iJETRM
International Journal of Engineering Technology Research & Management

SURVEY OF DEVELOPMENT OF PATTERN STORAGE SYSTEM USING DATA MINING TECHNIQUE

Jyoti Kaushal
Assistant Professor, Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies, Udaipur

Abstract –
In previous system the customer or user can gave the online feedback on every item . It will be positive ornegative. 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 patternsfrom the uncertain Big data; in other situations, users interested in only a tiny portion of these mined patterns. Inorder to reduce the calculation and to focus on mines in thelater stages, we propose data science solutions that use themining categorization technique of data to satisfy the user'slimitations as specified by the precise Big Data. The result ofdata mining technique which call pattern is stored in data warehouse.

Keywords:
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 associationsto human behaviour and interactions. It is a phrase used tomean a massive volume of both structured and unstructureddata that is so large it is difficult to process using traditionaldatabase and software techniques. Pattern mining consistsof using data mining algorithms to discover interesting and useful patterns in databases. Data mining algorithms can beapplied 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 wants to discover rare patterns, patterns with a high confidence, thetop patterns, etc. In the following examples of two types of patterns that can be discovered from a database. Discoveringfrequent 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 itemsshared by no less than minimum support transactions in theinput database.

Existing System- . Every organization or various websites generated large amount of data from a various source. Webmining 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 generatetens of Peta bytes of data in their web log files.

Proposed system - Paper discusses the importance of logfiles 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 processingsystem for a large database using MapReduce programmingmodel. Above two mechanisms help to process web log datain 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 collectdata and maintain it using data mining techniques. User canlogin and view the products, give feedback, and also can seereviews of other people. On selecting any item user can seeall the details of that item sitting at any geographical region.

2. REVIEW OF LITERATURE

1.“Big data analysis in e-commerce system using HadoopMap Reduce”

The analysis of log files is used for learning the userbehavior in E-commerce system. The analysis of such largeweb 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.

iJETRM
International Journal of Engineering Technology Research & Management

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

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

3. "Big data analytics: hadoop and tools"

Information technology gives utmost importance toprocessing of data. Some peta bytes of data is not sufficientfor 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 thatcan be analyzed to find patterns, trends. One technique thatcan 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 usefulpatterns that were impossible so far. With the help of Deepearning, 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 computingparadigm 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 clientof weak computational power verify that the server returnedcorrect mining result? In this paper, we focus on the specifictask of frequent itemset mining. We consider the server thatis potentially untrusted and tries to escape from verificationby using its prior knowledge of the outsourced data. We propose efficient probabilistic and deterministic verificationapproaches to check whether the server has returned correct and complete frequent itemsets.

6. "Efficient Incremental Itemset Tree for ApproximateFrequent Itemset Mining On Data Stream"

Mining frequent itemsets and association rules on data stream is an important and challenging task. Tree basedapproaches have been extensively studied and widely usedfor 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 onthe tree for fining frequent itemsets and association rules fordifferent support and confidence values. However the size oftree grows larger for unbounded data streams limiting the scalability.

7. "Frequent Itemset Mining Techniques"

Frequent Itemset Mining is one of the most populartechniques to extract knowledge from data. However, these mining methods become more problematic when they are applied to Big Data. Fortunately, recent improvements in thefield 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 andview the products, give feedback, and also can see reviews ofother people. On selecting any item user can see all the details of that item and also see the related Patterns of thatProduct . this data already stored on pattern warehouse andif 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 ruleData mining Technique. In that we use the Apriori Algorithmfor Pattern Mining .

1) Apriori Algorithm :

Apriori is a algorithm for frequent item stmining and association rule learning over transactional database. Itproceeds by identifying the frequent individual item in the database and extending them to longer and larger item setsas long as those item appear sufficient ofter in database. Thefrequent 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. Socompanies 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 andscoping out a strategy. Big Data management area brought new challenges in terms of data fusion complexity, storage ofdata, analytical tools and shortage of governance. We also can categorize.

Mathematical Model:

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

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

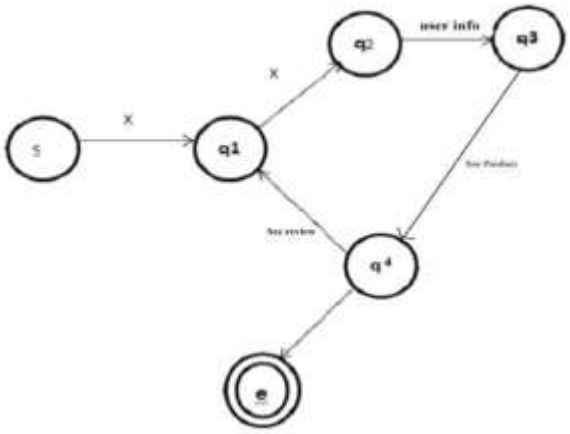
iJETRM
International Journal of Engineering Technology Research & Management

- **S(System)** = Is our proposed system which includesfollowing tuple.
- **s (initial state at time T)** = GUI of Login . The GUIprovides space to enter a valid id of user.
- **X (input to system)** :- Input Query. The user has tofirst 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 Timerequied for exection 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_{main}(main algorithm)** :- It contains Process P. Process P contains Input ,Output and subordinatesfunctions. It shows how the query will be processedinto different modules and how the results aregenerated.
- **DD (deterministic data)**:- It contains Database data. Here we have considered. SQLite which contains number of queries. Such queries are userfor 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 queriesuser 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 andtheir Related Patterns and numbers of Reviews. Since it is the only memory shared in our system, we have included it in the MYSQL.
- **CPU_{count}**: - In our system, we require 1 CPU forserver.
- **Success** = successfully recommended best system as per user's interest.
- **Failure** = Failed to be recommended.
- **Subordinate functions:**X=Set of InputX = {x1,x2,x3}

Where , x1= Registration and login Detailx2= Item Details
x3= Select any item for the sales history.

Y=Set of OutputsY = {y1,y2}
Where, y1= Sales details in daily, weekly, monthly, yearly inthe graph format.
y2= See the reviews of the items.
F_{main} = Set of procedureF= {f1,f2,f3,f4}

Where, f1= Take x1 & x2 inputf2= Give y1 output f3= Take x3 input
f4= Give y2 output



4. SYSTEM ANALYSIS

Apriori Algorithm-

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

iJETRM
International Journal of Engineering Technology Research & Management

```
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 isnot
        //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 Algmorith:

Step 1: Count the number of transactions in which each itemoccurs.

Step 2: Now remember we said the item is said frequentlybought if it is bought at least 3 times. So in this step we remove all the items that are bought less than 3 times fromthe above table and we are left with This is the single itemsthat 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 boughttogether

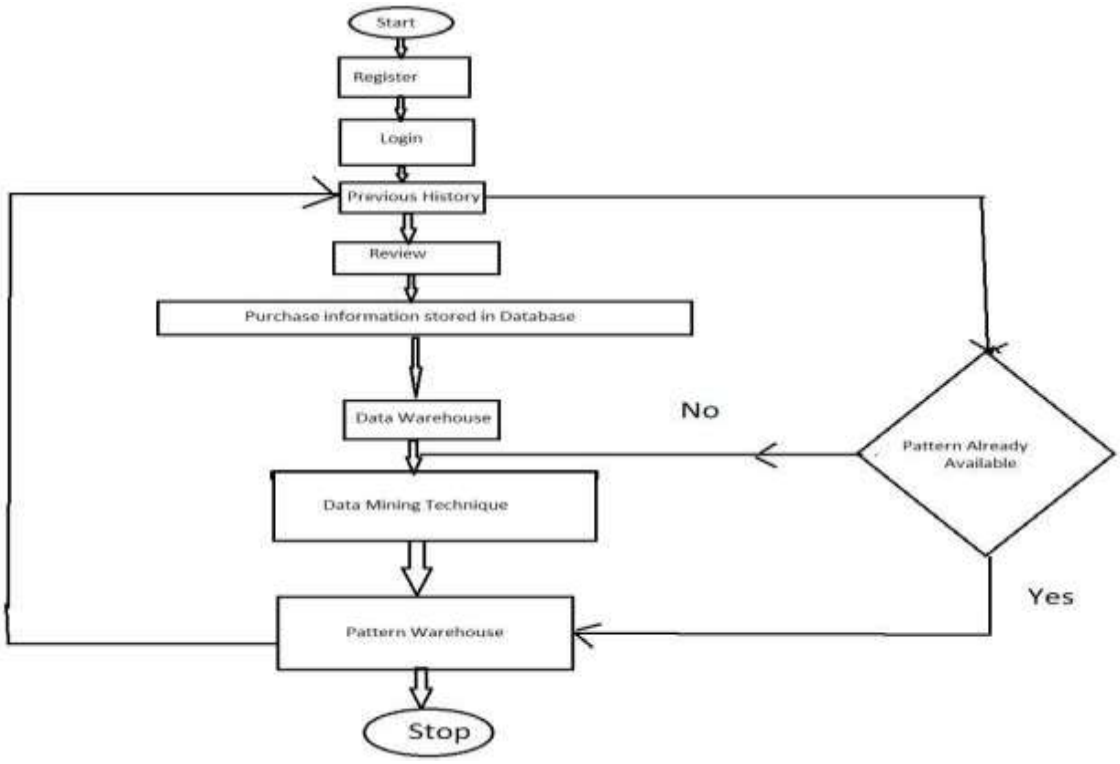
Step 5: Golden rule to the rescue. Remove all the item pairswith number of transactions less than.

Step 6: three and we are left with These are the pairs ofitems 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



iJETRM
International Journal of Engineering Technology Research & Management

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-commercesystem. 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 alarge 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 Thingsand 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; AmitManekar.

4."A Survey on Deep Learning in Big Data",2017()EEE),
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)temset Mining On Data Stream", 2016 (IEEE),Pavitra Bai S Assistant Professor, Dept., of InformationScience S.J.B Institute of Technology Bangalore.

7. "Frequent)temset Mining Techniques", 2016(IEEE), Tushar M. Chaure Department of Computer TechnologyYCCE Nagpur, India.

8. "Converting an E-commerce Prospect into a Customer using Streaming Analytics", 2016 ()EEE), Sahana Raj G ;Dr. BSathish Babu Department of Computer Science and Engineering Professor,

9. "Accelerating the Mobile Cloud:Using Amazon MobileAnalytics and K-Means Clustering", 2016 ()EEE), 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 ofSmart 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.