Survey on Mining of Data using Apriori Algorithm

Ms. Poonam Joshi[1] Ms. Reena Somani[2] Ms. Sejal Dmello[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Atharva College of Engineering, University of Mumbai, Malad, Mumbai 95, India

Abstract— Data mining is the process of extraction of concealed information from large databases. It is a new and a dominant technology with great potential to help companies to emphasis on the most important information in their data warehouses. The automated prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems. In today's time many algorithms are available to implement data mining. One such algorithm is the Apriori Algorithm proposed by R. Agrawal and R. Srikant in 1994. This project is based on the idea of designing such a system which can be instrumental in strategizing and decision making process in a more efficient way. The tools provided by this system will be helpful in predicting future trends and consumer behaviors, allowing the businesses to make pro-active and knowledge driven decisions.

Key words: Data Mining, Apriori Algorithm

I. INTRODUCTION

Data mining techniques can be categorized according to the objectives they follow and the results they offer, which obtains computer as a tool and makes use of the skill and knowledge significance to comprehend and explain the problem. Various data mining techniques such as, decision trees, association rules, apriori algorithm and neural networks are already presented and become the point of attention for several years. [1]. Data Mining is a detailed process of analysing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. [2].

A. Need

Real-time issues – your current systems aren't enabled to integrate disparate sources of data and keep historical records of those integrations, in near real-time. Scalability issues – you have tons of historical data you need to gather in to an easily accessible place, common formats, common keys, and common access methods. And you need to ensure that the system is scalable over the next 3 to 5 years. Self-Service BI – if you have a need to eventually reach this goal, where users can "visualize" and construct their own reports, then you probably need an enterprise data warehouse, along with its highly integrated historical facts from all the different sources in your organization.

B. Converter Modal and Analysis

The system is a full-fledged data warehouse which uses Data mining for Business Intelligence and thus has a wide variety of areas to benefit from like:

- Employee Module: The system will have employee database containing personal and professional information of all the employees currently under payroll and will provide all the information at one spot. It will also generate reports.
- Supplier Module: This module will facilitate order placement, delivery confirmation and transactions from all the loyal suppliers of the company and also generate reports of the same.
- Customer Module: This is where all the information of the existing customers will be available and it can also perform order placement, delivery confirmation and transaction management etc.
- Product Module: Details of all the available products, their types, in essence all the products available in real time in the Company warehouse.
- Inventory Control: It will help manage inventory and alert the user in case of some discrepancy or low level of certain product or raw material.
- Bill Generation: This module looks after the generation of bill as per the products purchased.
- Accounting Details: The employees can simply add the details of transactions into the system using the interface for future use and scrutiny.
- Mining – Using data mining the system will generate Sales Report (Daily, Weekly, Monthly, Quarterly, Half Yearly, and Yearly).
- Mining – Using data mining the system will generate Purchase Report (Daily, Weekly, Monthly, Quarterly, Half Yearly, and Yearly).
- Business Intelligence Module: This module will use the Apriori Algorithm to generate business intelligence reports for products and the type of customers who buy them.

II. PROBLEM FORMULATION

In today's world the IT industry is growing at a very fast pace in all types of industries. The usage of papers, pens and ledgers to manage inter-business transactions, maintain employee records and inventory etc. impedes one's company to grow and compete with other companies in the market. The above mentioned techniques to manage business is highly vulnerable to mutilation and loss therefore leading the company into chaos as to where the business is going in terms of profits, losses and the annual turnover. Hence a full automated procedure is required for the efficient working of a company. This automated process will be in charge of employee management, purchase, stock management, and accounts. In addition to all these functions the major feature is that this system will provide its customers with business reports and the necessary intelligence which will help he company in making strategies and making correct decisions. This system will also be the exact judge of how one's business is advancing and what other steps can be implemented so as to make it more efficient and profitable.

A. Need

The best and the most effective solution to the problem at hand is to develop an integrated system consisting of a data warehouse of the entire organization spanning databases of employees, products, bills, suppliers, orders etc. which will make management of all kind of data easy and effective and become helpful in matters of strategizing, accounting and management and using effective data mining tools in order to generate accurate reports and provide business intelligence for the betterment of the company.

III. PROPOSED METHOD

A. System Architecture

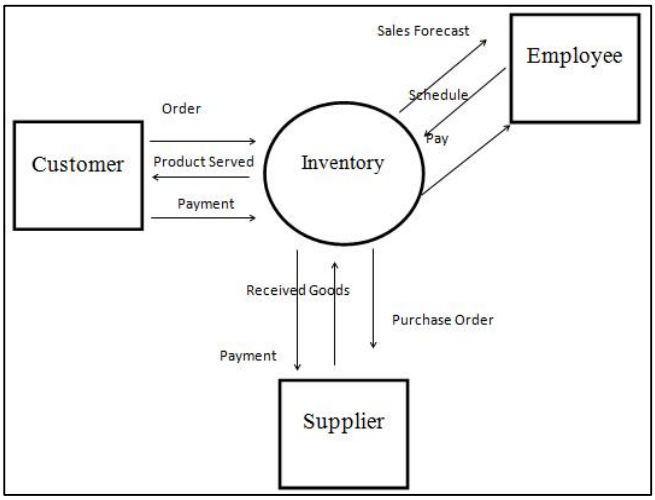

Fig. 1: Level 0 DFD of system

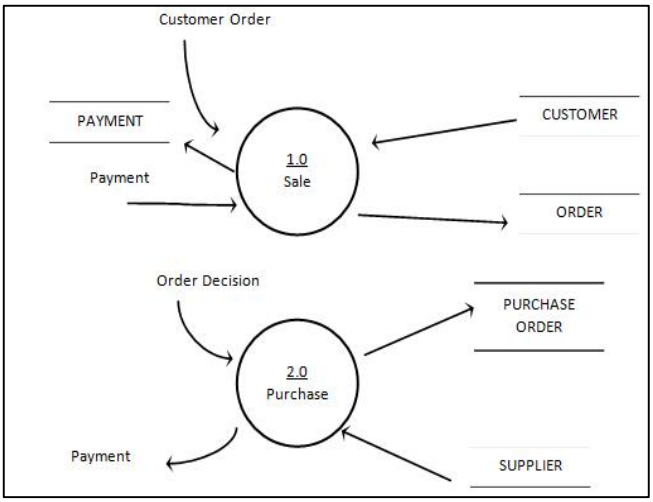

Fig. 2: Level 1 DFD of system

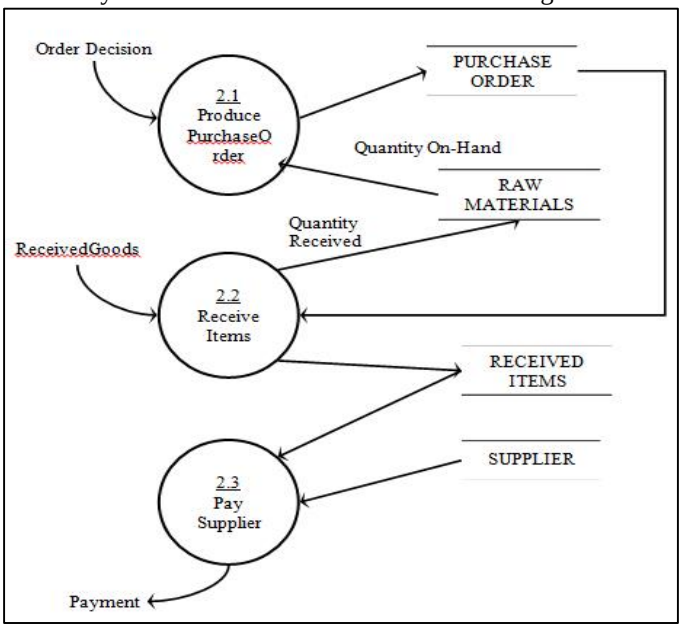

Fig. 3: Level 1 DFD of system

B. Algorithm Used

Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order.

Let the set of frequent itemsets of size k be Fk and their candidates be Ck. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1) Generate Ck+1, candidates of frequent itemsets of size k+1, from the frequent itemsets of size *k*.
2) Scan the database and calculate the support of each candidate of frequent itemsets.
3) Add those itemsets that satisfies the minimum support requirement to *Fk*+1. [3].

Initialize: k := 1, C1 = all the 1- item sets;
read the database to count the support of C1 to determine L1.
L1 := {frequent 1- item sets};
k:=2; //k represents the pass number//
while (Lk-1 ≠ Ø) do
begin
Ck := gen_candidate_itemsets with the given Lk-1
prune(Ck)
for all transactions t €T do
increment the count of all candidates in CK that are contained in t;
Lk := All candidates in Ck with minimum support ;
k := k + 1;
end
Answer: = ∪k Lk ; [4].

The first weakness of this algorithm is the generation of a large number of candidate item sets. The second problem is the number of database passes which is equal to the max length of frequent item set. [4].

C. Advantage of Apriori Algorithm

- Apriori algorithm is easy to understand.
- It is simple to implement.
- It uses large itemset property.
- It is easily parallelized.

D. Disadvantage of Apriori Algorithm

- It requires many database scans.
- It is less efficient and accurate.
- It takes more time and consumes more memory

E. Way to Improve Apriori Algorithm

- Transaction Reduction: transactions that do not consist of frequent itemsets are of no importance in the next scans for searching frequent itemsets
- Hash based itemset counting: hashing table is used for counting the occurrences of itemsets
- Partitioning: for any itemset i.e. frequent in database, then that itemset must be frequent in atleast one of the partition of database
- By adding attribute Weight and Quantity: means how much quantity of item has been purchased
- By adding attribute Profit: that can give the valuable information for business and customers.
- By reducing the number of scans
- By removing the large candidates that cause high Input/output cost.

IV. CONCLUSION

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. This Research work implements each of these phases. One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm. output of the system was in terms of memory usage and speed of producing association rules.

Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both apriori and FP-growth. In future the algorithm can be extended to web content mining, web structure mining, etc. The work can also be extended to extract information from image files. Probability the performance of the methods was illustrated using different gait databases.

REFERENCES

[1] Rachna Somkunwar. A study on Various Data Mini Approaches of Association Rules. Volume 2, Issue 9, September 2012.
[2] Divya Bansal, Lekha Bhambhu. Usage of Apriori Algorithm of Data Mining as an Application to Grievous Crimes against Women. Volume 4 Issue 9–Sep 2013.
[3] Xindong Wu et.al,"Top 10 Algorithms of Data Mining",Springer-Verlag London, 2007.
[4] Goswami D.N., Chaturvedi Anshu, Raghuvanshi C.S. An Algorithm for Frequent Pattern Mining Based On Apriori.