Big Data Analysis: Recommendation System with Hadoop Framework

Jai Prakash Verma
Assistant Professor
Intitute of Technology,
Nirma University, Ahmedabad
jaiprakash.verma@nirmauni.ac.in

Bankim Patel, Ph D
Professor & Director
SRIMCA,
UKA Trasadia University, Bardoli
bankim_patel@srimca.edu.in

Atul Patel, Ph D
Professor & Dean
CMPICA,
CHARUSAT University, Changa
atulpatel.mca@acchanga.ac.in

ABSTRACT

Recommendation system provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items. In this paper, we have proposed a recommendation system for the large amount of data available on the web in the form of ratings, reviews, opinions, complaints, remarks, feedback, and comments about any item (product, event, individual and services) using Hadoop Framework. We have implemented Mahout Interfaces for analyzing the data provided by review and rating site for movies.

Keywords

Big Data Analysis, Hadoop Framework, Recommendation System, Mahout.

I. INTRODUCTION

The large amount of data available on the web in the form of ratings, ranks, reviews, opinions, complain, remarks, feedback, and comments about any item (product, event, individual and services) can be used for making correct decision. Moreover lots of blog forums are available on the web where web users can give their opinion, reviews, and comments about the items [1, 2, 4]. The recommendation based on the ratings and summary of relevant text about the items can be used for decision making. The growth of e-commerce sites and online businesses are enhancing the requirements of a robust recommendation system.

As on 2012 in India, number of internet users reached 137 million [5]. Indian e-commerce market is growing, and has reached to $14 billion in 2012 from $6.3 billion in 2011 and it is likely to reach $38 million by 2015 [4, 5]. Now a days in India, almost all types of product can be bought online. As well as here all three types of business models like B2B, B2C, and C2C are available in e-commerce. These types of online sites provide platform to both businesses and customers for making transactions. Explosive use of internet and e-commerce applications generates huge amount of data. Text summarization has been studied extensively in NLP [2, 4, 5]. We require a robust system for summarizing these texts as well as ratings. Now a days many other communication portals like Facebook, WhatsApp, twitter, etc. are available where user can ask their friends and followers about any product, service, event, and issues before making any decision [4, 6, 7]. To handle these huge amount of data (also called Big Data), we require a robust system. Now the question remains: How do we analyze that amount of text data? The most popular answer to this is: Hadoop framework. Hadoop is an open-source framework for developing and executing distributed applications that process very large amount of data [14, 15, 16].

The term Big Data is defined by four dimensions represented by four V's (Volume, Variety, Velocity, and Veracity). Volume is represented by the amount of text data that we are using for summarization to generate recommendation. Variety represents different type of data extracted from different sources like blogs, Facebook, twitter as well as different review and opinion sites. Reviewer can write their reviews, remarks, feedback in any format like structure, semi-structured, and unstructured that should be handled by the system. Velocity represents the seed of data generation on internet. Now a days everybody is connected through internet as well increasing the popularity of ecommerce sites has become the main reason for increasing the speed of text data generation on the web. Veracity represents the trustworthiness of the data. Many times review, opinion, feedbacks are manipulated or sponsored by different stakeholders of online business for their personal interest [14, 15, 16].

Hadoop provides its own file system called HDFS (Hadoop Distributed File System). When we deploy our text data on Hadoop file system, Hadoop distributes all the data in different clusters and performs operations parallel. Hadoop also keeps multiple copies of data in case of hardware failure. HDFS represents data in the pair of {key, value} associated with specific key. Here we have data in the form of text then each word can identify with a key and its associated value represent its weightage or occurrence in the whole dataset. Applying the concept of map reduce where mapper class work as partitioner and reducer class work as combiner [14, 15, 16].

The rest of the paper is organized as follows. In the section II we discuss Recommendation system and its issues and challenges. In section III we discuss the opinion quintuple generation. Section IV and V, is discussed some related work in the same direction and proposed system. In section VI we discuss experimental study done for proposed recommendation system with Mahout running with Hadoop Framework.

II. RECOMMENDATION SYSTEM:

Recommendation system provides the facility to understand a person's taste and find new, desirable content for them

978-1-4799-6023-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CICT.2015.86

automatically. Although people's tastes vary, they do follow patterns. People tend to like things that are similar to other things they like as well as other similar behavioral person likes. Some times these types of patterns can be related with the relevancy of items. On the other hand, we could figure out what items are similar to what we already liked, again by looking to other's apparent preferences. In fact these are the two broadcast categories of recommender engine algorithms: user-based and item-based recommenders. These recommendations are based on two filtering techniques namely collaborative and content –based filtering techniques. Collaborative filtering produces recommendations based on, and only based on, knowledge of users relationship with items, product and services. These techniques require no knowledge of properties of items and characteristics. Content-based recommendations are based on attributes of items. Here suggestions are based on the content related to items and their aspects. One more approach has become more popular where both the filtering techniques can be applied on different levels of recommendation system, called hybrid filtering technique.

a. Collaborative filtering

Collaborative filtering approaches build a model from a user's past behavior (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users; then use that model to predict items (or ratings for items) that the user may be interested in. Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties. These approaches are often combined in Hybrid Recommender Systems [2, 6].

b. Content-based filtering

Content-based filtering is another approach for recommender system. These methods are based on description of the item and user preferences. The keywords are used to describe the item to indicate their characteristics that can be used for generating recommendations. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach has its roots in information retrieval and information filtering research [2, 6]. Here we have discussed a few issues that should be handled before summarization of these text data.

a) Regional and SMS type's language data:

Many time reviewer may use their regional or SMS type language to express their views and opinions. SMS type words: Meaning of these types of words are depend on the sound of them. So we can use word to sound convertor and then covert this sound to nearly sounded word in target language. Regional language words: First these type of words are converted in English. To handle regional words written in English, first we identify user profile, based on it, system can identify the language of the word. Based on the word sound and regional language dictionary we can get nearby English word.

b) Positive and negative reviews and opinions:

SentiWordNet is a lexical resource in which all the sentences can be classified into three categories like positive, negative, and its objective. SentiWordNet is freely available for research purposes. It generates three numerical score for each sentence Obj(s), Pos(s), Neg(s). Here Obj(s) represents the objective of sentence, Pos(s) represents positivity and Neg(s) represents negativity [7].

c) Review that may not express any meaning:

These review and opinion data are extracted from web that may be real time or offline. But there is a possibility that reviewer can write anything or irrelevant text that will not express any meaning. These types of reviews and text should be identified and removed from dataset. SentiWordNet can be used for the same to identify these types of words [7].

d) Sarcastic reviews:

Reviews like "what a great product, it will not work for two days" are called sarcastic reviews. These reviews and opinions should be handled before summarization of reviews because these reviews look like positive but actually they are negative. SentiWordNet can be used for the same to identify these types of words [7]

e) Internet slang words and emoticons:

Reviewer can use many internet slang words in their reviews, these words are basically used for expressing emotions. System should identify them and their emotion, then update rank or rating to the review accordingly.

f) Conditional sentences or opinions:

Reviewers may provide their review based on some condition that this product is good for these conditions but not good for other conditions. In these types of review, positivity and negativity should be decided based on the preferable conditions or circumstances.

g) Spam reviews:

Major problems with reviews and opinions provided by different sites are spam reviews. Many time manufacture or different types of business stakeholders sponsor these sites for manipulation of reviews and opinion. Now a days spamming opinion and review has become a business. Many research papers and review are available in the area of email, search engines, and blogs spamming [3, 8, 9].

c. Hybrid filtering

Hybrid filtering approach can be implemented in two ways. One is both content based and Collaborative filtering are applied separately and then combine the result as per need. Second, first we apply collaborative filtering and then apply content based filtering on the result [2, 6].

III. Opinion Quintuple Generation:

An opinion or review can be represented by a quadruple, (e, s, h, a, and t), where 'e' represents target entity (item) on which we are finding opinion or review, 's' is a string that is actual opinion or review about the target entity, 'h' represents opinion holder who reviewed or opinionated the target entity, 'a' represents aspect of sentiment (positive, negative and neutral) and 't' represents time on which opinion or review generated. This quadruple can be generated with following steps.

A. Entity Extraction:

Real Time Web data can be extracted form web as per requirement. These data should be categorized as per business need and specification. Here entity may be any product, service, and person on which we require recommendation or opinion. It is described with a pair, e: (T, W), where T is a hierarchy of parts, sub-parts, and so on, and W is a set of attributes of e. Each part or sub-part also has its own set of attributes [3].

B. String/ Opinion Extraction:

Review and opinion string, the actual text data that should be summarized for producing recommendations. In this step all the relevant text is extracted and categorized as per business need. Here we assign a unique id for each string/word for further reference.

C. Aspect Sentiment Classification:

In this step we determine whether the opinion sentiment is positive, negative, or neutral. This attribute helps to classify all the review in these category.

D. Opinion Holder Extraction:

Opinion and review holder information can be extracted this helps to assign the weightage to the review or opinion. We can assign rating to each reviewer as per his/ her past behavior, region, and religion etc. We require separate algorithm to assign rate to each reviewer.

E. Time Extraction and Standardization:

Time factor is also very important in review summarization. Time attribute represents the time on which review and opinion is generated. Based on the time period, review rank or weightage is assigned. Time attribute can also be used to identify spam reviews and opinions.

Following are the template for quadruple, (e, s, h, a, and t) (figure 1) in different item category. That can be used for storing data in ETL (Extraction, Transformation, and Loading) process of Recommendation System.

Product: {Digital Camera, Str_review, positive/ negative, Mr xyz, 11-05-2014}
Service: {Passport Apply, Str_review, positive/ negative, Mr xyz, 12-05-2011}
Person: {Shri Narendra Modi, Str_review, positive/ negative, Mr xyz, 26-05-2014}
Event: {Election Result 2014, str_review, positive/ negative, Mr xyz, 16-05-2014}
Opinion on any Issue: {IPL Match Fixing, Str_review, positive/ negative, Mr xyz, 14-06-2013}

Figure 1: List of Expected examples for quadruple, (e, s, h, a, and t)

IV. RELATED WORK:

Many papers are available in literature on Big Data Analysis for Recommendation System using different platforms.

Linyuan Lü,[1], reviewing recent developments in recommender systems and discuss the major issues. They have compared and evaluated available algorithms and examined their roles in the future development. In addition to algorithm, physical aspects are described to illustrate macroscopic behavior of recommender systems. Potential impacts and future directions are also discussed. They are emphasized that recommendation has great scientific depth and combines diverse research fields which makes it interesting for physicists as well as interdisciplinary researchers

Sandra Garcia Esparza [2], proposed a collaborative filtering approach for producing recommendation for different products using reviews and opinions available on twitter or other social communication sites. In this paper they have analyzed reviews provided by blipper (a review and opinion site) for four different products using collaborative filtering technique.

Derek Bridge [10], describe GhostWriter 2.0, which takes a case based approach for reusing reviews and opinion provided by different users/consumers on amazon online sell site. This paper also shows user trials for suggestions and recommendations that GhostWriter produces.

Hongyan Liu[11], proposed recommendation method that analyzes the difference between the ratings and opinions of the user to identify the user's preferences. This method considers explicit ratings and implicit opinions, an action that can address the problem of data sparseness. Based on these methods, they also conduct an empirical study of online restaurant customer reviews to create a restaurant recommendation system and demonstrate the effectiveness of the proposed methods.

Kushal Bafna [12], proposed feature based summarization of reviews generated by different users of ecommerce sites. This research work shows extraction of reviews and opinions from opinion rich web-sites and then detecting positivity and negativity using feature based summarization technique.

Li Chen [13], proposed a novel clustering method based on Latent Class Regression model (LCRM), which is essentially able to consider both the overall ratings and feature-level opinion values (as extracted from textual reviews) to identify reviewers' preference homogeneity. In the experiment, they tested the proposed recommender algorithm with two real-world datasets. More notably, they compared it with multiple related approaches, including the non-review based method and non-LCRM based variations

Klavdiya [17], study the implementation of a predictive model on cloud using Hadoop and MapReduce programming concept. This paper represents a cloud recommendation system using mahout machine learning algorithm.

Teng-Sheng [18], focused on clustering with the TF-IDF weighted mechanism of daily tweets and breaking news based on Apache Mahout as well as evaluating the effect of removing stop words from the dataset. This paper only considers English language tweets, news, and blogs.

Jai Prakash Verma [19], recommended summarization of reviews and feedback given by students or different stakeholders for an educational institution. In this paper a recommendation model based on text summarization is proposed using open source software and tools.

V. Proposed system:

In this paper, we propose a recommendation system for the large amount data available on the web in the form of ratings, reviews, opinions, complain, remarks, feedback, and comments about any item (product, event, individual and services) using Hadoop Framework. Here we recommended a hybrid filtering technique to filter different types of reviews, opinions, remarks, comments, complains etc. Because recommendations are based on ratings, ranks, content, reviewer's behavior, and timing of review generated by different reviewers. We study a recommendation based on numerical data like Ratings or ranks provided for different product or services. Recommendation by applying the weightage of summarized reviews and opinions on the rating of item are proposing as future work. Figure 2 depicts the steps require for recommendation system.

ETL
Feature Generation
Recommendation Algorithms/ModelGenration
Workflows & Scheduling
Support Testing of the model
Measure Performance

Figure 2: Recommendation System Processes.

Recommendation System Platform with Hadoop: Hadoop provide a framework for storage, management, and retrieval of the big amount of data known as Big Data. Figure-3 shows that many high-level languages, predictive analysis algorithms, and other tools for different task can be integrated with Hadoop framework. For the above mentioned processes for recommendation system we have configured Mahout and Hive with Hadoop framework. HDFS is a distributed file system designed to run on commodity hardware. It is highly fault-tolerant, designed to be deployed on low-cost hardware. It require a few POSIX for working with different environment systems. MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. Hive is a data warehousing solution built on top of Hadoop. It provides SQL-like query language named HiveQL. The Apache Mahout free machine learning library's goal is to build scalable machine learning tools and data mining framework for use on analyzing big data on a distributed manner

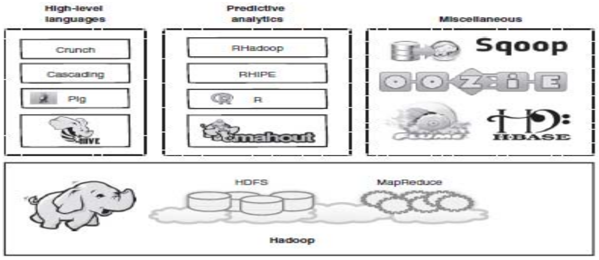

Figure 3: Hadoop Ecosystem from as per http://hadoop.apache.org/

Mahout Architecture and Algorithms: Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed, scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification [14]. As per mahout architecture (Figure-4), machine learning algorithms with different performance measurement methods are implemented in distributed computing environment. In this work these mahout interfaces are implemented with Hadoop framework. As well as the performance of the system are measured with different size of dataset. Figure-5 shows the list of algorithms, similarities, and neighborhood measures implemented with mahout on Hadoop framework.

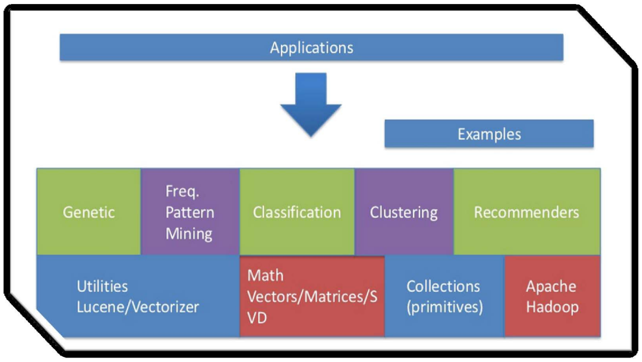

Figure 4: Mahout Architecture

Classification Clustering Recommender / Collaborative Filtering Evolutionary Algorithms Pattern Mining Regression Dimension reduction Similarity Vectors	Similarity Measures Pearson Correlation Spearman Correlation Euclidean Distance Tanimoto Coefficient Log Likelihood Similarity Neighborhood Measures Nearest N Users Algorithm

Figure 5: List of Mahout supported algorithm, Similarity and Neighborhood Measures

Evaluation for the different implementations is actually very time-consuming. The strength of Mahout lies in that it is possible to save time in the evaluation of the different combinations of the parameters. It provides standard interface for the evaluation of a Recommender System. Mahout provides classes for the evaluation of a recommender system. Figure 6 depicts two types of result evaluation techniques and their implementation parameters with mahout for recommendation system.

Prediction-based Measures	IR-based Measures
Class: AverageAbsoluteDifferenceEvaluator	Class: GenericRecommenderIRStatsEvaluator

Method: evaluate() Parameters: Recommender implementation DataModel implementation TrainingSet size (e.g. 70%) % of the data to use in the evaluation (smaller % for fast prototyping)	Method: evaluate() Parameters: Recommender implementation DataModel implementation Relevance Threshold (mean+standard deviation) % of the data to use in the evaluation (smaller % for fast prototyping)

Figure 6: Result Evaluation and Measuring Techniques

VI. Excremental Analysis

For experimental study we implemented recommendation system for movielans dataset, which is about the movie ratings given by different users with mahout on Hadoop framework. And analyzed with different size files. Following is the command line execution of item based recommendation algorithm on mahout with Hadoop framework. In table 1 we present CPU execution time with different size dataset.

$mahout recommenditembased -s SIMILARITY_LOGLIKELIHOOD -i /user/jaiprakash/dataset/dataset -o /user/jaiprakash/testdata/output --numRecommendations 25

Table 1: CPU execution time for different sized file.

File Name	Size	Time Taken
Dataset	288 bytes (0.000275MB)	Program took 168403 ms (Minutes: 2.8067166666666665)
r1.csv	24.6MB	Program took 631717 ms (Minutes: 10.528616666666666)
r2.csv	104.8MB	Program took 1408669 ms (Minutes: 23.477816666666666)
ratings.csv	129.4 MB	Program took 1684505 ms (Minutes: 28.075083333333332)

Following graph (figure-7) depicts that whenever file size is increasing the execution time is not increasing in the same ratio and we know that data size that are in the form of ratings, ranks, review, feedback are increasing drastically.

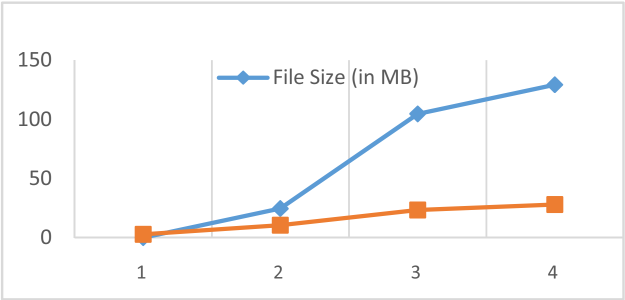

Figure 7: Execution time measurement with file size.

VII. Conclusion and future work

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 text data. We have implemented recommendation system for movielans dataset, on Hadoop framework and analyzed with different size files. Resultant graph is showing that whenever file size is increasing the execution time is not increasing in the same ratio and we know that data size that are in the form of ratings, ranks, review, feedback are increasing drastically. Here we are proposing Recommendation by applying the weightage of summarized reviews and opinions on the rating of item as future enhancement in this work.

References

[1] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang,Tao Zhoua, Recommender systems, Elsevier Journal - Physics Reports 519 (2012) 1–49

[2] Sandra Garcia Esparza, Michael P. O'Mahony, Barry Smyth, Mining the real-time web: A novel approach to product recommendation, Elsevier Journal - Knowledge-Based Systems 29 (2012) 3–11.

[3] Sentiment and Opinion Analysis By Bing Liu

[4] "Online Shopping touched new heights in India in 2012". Hindustan Times. 31 December 2012. Retrieved 31 December 2012.

[5] Website link http://en.wikipedia.org/wiki/E-commerce_in_India#cite_note-Online_shopping_touched_new_heights_in_India_in_2012-1

[6] Web Link http://en.wikipedia.org/wiki/Recommender_system

[7] Aurangzeb khan, Baharum Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs" 2011, IEEE

[8] Thiago S. Guzella*, Walmir M. Caminhas, "A review of machine learning approaches to Spam filtering", 2009, Elsevier Journal - Expert Systems with Applications 36 (2009) 10206–10222.

[9] Nikita Spirin, Jiawei Han, "Survey on Web Spam Detection: Principles and Algorithms", 2012, SIGKDD Explorations Volume 13, Issue 2, page 50-64.

[10] Derek Bridge, Paul Healy, "The GhostWriter-2.0 Case-Based Reasoning system for making content suggestions to the authors of product reviews", Elsevier Journal- Knowledge-Based Systems 29 (2012) 93–103

[11] Hongyan Liu, Jun He, Tingting Wang, Wenting Song, Xiaoyang Du, "Combining user preferences and user opinions for accurate recommendation", Elsevier Journal- Electronic Commerce Research and Applications 12 (2013) 14–23.

[12] Kushal Bafna, Durga Toshniwal, Feature Based Summarization of Customers' Reviews of Online Products, 2013, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

[13] Li Chen, Feng Wang, "Preference-based clustering reviews for augmenting e-commerce recommendation", Elsevier Journal - Knowledge-Based Systems 50 (2013) 44–59.

[14] Alex Holms, "Hadoop in Practice", 2012, Manning Publications co.

[15] Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman, "Mahout in Action", 2012, Manning Publications co.

[16] Pete Warden, "Big Data Glossary a guide to the new generations of data tools", 2011, O'Reilly.

[17] Klavdiya Hammond, Aparna S. Varde, "Cloud Based Predictive Analytics", 2013, IEEE 13'th International Conference on Data Mining Workshops.

[18] Teng-Sheng Moh, Surya Bhagvat, "Clustering of Technology Tweets and the Impact Stop Words on Clusters", 2012, ACMSE'12, March 29-31, 2012, Tuscaloosa, Alabama, USA, ACM, 978-1-4503-1203-5/12/03.

[19] Jai Prakash, Bankim Patel, Atul Patel,”Web Mining: Opinion and Feedback Analysis for Educational Institutions”, 2013, IJCA, Volume 84 – No 6, December 2013