Recommender System for E-Commerce Applications using Hadoop Framework

R Velumadhava Rao*, K Selvamani †

*Research Scholar, Anna University, India and velu_b4u@yahoo.com
†Assistant Professor, Anna University, India and smani@cs.annauniv.edu

Keywords: Maximum Hadoop, HDFS, Openstack, E-commerce, Recommender system.

Abstract

Nowadays, the usage of e-commerce applications increased rapidly with more number of users. The advancement in technologies like mobiles and other social media generated different variety and large quantity of data which is increasing a lot. The data collected and stored by social media is distributed and stored as an unstructured data in the server. The unstructured and an unformatted data is collected through the social media and other devices for ecommerce applications contains either useful information or irrelevant information. The information needs to be sorted out as the structured data and the relevant and useful information has to be retrieved and stored for further processing. The recommendation system developed for e-commerce application should improve the usage services with effective and convenient shopping experience by the users. In this paper we develop a Hadoop based e-commerce based recommender system for application of e-commerce with an implementation in Openstack.

1 Introduction

With the increase in the development of IT and internet services, electronic commerce is made easier and convenient for the customers for going with online shopping. It saves lot of time for the users and its cheap too for the users to shop items through e-commerce applications. Sometimes users face lot of issues in searching items and there is lot of confusion in buying the items. So it is necessary to recommend the user with the proper guidance in purchasing and to browse the items effectively. There are huge variety of e-commerce data which is displayed as ratings, reviews of the customer used the product, as well as feedback of users as well as the comments given by the customers [1, 2, 3]. Also there are lot of blogs and forums which are available for customers to provide with their feedbacks and ratings about the product used. The recommendation system will be based on these feedbacks, comments and user rating given for the particular item or product and this will be a decision maker. There is a huge requirement for the recommender system as there is an increase in more number of e-commerce sites and its usage. In the year 2012, the internet users who are in India alone are around 135 million [3]. The e-commerce business is growing in India rapidly with around $12 billion in the year 2012 which is double from 2011 which is around $6 billion. Now in the current recent years it almost touched around $30 billion [3]. In India everyone purchase items online includes electronic products, food items, other domestic products etc... The e-commerce can be divided into multiple business models such as B2C, B2B and C2C. These classification of models will provide the platform for both the customers who purchase items and for business peoples for transactions in every aspect. Nowadays, there are several communication media such as Whatsapp, Twitter, Facebook etc., sharing the information among the users about the product and there are followers based on the review comment [3, 5, 6]. There is a need for summarizing the text [2, 3] as there are huge amount of data is generated using the e-commerce applications and the internet technology. To handle this amount of data we use a Big Data technology and to process the data we use Hadoop. Apache Hadoop is an free and open source for implementing and processing the large amount of distributed data [7, 8, 9]. The terminology of Big Data is defined with three V’s (Volume, Variety and Velocity). Volume represents the quantity of data whereas Variety represents the several types of data which are extracted from different sources such as blogs, forums etc... The data can be of structured or unstructured as the system needs to handle this. Finally velocity represents the speed of data generation. There is another term which is represented as veracity which provides the trustiness of the data. Hadoop Framework is built up with its own distribute files system which is said to be a Hadoop Distributed File System (HDFS). When the data is transferred from local machine to hdfs, Hadoop distributes the data available in hdfs to different clusters and perform the required functions in parallel. Hadoop as default replicates with three copies of data for backup. Whenever the system failure or hardware failure occurs then the Hadoop uses the other data copy available within the system. Hadoop usually uses the (key, value) pair for processing. Hadoop uses the map reduce concepts for processing [7, 8, 9]. Openstack is open source software in cloud platform. Using openstack, the cloud users will deploy as IaaS. Openstack will help the users to run virtual machines and their instances which are used to handle several tasks in a cloud based environment. For example, in an e-commerce application the data are communicated and shared with the server on the remote site. This work can be processed by dividing it into multiple tasks such that the communication can happen with different users using several VM instances created. This will help in efficient communication with the individual users and the scaling process is also easier and quicker. In this research work, Hadoop based e-commerce
recommender Linux-based system for item/product is developed where the data are in the form of reviews, feedback, ratings, comments etc. We have introduced Mahout [4] for data analysis mechanism which is provided by the customers as reviews and ratings for the item or products. The remaining section of the paper is summarized as follows. In section 2, we more discuss on the proposed recommender system for e-commerce. In section 3, the implementation work done with Hadoop and openstack are discussed. In section 4, the conclusion and future work is considered.

2 Proposed Recommender System

In this section, a recommender system for e-commerce application is proposed. The data collected from the user will be in unstructured format. We use R tool to convert the unstructured format of data to the structure format. We didn’t include much detail about the R tool in this paper. The main task and objective of this system is to provide the user with the better product/items by understanding his choice of selection. Basically in choosing and selecting items different users have different perspectives. Basically the recommender system can be classified into two different categories. Firstly the user-based recommender and the other one is item/product-based recommender system. Some of the issues that may arise with this system are discussed below.

a) Most of the time the user who provide ratings to the item will make use of his own regional language for communication purpose. To overcome this, the conversion has to be made to English to understand what the user says. So there should be a mapping to the dictionary which helps the corresponding regional language to English.

b) A lexical resource such as SentiWordNet can be used to classify the +ve and -ve reviews and comments. SentiWordNet [6] can be used for identifying these activities and it’s a free resource available for these type of classification.

c) There are lot of possibilities that users can provide the review with the irrelevant content. These irrelevant data must be identified and it should be completely removed from ecommerce data set. SentiWordNet can be used for this purpose of classification also.

d) Users can provide the review with some condition mentioning the working of product only in this specific condition. So these type of classification needs to be done.

e) There are many reviews provided as a spam mentioning the product is good. But in real scenario the product may not be good. The organizations or the product manufacturer will sponsor the popular sites for making the product a successful one. There are many research works that are carried out in this area [9, 10]. There are several techniques available in the recommender system; mainly two techniques are used for filtering namely collaborative filtering technique and content-based technique. In collaborative based filtering techniques, the recommendations are based on users knowledge with items or products. In this technique there will be no knowledge on characteristics of the items or products. While in content-based technique, the recommendations are based on the items attributes and its contents.

2.1 Collaborative Recommendation Filtering Technique

In this technique, a model is developed based upon the past history of the users who purchased the items or products. It can be also with respect to the users who give rating/feedback to the item or product also with the other users feedback and comments given.

2.2 Collaborative Recommendation Filtering Technique

In this technique, a model is developed based upon the past history of the users who purchased the items or products. It can be also with respect to the users rating given to the item or product also with the other users feedback and comments given.

2.3 Content-based Recommendation Filtering Technique

In this technique, a series of discrete item or product characteristics are identified and analysed to provide the recommendation of other similar items with same properties. These techniques are combined with another recommender system which is a hybrid system [2, 5]. This technique is based on the item description and the preferences given by the user or customer. The characteristics of an item are indicated using the keywords which are used for creating and making recommendations. In this technique, the items are compared with the previous user ratings and the recommendation is provided by providing the best match with the items. Basically this method is used for the purpose of information retrieval mechanisms [2, 5] to be handled in the system. Some of the issues that may arise with this system are discussed below. In this research work, we develop a recommender system for processing huge amount of data which is available in the form of user ratings, user reviews, feedback etc., about the item or product available online. This system was developed using Hadoop framework. Here in this system we follow the hybrid filtering mechanism for filtering reviews and comments provided by the user. Figure 1 depicts the proposed system architecture of the recommender based e-commerce system.

![Figure 1: Proposed System Architecture](image-url)
Edata: It contains information tuples such as userId, ItemId, Rating, TimeDuration.
Euser: It contains information tuples such as userId, age, gender, and other user personal details.
E.item: It contains information tuples such as ItemId, ItemName, PckgDate, ExpiryDate etc...

Apache Hadoop provides the framework for storing and processing huge amount of data. HDFS is used as a file system which is designed for storing information in commodity hardware which is of very low cost. MapReduce is used as the programming model for data processing. The Apache Mahout is basically machine learning tool for analysis of big data.

3 Implementation Results
The software and hardware requirement for running this recommendation system chosen is given below.

a) Software Requirement
- Linus OS (We used CentOS 7)
- Openstack – devstack
- Ubuntu 14.04 Linus OS
- Hadoop - hadoop-2.1.7-bin.tar
- Java - jdk-7
- Mahout - mahout-distribution-0.7.tar
- e-commerce application shopping Dataset

b) Hardware Requirement
- 4GB RAM
- 100GB Hard Disk

The steps to be followed are given below
1. Install Openstack using devstack.
2. Open the Openstack dashboard and give username and password.
3. Launch a VM instance.
4. Choose Ubuntu 14.04 and launch is using VM
5. Install java to run Hadoop.
6. Run Mahout
7. Get the e-commerce data set.
8. Run the Mahout recommender system

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
<th>Model</th>
<th>Rating (1 to 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Television</td>
<td>Samsung</td>
<td>LED</td>
<td>4</td>
</tr>
<tr>
<td>Television</td>
<td>Sony</td>
<td>LED</td>
<td>4</td>
</tr>
<tr>
<td>Television</td>
<td>Onida</td>
<td>LED</td>
<td>3</td>
</tr>
<tr>
<td>Television</td>
<td>LG</td>
<td>LED</td>
<td>3.5</td>
</tr>
<tr>
<td>Television</td>
<td>Samsung</td>
<td>LCD</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Sample Data for User Rating on Products

Figure 2: User Ratings on Products

<table>
<thead>
<tr>
<th>Company</th>
<th>Model</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>LED</td>
<td>5</td>
</tr>
<tr>
<td>Sony</td>
<td>LED</td>
<td>5</td>
</tr>
<tr>
<td>Onida</td>
<td>LED</td>
<td>3</td>
</tr>
<tr>
<td>LG</td>
<td>LED</td>
<td>4</td>
</tr>
<tr>
<td>Samsung</td>
<td>LCD</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Recommended Televisions

Table 1 and Table 2 shows the sample data for user rating given for the products and the recommended product. Figure 2 shows the visual representation of user ratings on a product.

3 Conclusion
The data collected by the recommender system is in the form of reviews, comments, feedback etc..,This information cannot be utilized directly for analysis. In this paper we given a R tool for converting the unstructured data into a structure data. Much work about R tool was not discussed here. We also discussed about the different filtering techniques and the issues raised in the recommender system. For our testing we used the sample shopping dataset for analysis based on the user ratings. 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


[10] Nikita Spirin, Jiawei Han, “Survey on Web Spam Detection: Principles and Algorithms”, 2012, SIGKDD Explorations Volume 13, Issue 2, page 50-64. 1–49