International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Survey on Recommender System using Distributed Framework

Sonali B. Ghodake[1], R. S. Paswan[2]

[1, 2]Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: ***Recommender systems form a specific type of In-formation Filtering (IF) technique. It makes use of content based, collaborative filtering or hybrid approach. In this paper, all the approaches of recommender system is explained. Classification of data is vital part of recommender system. The Bayesian Classifier is one of the most successful machine learning algorithms in many classification domains. As scale of recommender system continues to expand, the number of items and users of recommender system is growing exponentially. The impact of this, the single node machine implementing these algorithms is time consuming and unable to meet the computing needs of large data sets. To improve the performance distributed processing of big data across multiple clusters of nodes is needed. Apache hadoop is a parallel distributed framework. Hadoop distributed file system allows distributed processing of big data across multiple clusters of nodes. The recommender input data will be encapsulated in Apache Mahout and enhance efficiency of recommendation using parallel distributed framework.***

Keywords: Recommendation, Collaborative filtering, Pearson correlation, Apache Mahout, Hadoop.

1. Introduction

Recommender systems are a subclass of information filtering system that predict the 'rating' or 'preference' that a user would give to an item. The concept of recommender systems was introduced to deal with the challenges of information overload, to scan through the large information sets, and to retrieve the most relevant information [1]. For example we can recommend any product, locations, and services, for example books, videos, music, TV programs, documents, research resources, and website. Through personalized recommendations, Recommender system suggested about the items or services to users on the basis of purchases of similar products or services by the other customers. As information overload has created potential problem, information filtering technique is required i.e Recommender system. Information need to be prioritized for user rather than just filtering the right information, to achieve these we need to classify data accurately[2].There are several classification techniques for classification of both multivariate and univariate dataset, but some of the basic techniques are Bayesian classifiers, Decision tree classifier, Support vector machine (SVM), K-Nearest-Neighbor classifier(KNN).

To handle large amount of data (also called Big Data), we require a robust system. Now the question left over: How do we analyze this huge amount of data? The most suitable solution to this is: Distributed framework. Hadoop is an open source framework. It develops and executes distributed applications that process very huge amount of data. Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques to enable the capture, storage, distribution, management, and analysis of the information [5].

The 3 Vs of Big Data management:

- **Volume:** There is more data than ever before, its size continues increasing.
- **Variety:** There are many different types of data, as text, sensor data, audio, video, graph, and many more.
- **Velocity:** Data is coming continuously as streams of data, and we are interested in obtaining useful information from it in real time [5].

Hadoop come up with its own file system called HDFS (Hadoop Distributed File System). HDFS utilizes master slave architecture. HDFS logically separates the file system metadata and application data. The metadata is stored on a delicate computer called as NameNode (known as master node in GFS) in HDFS [2]. Application data was stored on other computers named as DataNodes (known as slave node in GFS). A data file is subdivided into one or more blocks and these divided blocks are replicated and stored across several DataNodes. All of the nodes contained with a Hadoop cluster are fully connected. The NameNode retains the file system namespace and the mapping of file blocks to DataNodes. When an HDFS client read a file, it first contacts the NameNode to get the locations of data blocks comprising the file and then reads or access these data blocks from closest DataNode. Applying the concept of map reduce where mapper class work as practitioner and reducer class work as combiner [4].

2. Related Work

Prem Melville and Vikas Sindhwani [1] in their research paper titled as"Recommender Systems", present the concept of recommender system and its different approaches. Pros and Cons of different approaches of recommender system.

Satya Ranjan Dash, Satchidananda Dehuri [2] in their research paper titled as, "Comparative Study of Different Classification Techniques for Post Operative Patient Dataset", present the different classification techniques and comparison of these techniques.

KunhuiLin, Jingjin Wang and Meihong Wang [4] in their research paper titled as"A Hybrid Recommendation Algorithm Based on Hadoop", present the distributed collaborative filtering recommendation algorithm. They use combination of k-means and slope one algorithm on hadoop.

Nitesh V. Chawla and Darcy A. Davis[5] in their research paper titled as "Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework ", present the foundations of work that takes a Big Data driven approach towards personalized health care and demonstrate its applicability to patient centered outcomes and meaning full use. They use collaborative filtering method for future recommendation.

Xingyuan Li.[8] in their research paper titled as "Collaborative Filtering Recommendation Algorithm Based on Cluster ", proposed an improved collaborative filtering approach Cluster based collaborative filtering recommendation algorithms which improve scalability of Collaborative filtering algorithms and reduce data sets sparse of the recommended system.

Kala Karun. A, Chitharanjan K. [9] in their research paper titled as " A Review on Hadoop HDFS Infrastructure Extensions", reviews some of the major enhancements suggested to Hadoop especially in data storage, processing and placement.

Marios et al [10] in their research paper titled as"A Collaborative Recommender System Based on Space-Time Similarities", proposed a collaboration based recommender in an Internet of things environment. Their approach relies on user to object space-time interaction patterns.

3. Recommendation System

It was defined as "recommender systems form a specific type of information filtering (IF) technique that attempts to present information items (e.g. movies, music, books, news, images, web pages, etc.) that are likely of interest to the user".

To accomplish the task of recommendations, the recommender systems usually employ any of the following recommendation approaches:

A. Content-based Filtering
B. Collaborative Filtering
C. Hybrid Recommendation System

A. Content Based Filtering

Content based filtering approach provides recommendations based on the content items that were targeted by the users in the past searches. By comparing various candidate items with the items rated in past by different users, the best matching items are recommended [1].Fig 1 presents an illustrative example of content based filtering.

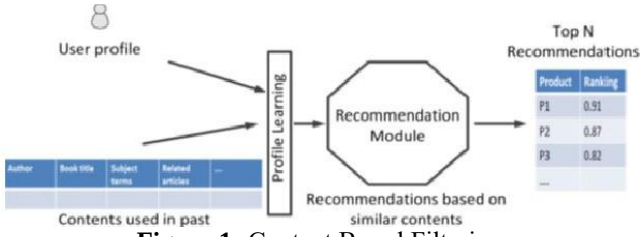

Figure 1: Content Based Filtering

B. Collaborative Filtering

Collaborative filtering methods are based on collecting and analyzing a large amount of information on user's behaviors, activities or preferences and predicting what users will like based on their similarity to other users. The input to collaborative filtering consists of user, item and rating to build recommendations using any of the following ways:

User Based Recommendations: User based recommendations are computed based on users with identical characteristics.

Item Based Recommendations:

Item based recommendations are computed based on similar items.

Slope-one: In this recommendation system, similarity metric is not considered as standard component. It is fast and simple approach for item recommendation.

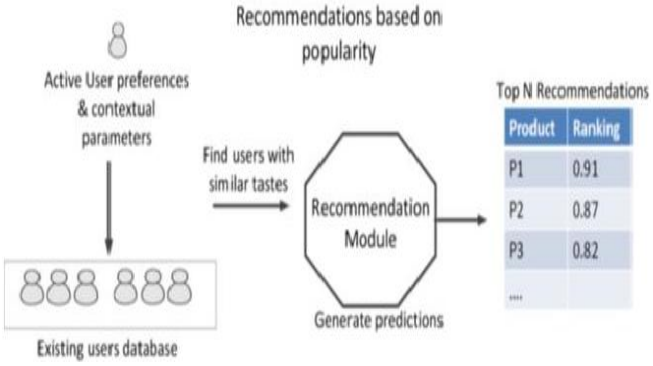

Figure 2: Collaborative Filtering

C. Hybrid Recommendation System

Recommender systems that use a combination of two or more filtering techniques are called hybrid systems. Hybrid system employs the content based and collaborative filtering. Such systems are claimed to have improved recommendation accuracy by overcoming the drawbacks of individual approaches as explained below.

Sparsity: The sparsity issue in recommender systems occurs due to the scarce data points required to de-scribe the exact context. Typically the number of items rated by each customer may be limited that eventually results in a smaller set of ratings by overall users [2]. The data sparseness issues arise

irrespective of the types of the recommender systems.

Cold start: Cold start is another challenge that recommender systems come across and it occurs for users or items that are new to the system and also because of the insufficient information. The variations of the cold start problem are new user problem, new item problem, and new system problem. The new users or new items do not have sufficient rating information in the system at the start. Consequently, for the new user or new items there may be insufficient records in the system to compare the similarity that eventually can result in zero similarity.

Scalability: Scalability refers to the capability to handle huge volumes of data in efficient and effective way.

4. Classification

The Classification has one of the major roles in Recommendation system. A classifier is a mapping between a feature space and a label space, where the features represent characteristics of the elements to classify and the labels represent the classes. A restaurant RS, for example, can be implemented by a classifier that classifies restaurants into one of two categories (good, bad) based on a number of features that describe it[2].

There are mainly two types of classifiers, supervised and unsupervised classification. In supervised classification, a set of labels or categories is known in advance and we have a set of labeled examples which constitute a training set. In unsupervised classification, the labels or categories are unknown in advance and the task is to suitably (according to some criteria) organize the elements at hand. One of the major goals of a classification algorithm is to maximize the predictive accuracy [2].

There are several classification techniques:

A. Decision Tree:
Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the items target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. It is one of the most successful techniques for supervised classification learning [6].A decision tree or a classification tree is a tree in which each internal (non leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes [2].

B. Support Vector Machine
Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM is used for both Linear and Non-linear data. It performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels.SVM finds this hyperplane using support vectors (essential training tuples) and margins (defined by the support vectors)[2].

C. K-Nearest-Neighbor Classifier (KNN):
K-Nearest Neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. It is a type of instance-based learning, or lazy learning. This algorithm is among the simplest of all machine learning algorithms [2].

D. Naive Bayes Classifier
The Bayesian Classifier is one of the most successful machine learning algorithms in many classification domains. Naive Bayes classifiers are simple probabilistic classifiers based on Bayes' theorem with strong independence assumptions between the features. It is a simple technique for constructing classifiers i. e models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is supervised learning method. The Naive bayes classifier is salable and usually provides better performance in terms of accuracy and coverage when applied to large databases [11].

Let D be a training set of tuples and their associated class labels. Each tuple is an n dimensional attribute vector.

Given features X=fX1, X2,, Xn. g
Predict a class label Y

$$P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)}$$

5. Generating Recommendations in Distributed Mode

To consider million preferences we require to implement recommender algorithm using distributed computing approach from Mahout, which is based on map reduce paradigm and Apache Hadoop [3]. It is also efficient in computing the recommendations in real time. It utilizes open source platform of Apache Mahout with Hadoop. This implementation is platform independent and performs distributed map reduce computation. This implementation framework indicates that use of Apache Hadoop with Mahout is most suitable for implementation of large scale and distributed generic recommender systems [3].

A. Apache Mahout

Apache Mahout is an open source project that is basically used for creating salable machine learning algorithms. It implements different popular machine learning techniques such as:

Recommendation
Classification
Clustering

Apache Mahout is used to incorporate its extensible collective library that enables us to implement recommendation algorithms. It Supports Distributed Naive Bayes and Complementary Naive Bayes classification implementations [3]. It provides a collaborative framework to generate recommendations. It requires both JDK and Maven. It enables to start exploring machine learning algorithms easily. The two main strengths of mahout are fast prototyping and evaluation. Mahout implementation framework is based on Java. Therefore the implemented code can be executed on any platform that uses Java virtual machine. Mahout provides a job to enable item and user based recommender in distributed mode. It is required to input onto Hadoop distributed file system i.e. HDFS to make it available to Hadoop because if the data is readily available on the local file system, it needs to be copied again into HDFS. Because Hadoop is a software [8] which runs across many nodes, so any data it uses must be available not on single machine but on multiple machines in the cluster. HDFS is an entity that can make data available to many nodes [3].

If we perform implementation of recommender using Ma-hout in distributed mode there are collections of mapper and reducer processes each with some intermediate result. The implementation process starts with the computation of co-occurrence matrix and user vectors. The difference in the implementations of recommender depends upon how they analyze the huge input data to identify the similarity between users and items that indicates the relevant preferences for that user [1]. It encapsulates recommender input data in Mahout and stores the resulting preferences. Data model implementation provides access to data required by recommender algorithms. There are many implementations to define similarity within Mahout [1].The recommender system framework can use any of the similarity metric such as Pearson correlation, Spearman correlation, log likelihood, Tanimoto coefficient etc.

6. Conclusion

Data in the form of reviews, opinions, feedback, remarks, and complaint treated as Big Data cannot be used directly for recommendation system. These data first filter/transform as per requirement. In the paper we discussed filtering techniques and issues related for handling data. We have discussed recommendation system on distributed i.e. hadoop framework.

References

[1] Prem Melville and Vikas Sindhwani, “Recommender Systems”, IBM T.J. Watson Research Center.
[2] Satya Ranjan Dash, Satchidananda Dehuri2, “Comparative Study of Different Classification Techniques for Post Operative Patient Dataset”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 5, July 2013.
[3] Lavannya Bhatia, S. S. Prasad, “Building a Distributed Generic Recommender Using Scalable Data Mining Library”, IEEE International Conference on Computational Intelligence Communication Technology, (2015).
[4] Kunhui Lin, Jingjin Wang, Meihong Wang, “A Hybrid Recommendation Algorithm Based on Hadoop”, The 9th International Conference on Computer Science Education (ICCSE 2014) August 22-24, 2014. Vancouver, Canada.
[5] Nitesh V. Chawla, PhD1 and Darcy A. Davis, “Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework”, June 25, 2013.
[6] Zhangguang Qian, Liuji Qing, Zhangrui Xue A, “Collaborative Filtering Recommendation Algorithm based on Correlation and Improved Weighted Prediction”, 2011 IEEE.
[7] Hao Ma, Irwin King and Michael R. Lyu, “Effective Missing Data Prediction for Collaborative Filtering”, ACM, July 2327, 2007,.
[8] Xingyuan Li, “Collaborative Filtering Recommendation Algorithm Based on Cluster”, 2011 International Conference on Computer Science and Network Technology.
[9] Xingyuan Li, “A Review on Hadoop HDFS Infrastructure Extensions”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
[10] Mario Muoz-Organero, Gustavo A. Ramrez-Gonzlez, Pedro J. Muoz-Merino and Carlos Delgado Kloos, “A Collaborative Recommender System Based on Space-Time Similarities”, IEEE CS 2010.
[11] Bankim Patel, Atul Patel, Atul Patel, “Big Data Analysis: Recommendation System with Hadoop Framework”, 2015 IEEE International Conference on Computational Intelligence Communication Technology.
[12] Mustansar Ali Ghazanfar and Adam Prugel-Bennett, “An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering”, IEEE 2013.
[13] Martin Wiesner and Daniel Pfeifer, “Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges”, Int. J. Environ. Res. Public Health, 2014.