

# Survey of Development of pattern storage system and analysis of online shopping using DMT.

Asha V.Bande<sup>1</sup>, Shrutika G.Nimadge<sup>2</sup>, Snehal Y.Katkar<sup>3</sup>, Sayali V.Jagtap<sup>4</sup>,

Prof. Vilas B. Khedekar<sup>5</sup>

<sup>1,2,3,4</sup>BE Student Department of Computer Engineering, JSCOE, Pune

<sup>5</sup>Assistant Professor, Department of Computer Engineering, JSCOE, Pune

\*\*\*\*\*

**Abstract** - In previous system the customer or user can give the online feedback on every item. It will be positive or negative. High volumes of valuable uncertain data can be easily collected or available at high speed in real-life applications. Users interested in mining all frequent patterns from the uncertain Big data; in other situations, users interested in only a tiny portion of these mined patterns. In order to reduce the calculation and to focus on mines in the later stages, we propose data science solutions that use the mining categorization technique of data to satisfy the user's limitations as specified by the precise Big Data. The result of data mining technique which call pattern is stored in data warehouse.

**Key Words:** Big Data, Hadoop, Pattern, Web Log File, Pattern Warehouse, Data Warehouse, DMT

## 1. INTRODUCTION-

Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations to human behaviour and interactions. It is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. Pattern mining consists of using data mining algorithms to discover interesting and useful patterns in databases. Data mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc. Data mining algorithms can be designed to discover various types of patterns: subgraphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns, high-utility patterns, etc. There are several definitions. For example, some researchers define an interesting pattern as a pattern that appears frequently in a database. Other researchers want to discover rare patterns, patterns with a high confidence, the top patterns, etc. In the following examples of two types of patterns that can be discovered from a database. Discovering frequent item sets The most popular algorithm for pattern mining is Apriori Algorithm (1993). It is designed to be applied on a transaction database to discover patterns in transactions made by customers which is stores. But it can also be applied on other applications. A transaction is defined a set of distinct items (symbols).

Apriori takes as input:

- (1) A minsup threshold set by the user.
  - (2) A transaction database containing a set of transactions.
- Apriori output is an all frequent item sets, i.e. groups of items shared by no less than minimum support transactions in the input database.

**Existing System**- Every organization or various websites generated large amount of data from a various source. Web mining is a process that extracts useful information from web resources. Log files are maintained by the web server. The challenging task for E-commerce companies is to analyze web log files. An e-commerce website can generate tens of Peta bytes of data in their web log files.

**Proposed system** - Paper discusses the importance of log files in E-commerce world. The analysis of log files is help full for learning the user behavior and large web log files needs parallel processing and reliable data storage system. The Hadoop framework provides reliable storage for (HDFS) Hadoop Distributed File System and parallel processing system for a large database using MapReduce programming model. Above two mechanisms help to process web log data in a parallel manner and compute results efficiently. Proposed system approach reduces the response time and load of the system. The main Aim of our system is to collect data and maintain it using data mining techniques. User can login and view the products, give feedback, and also can see reviews of other people. On selecting any item user can see all the details of that item sitting at any geographical region.

## 2. REVIEW OF LITERATURE

### 1. "Big data analysis in e-commerce system using Hadoop Map Reduce"

The analysis of log files is used for learning the user behavior in E-commerce system. The analysis of such large web log files need parallel processing and reliable data storage system. The Hadoop framework provides reliable storage by Hadoop Distributed File System and parallel processing system for large database using Map Reduce programming model.

## 2. "The electronic Commerce in the era of Internet of Things and BigData"

Nowadays big data is being used to create wealth in many fields, in particular e-commerce is playing an increasing role in modern life and the role of big data in this sphere is constantly evolving.

## 3. "Big data analytics: hadoop and tools"

Information technology gives utmost importance to processing of data. Some peta bytes of data is not sufficient for storing large amount of data. Large volume of unstructured and structured data that gets created from various sources such as Emails, web logs, social media like Twitter, Facebook etc.

## 4. "A Survey on Deep Learning in Big Data"

Big Data means extremely huge large data sets that can be analyzed to find patterns, trends. One technique that can be used for data analysis so that able to help us find abstract patterns in Big Data is Deep Learning. If we apply Deep Learning to Big Data, we can find unknown and useful patterns that were impossible so far. With the help of Deep learning, AI is getting smart.

## 5. "Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-mining-as-a-service Paradigm"

Cloud computing is popularizing the computing paradigm in which data is outsourced to a third-party service provider (server) for data mining. Outsourcing, however, raises a serious security issue: how can the client of weak computational power verify that the server returned correct mining result? In this paper, we focus on the specific task of frequent itemset mining. We consider the server that is potentially untrusted and tries to escape from verification by using its prior knowledge of the outsourced data. We propose efficient probabilistic and deterministic verification approaches to check whether the server has returned correct and complete frequent itemsets.

## 6. "Efficient Incremental Itemset Tree for Approximate Frequent Itemset Mining On Data Stream"

Mining frequent itemsets and association rules on data stream is an important and challenging task. Tree based approaches have been extensively studied and widely used for their parallel processing capability. Itemset Tree is an efficient data structure to represent the transactions for performing selective mining of frequent itemsets and association rules. The transactions are inserted incrementally and provide on-demand ad-hoc querying on the tree for finding frequent itemsets and association rules for different support and confidence values. However the size of tree grows larger for unbounded data streams limiting the scalability.

## 7. "Frequent Itemset Mining Techniques"

Frequent Itemset Mining is one of the most popular techniques to extract knowledge from data. However, these mining methods become more problematic when they are applied to Big Data. Fortunately, recent improvements in the field of parallel programming provide many tools to tackle this problem.

## 3. SYSTEM OVERVIEW

The main Aim of this system to collect data and maintain it using data mining techniques. User can login and view the products, give feedback, and also can see reviews of other people. On selecting any item user can see all the details of that item and also see the related Patterns of that Product. this data already stored on pattern warehouse and if data or pattern is not available the data mining technique is applied on data warehouse then this generated pattern is stored on pattern warehouse. Mainly we use Association rule Data mining Technique. In that we use the Apriori Algorithm for Pattern Mining.

### 1)Apriori Algorithm :

Apriori is a algorithm for frequent item stmining and association rule learning over transactional database. It proceeds by identifying the frequent individual item in the database and extending them to longer and larger item sets as long as those item appear sufficient ofter in database. The frequent item set determined by Apriori can be used to determine association rule

### 2)Storing Data :

There are four types of data models we have faced in Big Data area: 1- data that we can store them in relational 2- semi structured data same as XML 3- graph data such as those we use for social media and the last one is unstructured data such as text data, hand-written articles. So companies decided to implement their own file system (HDFS), distributed storage systems(Google Bigtable), distributed programming frameworks (MapReduce), and even distributed database management systems(Apache Cassandra) . Furthermore, Big Data management is a complex process especially when data are gathered from heterogeneous sources to be used for decision-making and scoping out a strategy. Big Data management area brought new challenges in terms of data fusion complexity, storage of data, analytical tools and shortage of governance. We also can categorize.

### Mathematical Model:

Let Assume S be the system which execute Analysis of data of E-commerce on Big data.

$$S=\{s,e,X,Y,T,F_{main},NDD,DD,Success,Failure\}$$

- **S(System)** = Is our proposed system which includes following tuple.
- **s (initial state at time T)** = GUI of Login . The GUI provides space to enter a valid id of user.
- **X (input to system)** :- Input Query. The user has to first enter the query i.e Product Name. The query may be ambiguous or not. The query also represents what user wants to search.
- **Y (output of system)** :- List of URLs with Snippets. User has to enter a query into search then search engine generates a result which contains relevant and irrelevant URL's and also display the Time required for execution of specific query.
- **T (No. of steps to be performed)** :- 4. These are the total number of steps required to process a query and generates results.
- **f<sub>main</sub>(main algorithm)** :- It contains Process P. Process P contains Input ,Output and subordinates functions. It shows how the query will be processed into different modules and how the results are generated.
- **DD (deterministic data)**:- It contains Database data. Here we have considered. SQLite which contains number of queries. Such queries are user for showing results. Hence, SQLite is our DD.
- **NDD (non-deterministic data)**:- No. of input queries. In our system, user can enter numbers of queries so that we cannot judge how many queries user enters into single session. Hence, Number of Input queries are our NDD.
- **Memory shared:** - MYSQL. MYSQL will store information like list of information about registration details , previous History, Product and their Related Patterns and numbers of Reviews. Since it is the only memory shared in our system, we have included it in the MYSQL.
- **CPU<sub>count</sub>:** - In our system, we require 1 CPU for server.
- **Success** = successfully recommended best system as per user's interest.
- **Failure** = Failed to be recommended.
- **Subordinate functions:**X=Set of Input

$$X = \{x1,x2,x3\}$$

Where , x1= Registration and login Detail

x2= Item Details

x3= Select any item for the sales history.

Y=Set of Outputs

$$Y = \{y1,y2\}$$

Where, y1= Sales details in daily, weekly, monthly, yearly in the graph format.

y2= See the reviews of the items.

F<sub>main</sub> = Set of procedure

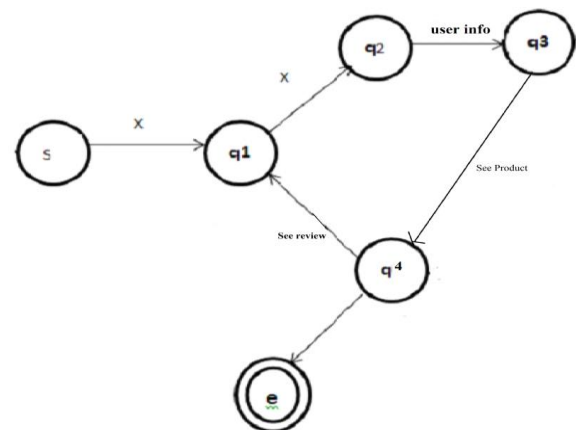
$$F = \{f1,f2,f3,f4\}$$

Where, f1= Take x1 & x2 input

f2= Give y1 output

f3= Take x3 input

f4= Give y2 output



#### 4. SYSTEM ANALYSIS

##### Apriori Algorithm-

Procedure Apriori(T,minSupport) { //T is the database and minSupport is the minimum support

L1={frequent items};

```

For(k=2;L(k-1)!=0;k++){
  Ck=candidates generated from L(k-1)
  //that is Cartesian product L(k-1) x L(k-1)
  and eliminating any (k-1) size item that is not
  //frequent
  For each transaction t in data base do{
    #increment the count of all
    candidates in Ck that are contained in t
    Lk=candidates in Ck with
    minSupport
  } //end for each
} //end for
return UkLk;
}
  
```

##### Steps of Apriori Algorithm:

**Step 1:** Count the number of transactions in which each item occurs.

**Step 2:** Now remember we said the item is said frequently bought if it is bought at least 3 times. So in this step we remove all the items that are bought less than 3 times from the above table and we are left with This is the single items that are bought frequently. Now let's say we want to find a pair of items that are bought frequently.

**Step 3:** We start making pairs from the first item

**Step 4:** Now we count how many times each pair is bought together

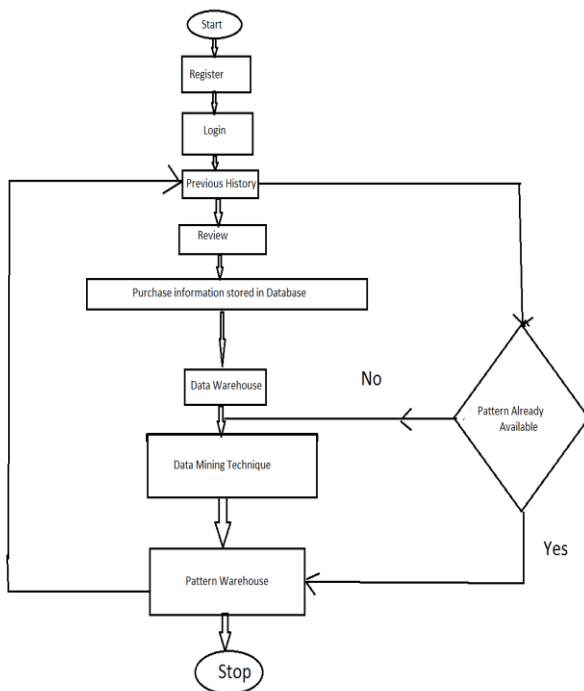
**Step 5:** Golden rule to the rescue. Remove all the item pairs with number of transactions less than.

**Step 6:** three and we are left with These are the pairs of items frequently bought together.

**Step 7:** To make the set of three items we need one more rule (it's termed as self-join).

**Step 8:** So we again apply the golden rule, that is, the item set must be bought together at least 3 times

**FlowChart**



**5. CONCLUSIONS**

Hence, we conclude that in our system the data retrieval time will be reduced by implementing data mining techniques. an e-commerce website can generate tens of Peta bytes of data in their web log files. The analysis of log files is used for learning the user behavior in an E-commerce system. The analysis of such large web log files needs parallel processing and reliable data storage system. The Hadoop framework provides reliable storage for Hadoop

Distributed File System and parallels processing system for a large database using MapReduce programming model.

**6. REFERENCES**

1. "Big Data Analysis in e-commerce system using Hadoop MapReduce" 2016(IEEE), S.Suguna:M.Vithya ; J.I.Christy Eunaicy.
2. "The electronic Commerce in the era of Internet of Things and BigData", 2017(IEEE) ,Jing Liu;Lili Sun; Russell Higgs; Yuanyuan Zhang; Yan Huang.
3. "Big data analytics: hadoop and tools", 2016(IEEE), Mrunal Sogodekar; Shikha Pandey; Isha Tupkari; Amit Manekar.
- 4."A Survey on Deep Learning in Big Data",2017(IEEE), Mehdi Gheisari; Guojun Wang; Md Zakirul Alam Bhuiyan.
5. "Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-mining-as-a-service Paradigm" 2015(IEEE), Boxiang Dong; Ruilin Liu; Hui (Wendy) Wang.
6. "Efficient Incremental Itemset Tree for Approximate Frequent Itemset Mining On Data Stream", 2016 (IEEE), Pavitra Bai S Assistant Professor, Dept., of Information Science S.J.B Institute of Technology Bangalore.
7. "Frequent Itemset Mining Techniques", 2016(IEEE), Tushar M. Chaur Department of Computer Technology YCCE Nagpur, India.
8. "Converting an E-commerce Prospect into a Customer using Streaming Analytics", 2016 (IEEE), Sahana Raj G ;Dr. B Sathish Babu Department of Computer Science and Engineering Professor,
9. "Accelerating the Mobile Cloud:Using Amazon Mobile Analytics and K-Means Clustering", 2016 (IEEE), Matthew Beck;Wei Hao; Alina Campan Department of Computer Science - Northern Kentucky University.
- 10." A Hierarchical Framework with Consistency Tradeoff", 2017 (IEEE) ,, Yingyi Yang, Yi You, Bochuan Gu.Department of Smart Grid Strategies for Big Data Management.
11. "The Application of Data Mining Technology to Big Data", 2017(IEEE), Jinlong Wang, School of Mathematics and Statistics.
12. "Data Analysis and Visualization of Sales Data",2016(IEEE),Kiran SinghDepartment of Computer Technology YCCE Nagpur, India.