Proposed Application of Distributed Association Rule Mining

Nishi Pancholi[1], Saurabh Shende[2], Isha Talekar[3], Yogita Ganage[4]

[123]Department of Information Technology, RGIT, Maharashtra, India

[4]Professor, Department of Information Technology, RGIT, Maharashtra, India

ABSTRACT

Data mining is a very paramount technique used to further divide the data into paramount information. This paper shows a detailed result predicated on the ARM technique which is utilized to further circulate a given pattern of data for the medical stores across sundry geographical locations. The customers are given the opportunity to get the medicines online as well as it is utilized by the store owners to determine the pattern and given data analysis.

Keywords: Association Rules, Frequent pattern, Data Mining, Apriori

I. INTRODUCTION

A wellknown work in Associative Rule Mining (ARM) has become a mature field of research in the market. So many research papers, articles are surveyed in the field of ARM and its benefits are discussed. This paper details some fundamentals about frequent itemset generation which avails to develop incipient algorithm for that process for great outcomes. The field of ARM is divided into: Positive rule mining, Negative rule mining. Major area of work in ARM is coming under these three categories. The positive rules are mined from a set of frequent itemsets collaboratively. Due to the lack of frequent itemset mining, the frequent itemsets are elongated to sundry formats like closed, maximum, sequential, involute frequent itemset. The above types of frequent itemset are fortified to constraint-predicated rule mining. The negative relationships between itemset are mined by rule mining process utilizing infrequent itemset. The interestingness measures play an essential role in the field of ARM kindred to the data mining process. These quantifications are discussed in paper.

II. LITERATURE SURVEY

Sundry journals and papers were studied to have a better understanding of the different concepts regarding Sodality Rule Mining and Data Mining. They are as follows:

Gurneet Kaur [2] presented a review on the rudimental concepts of ARM technique along with the recent cognate work that has been done in this field. The paper additionally discusses the issues and challenges cognate to the field of sodality rule mining. A minuscule comparison predicated on the performance of sundry algorithms of sodality rule mining has withal been made in the paper.

Priyanka Yadav et al [3] used data mining techniques and the Obnubilated Markov Model (HMM) to detect Credit Card Fraud. The clustering model used to relegate the licit and fraudulent transaction utilizing data cauterisation of region of parameter. Obnubilated Markov Model can detect whether an incoming transaction is fraudulent or not with low mendacious alarm. An HMM is initially trained with the mundane deportment of a cardholder. If an overviewed credit card transaction is not accepted by the HMM with very high probability, it is consider to be fraud based transaction. Concurrently, endeavor to ascertain that genuine transactions are not abnegated. The financial losses due to Credit card fraud affect not only the individuals but withal the merchants. Consequently, the security measures need to be taken to detect the Credit card fraud.

In this paper, Yasemin Atilgan and Firat Dogan [4] first investigate the data mining applications on centralized medical databases, and how they are utilized for diagnostic and population health, then introduce distributed databases. Determinately, the paper fixates on data mining studies on distributed medical databases.

In this paper, Research Philomath, Pramod Prasad [5] elaborates upon the utilization of sodality rule mining in extracting patterns that occur frequently within a dataset and showcases the implementation of the Apriori algorithm in mining sodality rules from a dataset containing sales transactions of a retail store. Utilization of a sodality rule mining driven application to manage retail businesses will provide retailers with reports regarding prognostication of product sales trends and customer demeanor. This will sanction retailers to make hands-on, erudition-driven decisions.

III. ASSOCIATION RULE

Initially it was largely incentivized to understand the market basket data, the results of which sanctioned companies to understand purchasing demeanor and, as a result, better target market audiences. ARM is utilizer centric as the objective is the elicitation of intriguing rules from which incipient erudition can be derived. ARM is to facilitate the revelation, heuristically filter, and enable the presentation of these inferences or rules for subsequent interpretation by the utilizer to determine their usefulness. ARM has been divided on basis of phases to achieve it's goal as follows:

Phase 1: Identify the sets of frequent items or itemsets or pattern within the set of transaction utilizing utilizer- designated support threshold.

Phase 2: Engender inferences or rules from these above patterns utilizing utilizer-designated confidence threshold.

The above two phases are engendered vigorous sodality rules from dataset. The first phase of the known ARM is called frequent itemset mining. That is prodigiously computational sumptuous than phase 2. The second phase is called sodality rule generation. That is, straight forward process. This phase computational intricacy is negotiable to compare with first phase. There are two major quandaries in second phase. The first quandary is rule quantity designates that algorithms can engender immensely colossal number of rules. The second quandary is rule quality designates that, all the rules are not intriguing.

The two major techniques on which ARM technique is based are called as minimal support and minimal confidence respectively. Support is defined as the records defined from A to B. Let us assume the item support is 0.1%, it means only 0.1 percent of the transaction will pertain this item. Confidence of an association rule is defined as the fraction entity of the number of transactions derived from a set that contain A. B to the total of the records that contain A. association rule A B is 80%, it means that 80% of the transactions that contain A have the ratio of containing B together. To illustrate this concept, a varied example of a superstore is evacuated. The set of items is I= {bread, egg, butter,} and a pertained database containing the items (1 shows that item is

present and 0 shows that item is not present in a transaction). An example rule for the superstore could be {bread, egg} => {butter} meaning that if bread and egg are bought, customers also buy butter.

T	Bread	Butter	Egg
T1	1	1	0
T2	1	1	1
T3	0	1	1

IV. FREQUENT PATTERN MINING

Patterns are set of different items, related structures and derived graphs in a dataset. The frequency of pattern is no less than a utilizer-designated threshold that is called frequent pattern or itemset. Finding frequent patterns plays a fundamental role in sodality rule mining, relegation, clustering, and other data mining tasks. Frequent pattern mining was first proposed by Agarwal et al [1] for market basket analysis in the form of sodality rule mining. The fundamental frequent pattern algorithms are relegated into two ways as follows:

1. Candidate generation approach (E.g. Apriori algorithm)
2. Without candidate generation approach (E.g. FP-magnification algorithm)

A. Candidate Generation Approach

i. Apriori Algorithm

First, the algorithm was derived as AIS. Later, the algorithm was ameliorated and called Apriori. The main amelioration has developed the monotonicity property of the fortification of sets [4]. After the amelioration, the monotonicity further got better by Mannila et al [39] and Agarwal et al [2]. The Apriori algorithm is predicated on candidate generation approach. The Apriori algorithm is implemented with sundry data structures in more detail.

ii. Extension of Apriori

Since the Apriori algorithm was evolved, there have been major studies on the enhancement or future extended vision of Apriori. The expanded algorithms are classified into following nine ways:

- Transaction reduction and mapping technique
- Hashing technique
- Partitioning technique
- Sampling approach
- Incremental mining
- Parallel and distributed mining
- Integrating mining with relational database systems
- Level-wise mining approach

i.Advantages and Disadvantages of Candidate Generation Approach

Advantages

1. It significantly abbreviates the size of candidate sets utilizing the Apriori principle. It uses sizably voluminous itemset property.
2. It is facilely parallelized.
3. It is facile to implement with all kind of authentic datasets.

Disadvantages

1. It engenders astronomically immense number of candidate sets.
2. When the farthest frequent itemsets is k, Apriori requires k passes of database scans. So, it will have low efficiency.
3. Repeatedly scanning the database and checking the candidates by pattern matching.

B. Without Candidate Generation Approach

i. FP-Growth Algorithm

Han et al moulded an FP-growth method that mines the vast set of frequent itemsets without use of candidate generation. It employed in a divide-and-surmount manner. In first scan, the database derives a list of frequent items in which items are injuctively authorized by frequency descending order. The database is compressed into a frequent pattern tree (FP-tree) utilizing frequency descending order list. The FP-tree is mined by beginning from each frequent length-1 defined pattern, constructing its conditional pattern base to work, then building its conditional FP-tree, and showcasing mining recursively on such a tree. The pattern growth is resulted in a way by the concatenation of the suffix pattern with the overcoming frequent patterns provoked from a conditional FP-tree. It exploits the least frequent items as a suffix, offering good selectivity. The performance studies of FP-growth exhibit that the method significantly abbreviates search time.

ii.Extended Algorithms

There are many substitutes and allowances to the FP-growth approach, including

1. Depth first generation of frequent itemsets.
2. H-Mine (Hyper-structure Mining) algorithm.
3. Building alternative trees.

a) Benefits of using Without Candidate Generation Approach

1. It conserves consummate information for frequent pattern mining.
2. It minimizes impertinent information or infrequent items are gone.
3. The frequency descending authoritatively mandating is more liable to be shared.
4. It does not make transaction set more immensely colossal than the pristine database.
5. It is much more expeditious than Apriori algorithm.

TABLE: PERFORMANCE REVIEW OF SOME ALGORITHMS

Association Rule Mining Algorithm	**Benefits**	**Limitations**
AIS	1. An estimation is utilized in the algorithm to prune those candidate itemsets that have no hope to be sizably voluminous. 2. It is congruous for low cardinality sparse transaction database.	1. It is constrained to only one item in the consequent. 2. Requires Multiple passes over the database. 3. Data structures required for maintaining sizably voluminous and candidate itemsets is not designated.
Apriori	1. This algorithm has least recollection Consumption. 2. Easy implementation. 3. It utilizes Apriori property for pruning Consequently, itemsets left for further support checking remain less.	1. It requires many scans of database. 2. It sanctions only a single minimum support threshold. 3. It is auspicious only for diminutive database. 4. It expounds only the presence or absence of an item in the database.
FP- growth	1. It is more expeditious than other sodality rule mining algorithm. 2. It utilizes compressed representation of pristine database. 3. Reiterated database scan is eliminated.	1. The recollection consumption is more. 2. It cannot be utilized for interactive mining and incremental mining. 3. The resulting FP-Tree is not unique for the same logical database

V. EXISTING SYSTEM

The current management of individual medical store consists of a standalone desktop application which is limited to manual maintenance of records, cash flows, stock, etc. The analysis of their data is also done manually. There was a central server present which collects the data from all the stores and stores it for future reference. This data which is stored is just meant for retrieval of past records, no further analysis were done on it like how much is the sales of a particular medicine, what is the demand for it and in which area the demand is high as well as no frequent patterns were found. The shopkeeper has to enter its data on the server and then the processing is done on server side. The framework of the system is shown below.

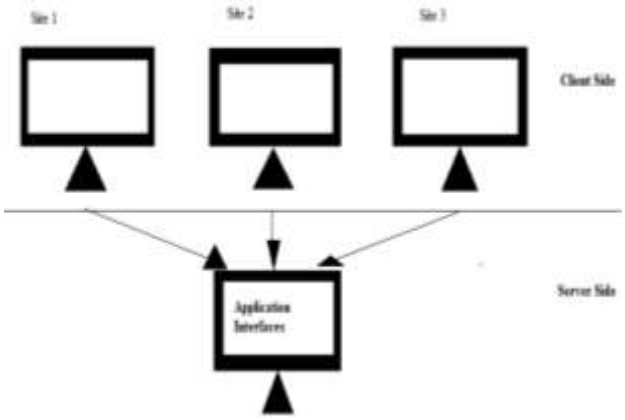

FIG V.I: ARCHITECTURE OF CENTRALIZED SYSTEM

Drawbacks of existing system:

1. The current system of medical stores is centralized which is inefficient.
2. Collaboration and communication among medical stores does not take place.
3. There is a loss of products due to expiration which leads to wastage.
4. Analysis of data done by medical stores is time consuming and is not done in an efficient manner.

VI. PROPOSED SYSTEM

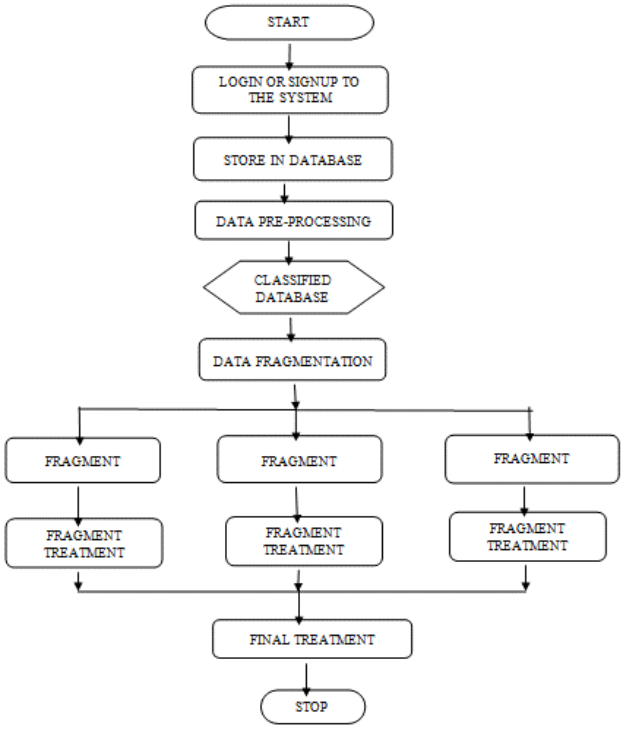

FIG VI.I: PROCESS FLOW DIAGRAM

STAGES:

- The first stage consists of data pre-processing where the raw data would be taken from the database and all the pre-processing stages would be called here.
- The second stage consists in fragmenting the data to be searched for association rules. Here, the partitioning takes place horizontally on the basis of the location of the sites.
- The third stage is computing frequent item sets and association rules individually. Local count is calculated on each site and then it can be globally distributed.

VII. SYSTEM ARCHITECTURE

Frequent Pattern Finding Algorithm:

To solve this problem we use 2 algorithms stated below:

I. Apriori Algorithm

II. Count Distributed Algorithm

Statistical analysis of the data based on algorithms:

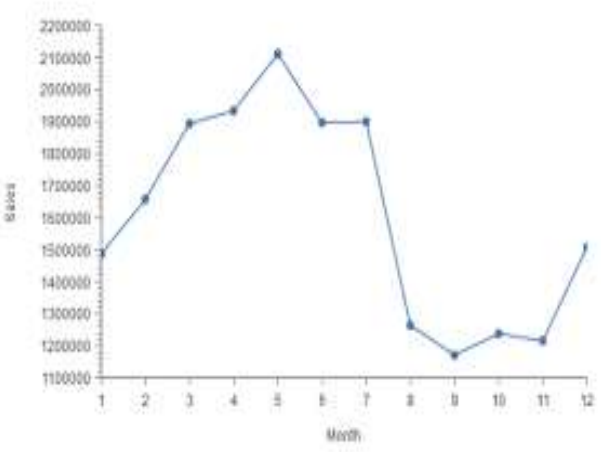

FIG VII: Monthly Sales

VII.I. Apriori Algorithm

There have been many algorithms developed for mining frequent patterns which can be classified into two categories:

- Candidate generation-test
- Pattern growth method.

The first category, the candidate-generation and test approach such as the Apriori algorithm is directly based on an important property of frequent item sets: if a pattern (set) with k items is Apriori Algorithm. Let Fk be the set of frequent item sets of size k. It first scans the database and searches for the frequent item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. All non-empty subset of frequent itemset must be frequent. The anti-monotonicity of support measure is the key concept of Apriori algorithm. Apriori assumes that

All subsets of a frequent itemset must be frequent (Apriori property). If an itemset is infrequent, all its supersets will be infrequent. It then iterates on the following three steps to extract all the frequent item sets.

1. From the frequent item sets of size k, generate Ck+1, candidates of frequent item sets of size k+1.
2. Calculate the support of each candidate of frequent item sets from the database.
3. Satisfy the minimum support requirement by adding those item sets to Fk+1.

Apriori Algorithm states:

Ck: Candidate itemset of size k

Lk: frequent itemset of size k

L1 = {frequent items};

for (k = 1; Lk! = ∅; k++)

do begin Ck+1 = candidates generated from Lk;

for each transaction t in database

do increment the count of all candidates in Ck+1 that are contained in t

Lk+1 = candidates in Ck+1 with minimum support

end

return ∪k Lk;

VII.II. Count Distribution Algorithm

This algorithm is the parallel version of the sequential algorithm Apriori. This algorithm partitions distributed horizontally and equitably the database on all processors. This algorithm is detailed as follows.

At each step, each processor Pp generates independently support candidates by accessing its local database Dp. It transmits the following the local support of the candidates bya broadcast to other processor that calculate the global support of item set. Once all global frequent item sets L(k) has been determined, each processor construct all candidates C(k+1) in parallel with next step. This process repeated until all frequent item sets are found.

According to Dunham the algorithm is shown below.

Input: I // itemsets

p1, p2, …pp //processors

D = D1, D2, …Dp //database divided into partitions

s //support

Output: L //large itemsets

Count distribution algorithm:

At each processor p1; //perform count in parallel

k = 0; // It is used as the scan number

L = ∅; C1 = I; //initial candidates are set to be the items

repeat k = k + 1; Lk = ∅;

for each Ii ∈Ck

do c 1 i = 0; //initial counts for each itemset are 0

for each tj∈ D1

do for each Ii ∈Ck

do if Ii ∈ti

then c 1 i = c1 i +1;

broadcast c1 i to all other processors;

for each Ii ∈Ck do //determine global counts

ci = ∑p l=1 = c1 i;

for each Ii ∈Ck

do if ci ≥ (s × | D1 ∪ D2 ∪ … ∪Dp |)

then Lk = Lk ∪ Ii; L = L ∪ Lk;

Ck+1 = Apriori-Gen (Lk)

until Ck+1 = Ø.

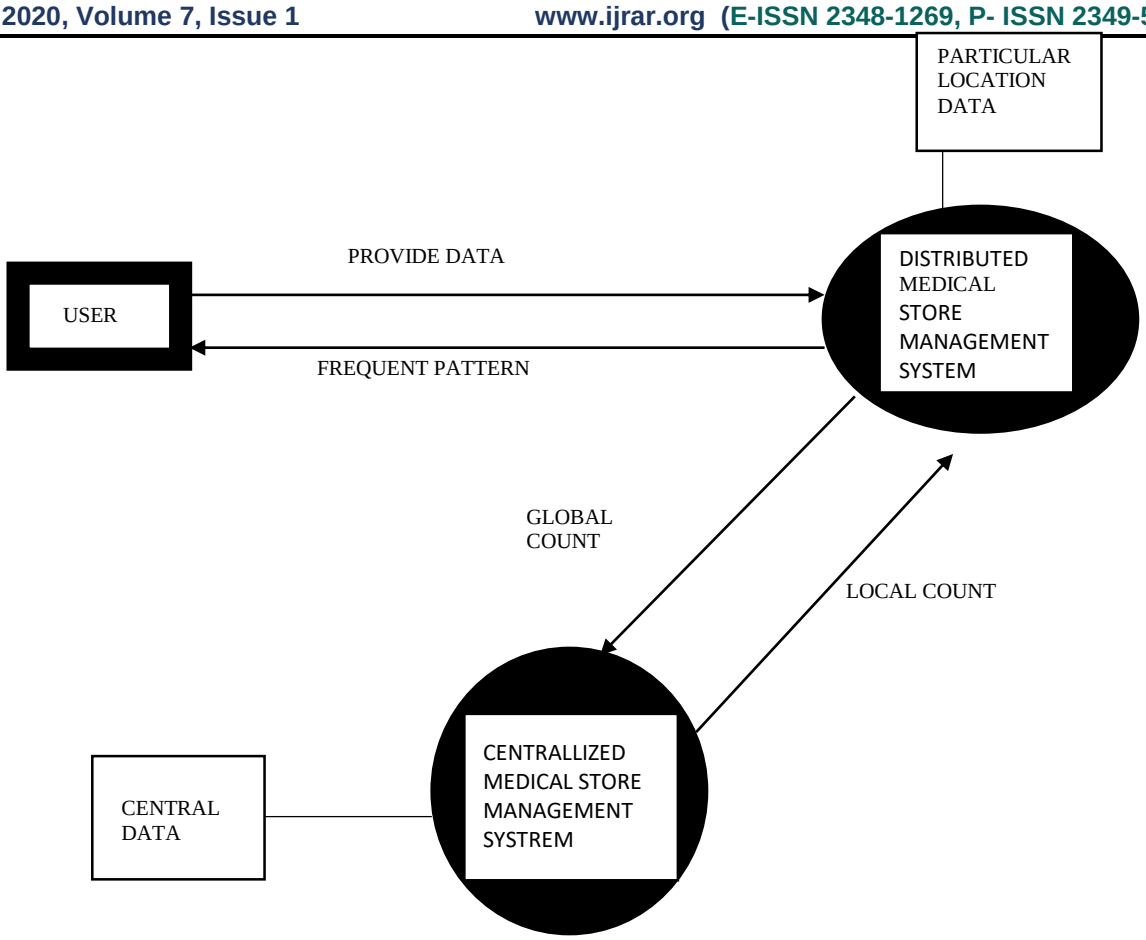

FIG: SYSTEM ARCHITECTURE

VIII. CONCLUSION

This paper can give a detailed view of the application based on ARM techniques. This application shows two operators: A customer and an owner.

The customer is responsible for ordering the medicines online through this application. The store owner can analyse the details of the stock and plan how to maximize their profit. This application also considers the geographical location as a parameter.

REFERENCES

[1] Jyoti Arora and Sanjeev Rao, "An Efficient ARM Technique for Information Retrieval in Data Mining".

[2] Gurneet Kaur, "Association Rule Mining: A Survey" International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (2), 2014, 2320-2324.

[3] Priyanka Yadav, Pavan Wangade, Manish Thakur, Mohammed Fakih, Gayatri Hegde, "PROPOSED DISTRIBUTED DATA MINING IN CREDIT CARD FRAUD DETECTION" IRJET, ISSN: 2395-0056 vol. 3, April 2016.

[4] Yasemin Atilgan and Firat Dogan, "Data Mining on Distributed Medical Databases: Recent Trends and Future Directions" Dogus University, Computer Engineering Department, Research Assistant Acibadem, Istanbul, Turkey