I.J. Information Engineering and Electronic Business, 2017, 5, 18-27
Published Online September 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2017.05.03

A Framework for Development of Recommender System for Financial Data Analysis

Pradeep Kumar M. Kanaujia, Manjusha Pandey and Siddharth Swarup Rautaray
School of Computer Engineering, KIIT University, Bhubaneswar - 751024, Orissa, India
Email: pkanaujia26@gmail.com, manjushafcs@kiit.ac.in, siddharthfcs@kiit.ac.in

Received: 10 April 2017; Accepted: 11 June 2017; Published: 08 September 2017

Abstract—The huge amount of data is being created every day by various organisations and users all over the world. Structured, semi-structured and unstructured data is being created at a very rapid speed 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. 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. The presented paper proposed a collaborative filtering based recommender system for financial analysis based on Saving, Expenditure and Investment using Apache Hadoop and Apache Mahout. 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. Due to enormous amount of data involved, Apache Hadoop framework is used for distributed processing. Collaborative filtering and Apache Mahout is used for analysing the data and implementation of the recommender system.

Index Terms—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. 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 [1][2]. The following Fig. 1 depicts the seven most important characteristics of Big Data that are: Volume, Velocity, Variety, Veracity, Validity, Volatility and Value [3][4].

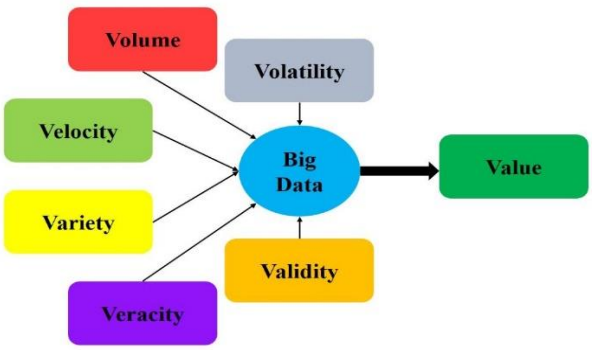

Fig.1. Seven characteristics of Big Data (Seven V's)

The huge amount of data that is being created every day by various organisations and users all over the world adds up to the Volume of Big Data. Velocity is the rate at which the data is being created and modified because of ever increasing usage of computing devices by mankind. The data changes rapidly over time and there may be chances that the data which is valid for a particular instant may become obsolete for another instant.

Variety is the type and format of the data i.e. structured which follows a specific format like online transactions, semi-structured that follow flexible format like emails and unstructured that do not follow any specified format like images and videos [5]. The following Fig. 2 represents the variety of data along with examples of those.

Veracity means the authenticity of data. It defines how much sure an organisation is with this data? Or tells whether the data is, the data that is required? The data should produce meaningful information when it is used for some problem area or research task. The correct and consistent data which can produce meaningful

information can only be used for further analysis thus making Veracity an important characteristics of Big Data Analytics. Validity of data is different from veracity. Validity means whether the data is accurate and correct for a particular problem usage. There may not be any veracity problem in the data but it may be invalid data for that particular instant or problem task. A dataset which is valid for one instant of time and problem may become invalid for another instant of time and another problem. Volatility of big data defines the nature of data that tends to change rapidly and in unpredictable way. Organisation can follow the retention policy for the data through which data can be easily removed once retention period expires.

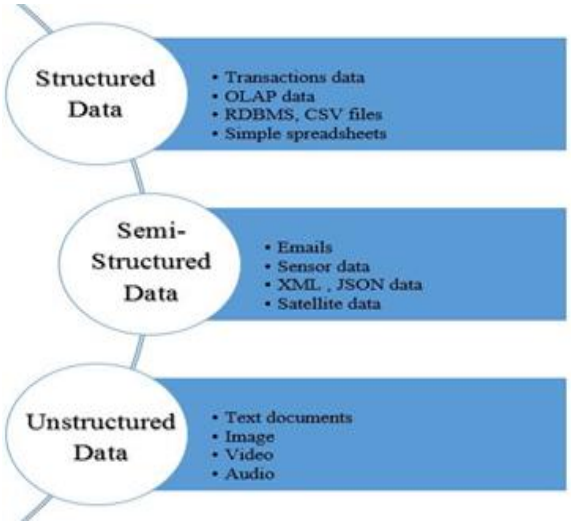

Fig. 2 Categories of data types with examples

Value is the usefulness of data based on the face value of data that has been generated by an organisation and would be used by them. This is a special V as it is the desired end result of data analysis process. Initial Data Analytics uses simple visualisation and statistical techniques like slicing and dicing of data [6] to generate useful pattern. Slicing and dicing are used to breakdown larger data into smaller subset that are easier to manage and explore for generation of analytical results. Advanced Analytics uses algorithms for complex analysis of structured, semi-structured and unstructured data. Big Data Analytics also includes advanced statistical methods, machine learning, neural networks, video analytics, text analytics and advanced data mining techniques [7].

The large amount of data generated from different users from heterogeneous sources in the form of reviews, ratings, feedbacks, shopping details, etc., 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 [8].

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 [9] 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 research paper is divided into six sections. The Section 1 gives introduction to the emerging research field of big data analytics with its one of the most important application known as recommender systems. Section 2 discusses the literature survey on recommender system. Section 3 briefs about the proposed framework for Recommender system for financial analysis using big data technologies. Section 4 is about the experimental setup for the implementation of the proposed system. Section 5 outlines the result analysis of the implemented system. Followed by the conclusion in section 6 on the proposed recommender system accumulating all the experiences while performing the implementation.

II. Literature Survey

Recommendation systems have proved to be very useful and their popularity is increasing day by day. Recommender systems are a kind of information processing applications that attempts to find unknown "rating" or "preference" for an item that a user may give. Those are 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 [9]. 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 such as 'where to invest money', 'how to plan savings', 'where to minimise and maximise expenditures', 'what product to buy', 'which place to visit'.

The recommender system aims to provide a list of personalized suggested items to its users. Rating or preference value is predicted for items which user has not observed or rated yet. Those items which contains highest predicted value is included in the recommendation list for the user. A vast class of performance measurement techniques are available that can be used for evaluation of the obtained recommendations. To give quality predictions for user's preferences and choices, a recommender system requires a good dataset. The input dataset plays an important role in determining the user's past behaviour and habits regarding the item into consideration. Table 1 shows example of some websites that use recommender systems.

Table 1. Some websites that use recommender system

Site	Recommended object
Amazon	Books/other products
Facebook	Friends
Netflix	DVDs
IMDb	Movies
LinkedIn	Jobs

Recommendation techniques can be classified into following three types: collaborative filtering, content-based filtering and hybrid approach.

A. Collaborative filtering

Collaborative filtering [10] plays significant role in recommendation process and that is why collaborative filtering is the most extensively used technique for designing recommendation systems. 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 analysing a large amount 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. Collaborative filtering is based on the assumption that people who liked or preferred an item in the past will like that item in the future too, and that they will like similar kinds of items also. A popular example of collaborative filtering is item based collaborative filtering i.e. a user who purchased a product 'x' will also purchase product 'y'. This technique is used by Amazon's recommender system [11].

Following Table 2 show an example of collaborative filtering using three schemes A, B and C of a bank. It also contains preferences of four users for these schemes. 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 Jean will invest in scheme C or not. Using collaborative filtering, one could be able to find that there is similar pattern in preferences of Roy and Jean. Roy gave high preference to scheme C so there is more possibility that Jean may also give high preference to scheme C. Hence, Jean may invest in scheme C.

Collaborative filtering techniques are classified as: memory based and model based. Memory based collaborative filtering techniques uses all or small subset of database of the user items to give prediction. The most extensively used algorithms for collaborative filtering is the k Nearest Neighbors (kNN).

Table 2. Example of a bank's three investment schemes and preferences by users

Schemes	Users			
	Roy	Jack	Tacy	Jean
A	5	4	-	5
B	2	-	4	1
C	5	1	2	?

Some collaborative filtering approaches use matrix factorization [12], a low-rank matrix approximation technique [13]. Dimensionality reduction techniques [14] can also be used for memory based collaborative filtering. An example of memory based approach is user-based nearest neighbour algorithm [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 approach is beneficial in terms of both speed and scalability. Model based approach [16] also improves prediction accuracy of algorithm. An example of model based approach is Kernel Mapping Recommender.

Advantages of collaborative filtering [17] is that it does not require any content information regarding user and item as a result it doesn't depend on 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. 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. Scalability: Since enormous number of users and items are involved, a large computation power is required to give recommendations. Sparsity: The number of products or items purchased from popular ecommerce websites is very large [18]. There may be many cases that the most active users might only have rated a small subset of the overall products dataset. Therefore, most popular product may have very few

ratings due to ignorance of user's to give rating and feedback.

B. Content-based filtering

A content based filtering [19] system selects items based on the relationship between the description of the items and the profile of the user's preferences. Recommendations given by content based filtering [20] is based on users past experiences. A profile of user's orientation and items description is required in content based filtering technique. A content based description or profile of users is created by the system by using a weighted vector list of item features. The value of weight represents the interest or likeness of the user towards each feature of the item. Weighted vector can be calculated from each item's rated content vectors using various techniques. Simple technique like average values of the rated item vector 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.

Advantage of this techniques is that content based recommender system is not user dependent for prediction instead it uses ratings on items given by the current. These exclusive ratings for different items and their features helps in creating user's profile. Another advantage is that by providing clear view of how recommendation is made, content based recommender system provides transparency to the users. Items that has not been purchased or used by any of the user can also be recommended using it. This will benefit a new user of the system. Content-based filtering also have some disadvantages. Sometimes it is tough to generate attributes for items of certain range. It many times leads to overspecialisation problem because same types of items are recommended repeatedly. Sometimes it is difficult for the system to give correct recommendation due to lack of ratings. The system does not have facility to get ratings for unrated items from the user due to which recommended item may be useless.

C. Hybrid filtering

Table 3. 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	Na ïve Bayes
Regression based methods	Cosine Similarity Function
Dimensionality reduction techniques	Decision Trees

Hybrid approach is an approach in which collaborative filtering and content-based filtering are combined to improve accuracy and effectiveness of recommendations. Following Table 3 lists few algorithms that are used for collaborative and content-based filtering.

Hybrid filtering technique can be implemented in following ways as shown in Fig. 3. By making collaborative based and content based predictions separately and then combining them; by adding content based capabilities to a collaborative based approach; by adding collaborative based approach into content based approach; or by merging the approaches into one model. An example of hybrid recommendation system is Netflix [21]. It recommends movies by making comparison between purchasing habit and interest of similar users (collaborative filtering) as well as it recommends movies having similar features to the movies that a user has previously liked or rated (content based filtering).

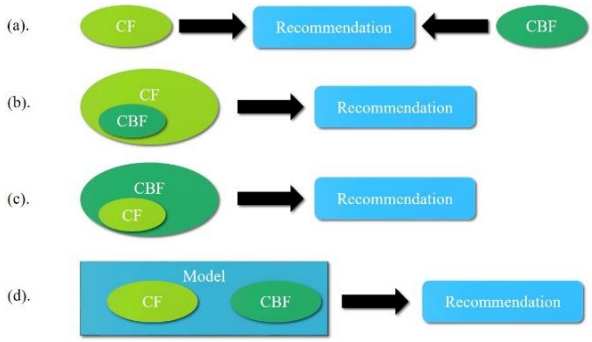

Fig.3. Different ways to implement Hybrid approach

Advantage of Hybrid filtering technique is that it solves the sparsity and the cold start problems in recommender systems.

Major challenges in recommender systems are [22] data sparsity, scalability, diversity and vulnerability to attacks. Due to large datasets, the user-item matrix used for filtering could be very large and sparse. It can degrade the performance of recommendation process. Data sparsity leads to the cold start problem. Such systems suffers from scalability problems as the number of users and items increases. Recommender system are expected to increase diversity because they help us to discover new product but some algorithms, may accidentally do the opposite. Most recommender systems are vulnerable to harmful attacks trying to promote or hamper some items.

Several related work have been done in past focusing on recommendation systems and filtering techniques. Adomavicius et al., [23] presents overview of recommendation systems and describes the three methods of recommendation systems which are collaborative, content based and hybrid recommendation approaches. It also outlines some of the limitations of existing recommendation methods and discusses some extensions that can be employed to develop more capable systems.

Sarwar et al., [24] analysed different item-based recommendation algorithms. It discussed different techniques to find item-item similarity and used it to obtain recommendations. The obtained result was compared with basic k-nearest neighbour and it was

concluded that item based techniques provide better performance and quality suggestions than user based techniques.

Ponnam et al., [25] presented a movie recommender system using item-based collaborative filtering. In this first User item rating matrix was examined and the relationships among various items were identified, and then these relationships were used in order to compute the recommendations for the user. The dataset that was utilized was the Netflix Data set which is available in the Group Lens which has collected and made accessible of the user item rating data sets from the Movie Lens web site. Adjusted cosine similarity method was used for calculating the similarities between the items. Then these similarity weights were used to calculate the predicted rating of the movies or items that are not rated by the user. Finally the top most N number of recommendations were the output to the users as recommendations.

Weijie et al., [26] presents a new improved collaborative filtering based on item similarity modified and item common ratings which take full advantage of the sectional data of item-user matrix information to modify the similarity calculation and rating prediction. Extensive experiments have been conducted on two different dataset to analyze the proposed approach. The result of the paper shows that the presented approach can improve the prediction accuracy of the item-based collaborative filtering not only on different neighbors, but also on different training ratio data set. The two data sets from the two different movie domains used were MovieLens rating data set and EachMovie data set. Pearson correlation based similarity method was used for similarity calculation.

Wei et al., [27] proposed a new collaborative filtering recommendation algorithm by using items categories similarity and interestingness measure to overcome the limitations of data sparsity and inaccurate similarity in personalized recommendation systems. Authors experimentally evaluated a top-N recommendation algorithm that uses items categories similarity and item-item interestingness to compute the recommendations. The data set used in this was also MovieLens data set.

III. Proposed Framework

The proposed framework is a collaborative filtering based Recommender System for financial data analysis based on Saving, Expenditure and Investment using Apache Hadoop and Apache Mahout.

Every person who is engaged in some work, employee of any organization or running any business get money in turn for their service or goods provided to that organization or client in the form of salary. Salary of persons varies according to the post and work in which they are engaged or employed in. For example an IT professional may get above Rs. 50,000; a person in government services of officer level may get above Rs. 35,000; a company owner may get profit in the form of money which can be above Rs. 1 Lakh and likewise others may get varying salary at the end of every month. People utilises their salary for different purposes according to their requirements. Person having low salary does not have much option on spending their salary as they spend the salary for the need and survival of their family. But the person who have better or higher salary need to look for better options and alternatives before spending their salary on something. For example a person wants to purchase a home theatre. Then he/she would like to compare different brands and different shop's offers and prices for home theatre. Another example is if a person want to invest their money in any bank. Then it would be easier for that person to select suitable scheme if he knows different investment schemes available in different banks. 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 and which will lead to better profit and wealth generation.

The proposed framework will implement a recommender system for financial data analysis on the salary or income or budget of a person. A person's money or finance will be represented as Salary and all the analysis will be done on it. It works by first dividing Salary into three main components i.e. Expenditure, Investment and Saving. 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. Fig. 4 shows the framework for development of recommender system for financial data analysis.

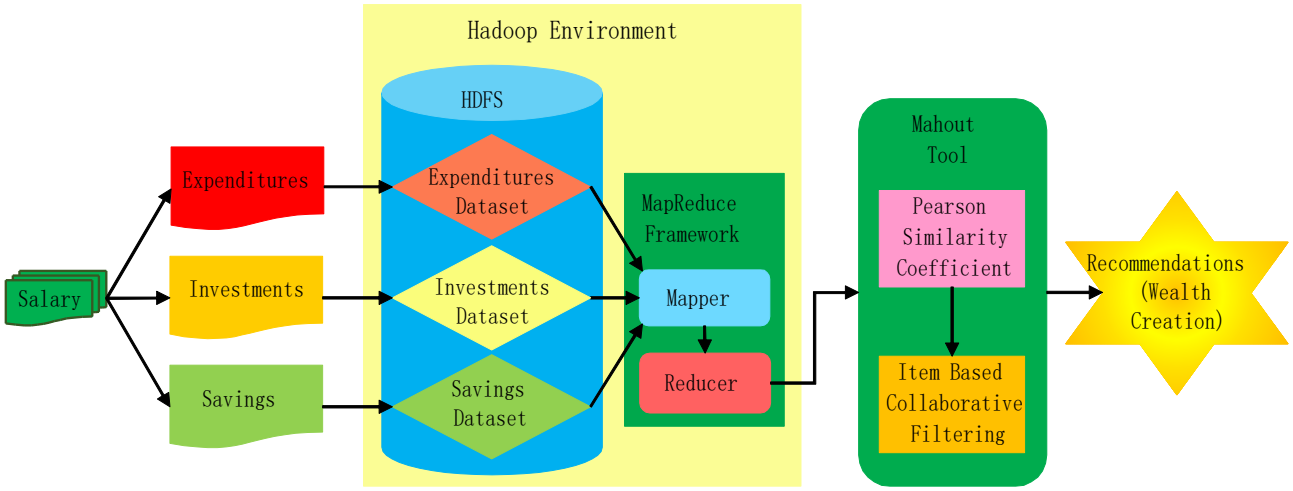

Fig.4 Proposed framework for recommender system for financial data analysis

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.

The proposed system for Recommendation system for financial analytics will be implemented using Apache Hadoop Framework [28] and Apache Mahout. The Item-based collaborative filtering technique [29] will be used for filtering and analysing the dataset. The most significant advantage of this system is a user can get recommendation related to three uses of money i.e. Expenditure, Saving and Investment. It provides fast and easy way to get recommendations. Another advantage of this system is the capability to analyse large dataset of users and their preferences quickly. Even though the system has number of benefits, there are few disadvantages such as the system is beneficial for users having good range of salary. Another demerit is that a user must have rated at least two items to get recommendation. There are few constraints that must be taken into consideration for the system. First most important constraint is that only salary of a user will be considered as input for financial analysis. It requires three different dataset for three components i.e. Expenditure, Savings and Investment. Dataset must be in .csv format and commas must be equally balanced. Only three fields are required i.e. Users, Items and Rating or Preference value in the dataset file.

IV. Experimental Setup

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. Table 4 shows the software and hardware specification for the implementation of the proposed framework.

Table 4. Minimum system requirements for proposed framework

Software Requirements	▪ Java ▪ Apache Hadoop and MapReduce ▪ Apache Mahout ▪ Eclipse IDE ▪ Yum, Rpm ▪ SSH must be installed and SSHD must be running to use the Hadoop scripts
Operating Systems Requirements	▪ Ubuntu v14.x (64-bit) ▪ CentOS v7.x (64-bit) ▪ Windows 7/8/10 (64-bit)
Browser Requirements	▪ Firefox ▪ Google Chrome ▪ Internet Explorer 9.0 or higher
Hardware Requirements	▪ CPU: Quad-/Hex-/Octo-core CPUs, running at least 2-2.5GHz ▪ RAM: 8GB or more ▪ Hard Disk: 20GB or more ▪ Network: 1GB Ethernet ▪ OS: 64-bit

The most important and necessary requirement for running the proposed recommender system is Java and Hadoop environment setup. Latest version of Java must be installed in the system for running Hadoop framework. Hadoop and MapReduce environment must be configured and running correctly. Apache Mahout Libraries are the core requirements for implementing prediction and recommendation steps. It can be downloaded and installed form Apache website. Eclipse is an integrated development environment for developing Java applications. It will be used to create MapReduce jobs in .jar file. All the software requirements listed in above table must be installed properly. Any stable version of operating system can be used for implementing the proposed system like Ubuntu, CentOS, Windows, etc. Latest version of web browser like Firefox, Chrome, Internet Explorer, etc. is required. There are not much hardware requirements for the system. Few electronics hardware will be used such as fast and efficient processor with latest features, sufficient storage device, high speed internet connection and 64 bit architecture computer system.

Apache Hadoop [30] is a 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. The modules of Hadoop framework 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. A snapshot of Hadoop running on single node system is shown in Fig. 5.

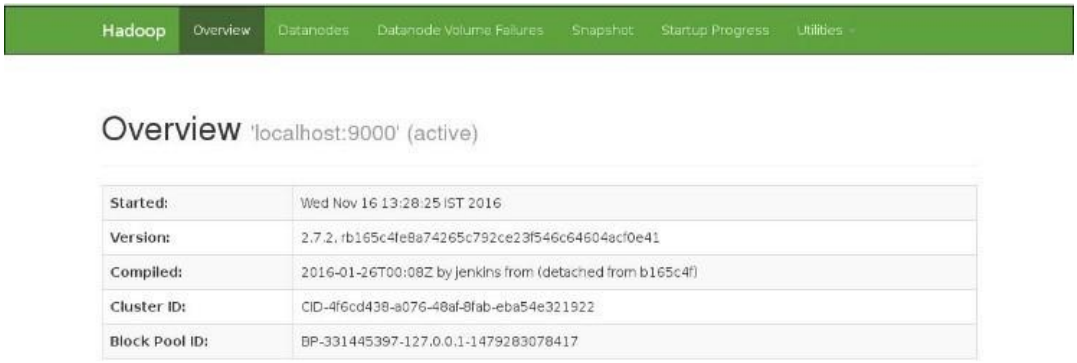

Fig.5 Hadoop overview screen

MapReduce [31] 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. 6 provides a snapshot of Hadoop [32] interface which will be used for implementation of the proposed system.

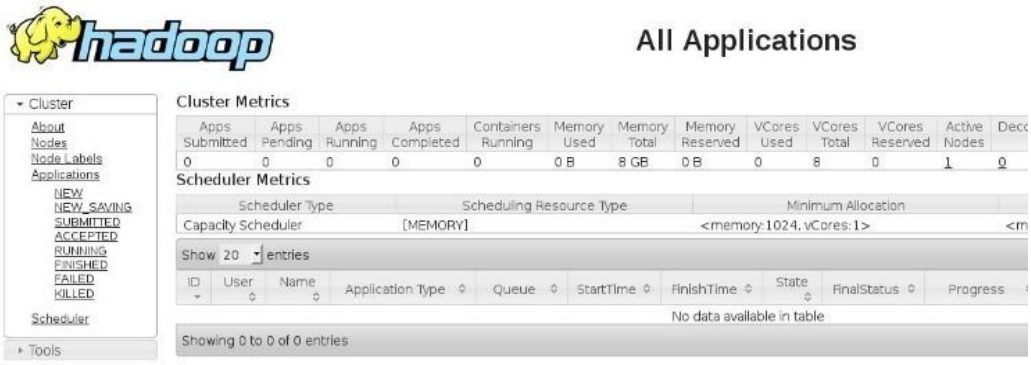

Fig.6. Hadoop interface screen

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 [34], Similarity Vectors, Euclidean Distance, Similarity Measures, Spearman Correlation, Tanimoto Coefficient, Log Likelihood Similarity, etc.

V. Result Analysis

The proposed system a framework for development of recommender system for financial data analysis 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.

Different datasets related to three components i.e. Expenditure, Investment and Salary will be used. Amazon dataset [33] have been used for Expenditure component. Amazon dataset contains information related to different product categories like Movies and CDs; Books; Electronics; Clothing, Shoes and Jewellery; Home and Kitchen; Sports and Outdoors; Cell Phones and Accessories; Health and Personal Care; Office Products; etc. Dataset for Investment and Saving component has been generated to be used as test dataset.

Item-based collaborative filtering is one of the most favourable technique used for developing recommender systems [34]. 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 system works by first getting salary range from user. A list of items whose price is under the income range specified by the user will be generated. Then similarity between only those items which fall under the given price range will be calculated and recommendation will be made from the obtained list of items to 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. 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. 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 [35] 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; $\overline{R_i}$ 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; $\overline{R_j}$ is average ratings on item j.

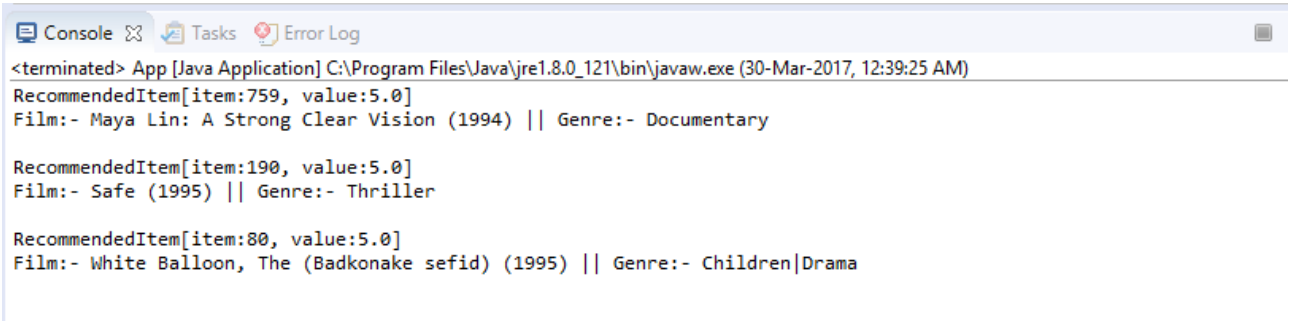

Fig.7.Recommender system output for movie test dataset

The Fig. 7 shows the output of recommender system for expenditure component on the basis of a user's preferences for movies. Three movies has been recommended to the user based on similarity value calculation using Pearson's correlation coefficient between movies rated earlier by the user. movies_rating dataset from Amazon data repository is used which contains different user ids, movies ids and ratings given by different users to the movies. The output screen show recommended items and value of rating for that item [36]. Name of movie as Film and its Genre is also displayed for better understanding to the user about the recommended item.

VI. Conclusion

After studying and comparing different techniques, Collaborative filtering based recommender system is being proposed for financial analysis using big data tools. This system can be used by organizations and enterprises for their employee's salary analysis or management. There is no way of getting correct suggestion and direction about using the salary or money for saving, investment and expenditure. 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 a framework for development of recommender system for financial data analysis 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 recommender 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] What is big data? In: Big Data Now: 2012 Edition, 1st edn., p. 3. O'Reilly Media, Inc, 1005 Gravenstein Highway North, Sebastopol, CA 95472 (2012)

[2] Gartner IT Glossary Big Data. http://www.gartner.com/it-glossary/big-data/

[3] M. A. u. d. Khan, M. F. Uddin and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the, Bridgeport, CT, 2014, pp. 1-5.

[4] van der Aalst, W.M.: Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In: Asia-Pacific Conference on Business Process Management, pp. 1-22 (2013). Springer

[5] Gray, J.: Data management: Past, present, and future. IEEE 13 (1996)

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

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

[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] 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.

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

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

[18] 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.

[19] 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.

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

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

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

[23] Adomavicius, G.; Tuzhilin, A., "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering, vol.17, no.6, pp.734-749, June 2005.
[24] Sarwar, B., George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms", In Proceedings of the 10th international conference on World Wide Web, ACM, pp. 285-295, 2001.
[25] L. T. Ponnam, S. Deepak Punyasamudram, S. N. Nallagulla and S. Yellamati, "Movie recommender system using item based collaborative filtering technique", 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, pp. 1-5, 2016.
[26] W. Weijie, Y. Jing and H. Liang, "An Improved Collaborative Filtering Based on Item Similarity Modified and Common Ratings", 2012 International Conference on Cyberworlds, Darmstadt, pp. 231-235, 2012.
[27] S. Wei, N. Ye, S. Zhang, X. Huang and J. Zhu, "Item-Based Collaborative Filtering Recommendation Algorithm Combining Item Category with Interestingness Measure", 2012 International Conference on Computer Science and Service System, Nanjing, pp. 2038-2041, 2012.
[28] deRoos, D., Zikopoulos, P.C., Brown, B., Coss, R., Melnyk, R.B., "Hadoop For Dummies," pp. 13, John Wiley & Sons, Inc., 2014.
[29] J. B. Schafer, D. Frankowski, et al.,"Collaborative filtering recommender systems", The Adaptive Web, pp. 291–324, 2007.
[30] 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.
[31] Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, pp. 107-113, 2008.
[32] McAuley, Julian, et al. "Image-based recommendations on styles and substitutes." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015.
[33] 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.
[34] 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.
[35] Zaied, Abdel Nasser H., Gawaher Soliman Hussein, and Mohamed M. Hassan. "The role of knowledge management in enhancing organizational performance." International Journal of Information Engineering and Electronic Business 4.5 (2012): 27.
[36] Olaiya, Folorunsho, and Adesesan Barnabas Adeyemo. "Application of data mining techniques in weather prediction and climate change studies." International Journal of Information Engineering and Electronic Business 4.1 (2012): 51.

Authors' Profiles

Pradeep Kumar M. Kanaujia has received B.Tech degree in Computer Science and Engineering from UPTU, Lucknow. He is currently an M.Tech, Computer Science and Engineering student in School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India. His areas of interests are Data Mining and Data Analytics.

Dr. Siddharth Swarup Rautaray has done Ph.D (Computer Science) and is an IEEE member. He is currently working as Professor at the School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India. He has more than a decade of teaching and research experience. Dr. Rautaray has published number of Research Papers in peer-reviewed International Journals and Conferences. His areas of interests are Image Processing, Data Analytic and Human Computer Interaction.

Dr. Manjusha Pandey has done Ph.D (Computer Science) and is an IEEE member. She is currently working as Professor at the School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India. She has more than a decade of teaching and research experience. Dr. Pandey has published number of Research Papers in peer-reviewed International Journals and Conferences. Her areas of interests are WSN and Data Analytics.

How to cite this paper: Pradeep Kumar M. Kanaujia, Manjusha Pandey, Siddharth Swarup Rautaray, "A Framework for Development of Recommender System for Financial Data Analysis", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.9, No.5, pp. 18-27, 2017. DOI: 10.5815/ijieeb.2017.05.03