Figure 2.

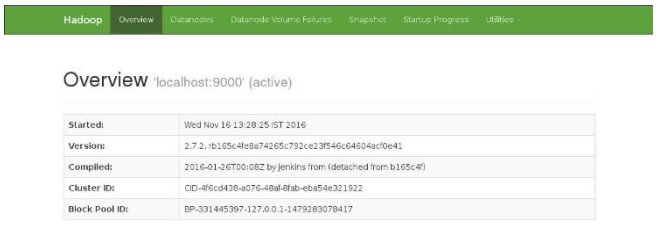

Fig. 2. Hadoop overview screen

Apache Hadoop is a great significant tool for Big Data Analytics. Hadoop is a framework for storage of large amount of data or information on clusters of commodity hardware and processing of that data. A cluster is a group of interconnected computer systems (known as nodes) that can work on a common analytic problem [25]. The modules of Hadoop framework [26] is designed with an assumption that any failure in hardware will be automatically handled by the framework. Hadoop consists of two main components: Hadoop Distributed File System (HDFS) and MapReduce. Hadoop contains two main components: storage and processing. The Hadoop Distributed File System (HDFS) is the storage component. HDFS is a distributed, scalable, fault tolerant and portable file system written in Java for storing and managing huge amount of data. Hadoop creates multiple replicas of the work and distribute them among nodes (machine) in the clusters and HDFS stores the data that may be used for processing. It enables reliable and rapid access of the data. MapReduce [27] is a framework for performing distributed data processing using the MapReduce programming model. It is the processing component of Hadoop. MapReduce consists of two basic steps: a map step and a reduce step. Map step takes input tasks and splits them into smaller sub-tasks. It then perform some required operations on each sub-task and gives some intermediate outputs. Reduce step combines the intermediate outputs from the map step and gives final output.

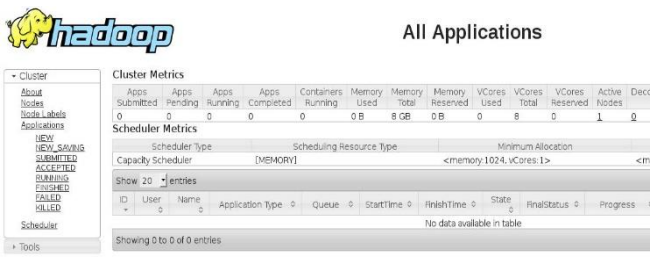

Fig. 3. Hadoop interface screen

Figure 3 provides a snapshot of Hadoop interface which will be used for implementation of the proposed system. Apache Mahout provides distributed, scalable machine learning algorithms mainly used for Collaborative filtering. Apache Mahout's core algorithms are implemented on Apache Hadoop using the MapReduce paradigms. List of Apache Mahout supported algorithms includes: Classification, Clustering, Pattern Mining, Regression, Dimension reduction, Pearson Correlation, Similarity Vectors, Euclidean Distance, Similarity Measures, Spearman Correlation, Tanimoto Coefficient, Log Likelihood Similarity, etc.

Item-based collaborative filtering is one of the most favourable technique used for developing recommender systems. In Item-based collaborative filtering approach only those items will be considered which are most similar to the given item for which rating is to be predicted. Item similarity weights will be used to obtain K most similar items and then unknown rating is predicted. The top N items which have the highest predicted rating will be recommended to the user. The steps involved in predicting recommendations are as follows:

A. Similarity weight computation

The similarity weight is an important parameter in the collaborative item based filtering approach [28]. The most popular similarity measure is Pearson correlation coefficient shown in Equation (1) and it will be used for this work. It is defined as:

$$sim(i,j) = \frac{\sum_{u \in U_{ij}} (R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \in U_{ij}} (R_{u,i} - \overline{R_i})^2} \sqrt{\sum_{u \in U_{ij}} (R_{u,j} - \overline{R_j})^2}} \quad (1)$$

where $sim(i,j)$ is the similarity between item i and item j; U_{ij} is the common user set, who rated on both item i and item j; $\bar{R}_i$, $\bar{R}_j$; are the average ratings of item i and item j respectively; $R_{u,i}$, $R_{u,j}$ are the rating of user u on item i and item j respectively.

B. Selection of K most similar neighbors

In this collaborative filtering technique, the quality of recommendations depends on the number of neighbors or similar items which were used to obtain the unknown prediction or rating for an item into consideration. Hence neighbors selection must be done more carefully to generate quality recommendations [29]. Hence K most similar neighbors will be selected which have highest similarity than others.

C. Recommending the top N items

In this step the unknown rating or preferences for the items are predicted which user has not rated in past. Out of those items whose unknown ratings and preferences were predicted, N items with highest predicted value are recommended to the user. The value of N should be selected with caution so that user gets better quality recommendations. Prediction function shown in Equation (2) for a given user u and item i is defined as:

$$P_{u,i} = \overline{R_j} + \frac{\sum_{j \in kNN_i} sim(i,j) \times (R_{u,j} - \overline{R_j})}{\sum_{j \in kNN_i} (|sim(i,j)|)} \quad (2)$$

where $P_{u,i}$ represents the predication for user u on item i; $\bar{R}_i$ is the average ratings on item i; kNN_i is the nearest neighbor item set of item i; $sim(i,j)$ is the similarity between item i and its neighbor j; $R_{u,j}$ is the user u rating on item i; $\bar{R}_j$ is average ratings on item j.

V. CONCLUSION

There is no way of getting correct suggestion and direction about using the salary or money for saving, investing and expending. People are not so much aware of several alternative schemes, offers and product that are more profitable, economical, safe and reliable. Many of them follow their friends, relatives and neighbours suggestions. But the decisions which was right for one can be wrong for another because there are several factors which differentiates every person. The proposed system for Real time financial analysis using big data technologies will aim to provide correct and efficient suggestions to a person which can lead to profit and wealth management. Although this proposed system is beneficial for many people, it also have some shortcomings. It is more useful when persons are having high salary which is to be managed properly. Low salary people doesn't need such recommendation systems as there is not much money to plan or manage properly. Another disadvantage is that in spite of having high salary, many people doesn't have much time to use such kind of recommended systems before using their money. In today's fast moving world one do not spend much time in thinking and deciding about expenditures, savings and investments. Instead they choose from one or two options whichever best suites them. In such scenarios, Recommender system play no role and remains unused.

REFERENCES

[1] Chandarana, P., Vijayalakshmi, "Big data analytics frameworks", 2014 IEEE International Conference On Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 430-434, 2014.

[2] Thomas Erl, W.K., Buhler, P., Big Data Fundamentals Concepts, Drivers and Techniques, 1st ed., pp. 29, USA: Prentice Hall, 2015.

[3] Anuradha, J., et al., "A brief introduction on big data 5vs characteristics and hadoop technology", Procedia computer science 48, pp. 319-324, 2015.

[4] Thorat, Poonam B., R. M. Goudar, and Sunita Barve. "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," International Journal of Computer Applications 110.4, 2015.

[5] Francesco Ricci and Lior Rokach and Bracha Shapira, "Introduction to Recommender Systems Handbook, Springer, pp. 135, 2011.

[6] F. Ricci, L. Rokach, B. Shapira, P.B., "Kantor Recommender Systems Handbook", 1st ed., XXX, pp. 20, 2011.

[7] J. B. Schafer, D. Frankowski, et al.,"Collaborative filtering recommender systems", The Adaptive Web, pp. 291–324, 2007.

[8] X. Luo, Y. Xia, Q. Zhu, "Applying the learning rate adaptation to the matrix factorization based collaborative filtering", Knowledge Based Systems 37, pp. 154–164, 2013.

[9] John S. Breese; David Heckerman & Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (UAI'98), 1998.

[10] G. Linden, B. Smith and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," in IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan/Feb 2003.

[11] I. Markovsky, "Low-Rank Approximation: Algorithms, Implementation, Applications," Springer, 2012.

[12] Takács, G.; Pilászy, I.; Németh, B.; Tikk, D., "Scalable Collaborative Filtering Approaches for Large Recommender Systems," Journal of Machine Learning Research, vol. 10, pp. 623–656.

[13] Rennie, J.; Srebro, N., et al., "Fast Maximum Margin Matrix Factorization for Collaborative Prediction," Proceedings of the 22nd Annual International Conference on Machine Learning, ACM Press, 2005.

[14] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Application of dimensionality reduction in recommender system – a case study," ACM WebKDD Workshop, 2000b, pp. 264–272.

[15] Breese, John S.; Heckerman, David; Kadie, Carl, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Microsoft Research, 1998.

[16] "Kernel-Mapping Recommender system algorithms," Information Sciences, pp. 81–104, 2015.

[17] Rubens, Neil, et al., "Active Learning in Recommender Systems," Recommender Systems Handbook. Springer, US, 2016.

[18] Elahi, Mehdi, et al., "A survey of active learning in collaborative filtering recommender systems," Computer Science Review, Elsevier, 2016.

[19] Sanghack Lee and Jihoon Yang and SungYong Park, "Discovery of Hidden Similarity on Collaborative Filtering to Overcome Sparsity Problem," Discovery Science, 2007.

[20] Zhou, Jia, and Tiejian Luo. "A novel approach to solve the sparsity problem in collaborative filtering." Networking, Sensing and Control (ICNSC), 2010 International Conference on. IEEE, 2010.

[21] Van Meteren, Robin, and Maarten Van Someren. "Using content-based filtering for recommendation." *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*. 2000.

[22] Peter, Brusilovsky, "The Adaptive Web," pp. 325, 2007.

[23] M. Balabanovic, Y. Shoham, "Content-based, collaborative recommendation", Communications of the ACM, pp. 66–72, 1997.

[24] Gomez-Uribe, Carlos A., and Neil Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, 2016.

[25] deRoos, D., Zikopoulos, P.C., Brown, B., Coss, R., Melnyk, R.B., "Hadoop For Dummies," pp. 13, John Wiley & Sons, Inc., 2014.

[26] Patel, A.B., Birla, M., Nair, U., "Addressing big data problem using hadoop and mapreduce," 2012 Nirma University International Conference on Engineering (NUiCONE), IEEE, pp. 1-5, 2012.

[27] Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, pp. 107-113, 2008.

[28] Alam, M.I., Pandey, M. and Rautaray, S.S., *"A Proposal of Resource Allocation Management for Cloud Computing"*, International Journal of Cloud Computing and Services Science (IJ-CLOSER), 3(2), pp.79-86, 2014.

[29] Dey, Monali, and Siddharth Swarup Rautaray, "Disease Predication of Cardio-Vascular Diseases, Diabetes and Malignancy in Lungs Based on Data Mining Classification Techniques.", International Journal of Computer Science International Journal of Computer Science and Engine and Engineering Open Access, 2014.