Data Mining Using Apriori Algorithm

Akshay Palekar#1, Komal Narwekar#2, Pooja Manjeshwar#3, Yugandhara Rane#4

1Computer Engineering Department, Atharva College of Engineering
Malad Marve Road, Charkop Naka, Malad (West), Mumbai-400095, Maharashtra, India
2Computer Science Department, University at Albany - SUNY
1400 Washington Ave, Albany, NY 12222, United States
3Information Technology Department, Atharva College of Engineering
Malad Marve Road, Charkop Naka, Malad (West), Mumbai-400095, Maharashtra, India
4Computer Science Department, New York Institute of Technology
1855 Broadway, New York, NY 10023, United States

Abstract — The abstraction of masked information from large databases is called ‘Data Mining’. This latest technology is the key to help companies understand and classify important information in their data warehouses. Data mining offers a better scope of extraction of information as it goes beyond the approach of the traditional Decision support systems which relied mainly on analyzing the past transactions. Several data mining algorithms have been developed to maximize the efficiency of information extraction. One of the most popular data mining algorithms is Apriori Algorithm which was proposed by R. Agrawal and R. Srikant. Our main objective behind designing this system is helping the user in decision making. This system will incorporate features which will help in predicting future trends and customer behavior which will in turn enable organizations to make more informed decisions.

Keywords — Data mining, Decision Support System, Apriori Algorithm

I. INTRODUCTION

In today’s world information technology is growing rapidly in all fields. In the past, organizations use to manage transactions, records and inventories using the traditional system of pen and paper. This system impeded the ability of the organization to succeed and compete with other companies. One of the most fundamental flaws of this system was that the record books were prone to damage or loss which in turn hampered the company’s business. The flaw in the traditional system led to the growth of IT in this field. There was a need to automate processes such as employee and inventory management. Software was needed in order to keep track of the company’s transactions and also provide reports to help in more efficient decision making.

II. LITERATURE REVIEW

A. Fast algorithms for mining association rules in large databases

This paper analyzes the different algorithms used in data mining implemented on a data warehouse. The basic problem is to extract association rules between items. The paper suggests that data mining algorithms such as Apriori outperform the earlier known algorithms. It proposes to combine two algorithms to make a new algorithm called as Apriori Hybrid. Empirical evaluation has shown this to be linearly dependent on the number of transactions.

B. Top 10 Algorithms of Data Mining

The paper gives detailed information about the Apriori algorithm that we are using in our project. This paper presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriori, EM, Page Rank, Ada Boost, ANN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, we provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development.

C. An Algorithm for Frequent Pattern Mining Based On Apriori

This paper also gives detailed information about the Apriori algorithm that we are using in our project. The paper suggests that frequent pattern mining is a heavily researched area with a large number of applications. Many numbers of algorithms have been devised in order to determine frequent patterns. Here three different frequent pattern mining methods (Record filter, Intersection and Proposed Algorithm) are given based on classical Apriori algorithm.

III. PROPOSED SYSTEM

The proposed system aims to make the daily tasks of sales, purchases and inventory control fully automated in an inexpensive and low maintenance way which will enable the company to derive maximum profit.
A. Employee Module
This module is where the employee database will reside. The database will contain personal and professional information of all the employees currently under payroll of the company and it will also generate reports.

B. Supplier Module
It will enable order placement, delivery confirmation and transactions from all the suppliers of the company.

C. Customer Module
Information regarding the existing customers will be available here.

D. Product Module
Particulars of the products, their types and quantity are made available through this module.

E. Inventory Control
It will manage the inventory and notify the user in case of some inconsistency.

F. Bill Generation
This module looks after the generation of bill as per the products purchased.

G. Accounting Details
The employees can simply add the details of transactions into the system using the interface.

H. Business Intelligence Module
This is the main module where the Apriori Algorithm will be used to generate business intelligence reports for products and the type of customers who buy them.

IV. APRIORI ALGORITHM
Apriori is a seminal algorithm for finding frequent itemsets using antimonotonicity 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. [2]

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; [3][4]

V. FUTURE SCOPE
The current system handles only Employee information; hence there is further scope to add tools like employee management Human Resources. Various other features like taxation, customer care, raw material handling etc. In future, more efficient data mining algorithms can be used to increase the efficiency of the system. This system can also be designed to work on today’s smartphones so as to make it more user friendly and accessible.

VI. CONCLUSION
This application is able to manage the entire database of an organization and it enables users to manage all the records at one place digitally without the hassle of pen and paper. It also provides Sales forecasts and Business Intelligence for to the user. The application provides information about sales and Purchases as per products, suppliers and customers. All this is possible in a simple and cost effective way without the need of any expensive software.
ACKNOWLEDGEMENT

We take this opportunity to express gratitude to all the Department faculty members for their help and support. We also thank our parents for the unceasing encouragement, support and attention. We would also like to place on record, our sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

REFERENCES


