Distributed Apriori for Malls

Devare Jayvant¹, Minal Hadawale², Anuja Gajare³, Mukta Gajare⁴, Sonal Auti⁵
¹Asst.Professor, ²³⁴⁵Student

Abstract—Data mining is used for knowledge discovery in database i.e. data is mined from database and it generates information data which is used by humans effectively and efficiently. Data mining has area known as Association rule mining, which is based on pruning candidate keys. An Apriori algorithm is commonly used Association Rule mining. Apriori algorithm was already been used for single shopping mall. This paper introduces a new way to use Apriori algorithm for distributed shopping mall system. The purpose of this system is to increase the profits of mall administration and develop a customer relationship.

Keywords—Data mining, Association Rule mining, Distributed System, Apriori Algorithm, Candidate Key.

I. INTRODUCTION

As we are familiar with malls, where all customer can buy various products at the same place. In already done research, An Apriori algorithm was used in single malls for generating Association Rule for various products and then categorizing the products based on frequent item set. In this paper we will be implementing an Apriori algorithm for distributed malls.

A. Data Mining:

Data mining is process of discovering the information from database and uses it further. The information is nothing but the patterns present in the database. The data mining is completely based on the patterns present in database. The main purpose of data mining is to extract the informative data from database and convert it into understandable structure for further use.

Data Mining contains following steps:

i) Data cleaning:
   It is the preprocessing done on database, so that the data entry which does not generate the appropriate pattern i.e. inconsistent data is removed.

ii) Data integration:
   It is process in which various data sources can be combined.

iii) Data selection:
   It is the process of selecting relevant data from database and using it.

iv) Data transformation:
   Data is converted in to forms which will be used for data mining by performing aggregation operation.

B. Association rule Mining:

Basic objective of finding association rule is to find all co-occurrence relationship called association. Since it was first introduced in 1993 by Agarwal et. , it has attracted a great deal of attention. The classic application of association rule mining is market basket data analysis, which aims to discover how items purchased by customer in a supermarket (or store) are associated.

Support: The support of a rule, X→Y, is the percentage of transaction in T that contains X U Y, and can be seen as an estimate of the probability, Pr(XUY).

\[ \text{Support} = \frac{\text{count}(XUY)}{n} \]

Confidence: The confidence of a rule, X→Y, is the percentage of transaction in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, Pr(Y|X). It is computed as follows.

\[ \text{Confidence} = \frac{\text{count}(XUY)}{\text{count}(X)} \]

C. Distributed System:

A Distributed system is collection of networked computers or autonomous computers that communicate and co-ordinate with each other for performing a particular task. These networked computers can communicate with each other by message passing and share resources of the system, to provide an integrated computing environment.

II. OUR ALGORITHMS

A. Apriori Algorithm

Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support. A large no of association rule mining algorithms have been developed with different mining efficiencies. Any algorithm should find the same set of rules through their computational efficiencies and memory requirements may be different.
The best known mining algorithm is apriori algorithm. The apriori algorithm works in two steps:

i) Generate all frequent itemset:

A frequent itemset is an itemset that has a transaction support above minimum support.

ii) Generate all confident association rules from frequent itemset:

A confident association rule is a rule with confidence above minimum confidence. A Apriori algorithm relies on apriori or downward closure property to generate all frequent itemset.

**Downward Closure Property:**

If an itemset has minimum support, then every non-empty subset of this itemset also has minimum support.

**Algorithm:**

Ck: Candidate item set of size k
Lk : frequent item set of size k
L1 = {frequent items};
for (k = 1; Lk !=null; 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 min_support
end
return U k Lk;
end

**Step 1:** self-joining Lk-1

insert into Ck

select p.item1, p.item2, ..., p.itemk-1, q.itemk-1 from Lk-1 p, Lk-1 q

Where p.item1=q.item1, ..., p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1

**Step 2:** pruning

for all item sets c in Ck do
for all (k-1)-subsets s of c do
if (s is not in Lk-1) then delete c from Ck

**JOIN Step:**

To Find Lk set, we have to join Lk-1 with itself to generate Candidate k-item sets. The items are sorted into lexicographic order in items set i.e.li[1]<li[2]<....<li[k-1], Members l1 and l2 can be joined if (l1[1]=l2[1])^.........^(l1[k-1]<l2[k-1]). The condition l1[k-1]<l2[k-1] ensure that there is no duplication.

**PRUNE Step:**

Firstly, to get frequent 1-itemset, scan database once. From length k frequent itemsets generate length (k+1) candidate item set. Compare candidate and database. Stop when none of the frequent or candidate set can be generated.

### III. APPROACHES

In this section, the main approaches used in our system are presented. Apriori algorithm is adopted to find the frequent item sets, then in order to get the global support and confidence without privacy disclosure, secure computation is used. For the knowledge hiding, an improved algorithm is mentioned to get satisfy result.

**A. Secure computation**

The basic idea of [2] Secure Multiparty Computation(SMC) is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. Security sum is a very simple and useful method which base on SMC. It is used for get the sum of data from the different site.

**B. Association rule hiding**

There are two modification schemes that incorporate unknowns and aim at the [3][4] hiding of predictive association rules, i.e., rules containing the sensitive items on their LHS(left-hand-side).
The algorithms presented require reduced number of database scans and exhibit an efficient pruning strategy. However, by construction, they are assigned the task of hiding all the rules containing the sensitive items on their LHS, while the algorithms in the work of can hide any specific rule. The first strategy, called ISL (Increase Support of LHS), decreases the confidence of a rule by increasing the support of the item sets in its LHS.

The second approach, called DSR, reduces the confidence of the rule by decreasing the support of the item sets in its RHS (left-hand-side). Both algorithms experience the item ordering effect under which, based on the order that the sensitive items are hidden, the produced sanitized databases are different.

### IV. IMPLEMENTATION

Consider one system having various branches which are located geographically away from each other. Each branch will send their daily sales report; i.e. report of details of transaction that contains details about products purchased by customer on the daily basis by the branch, to the central branch. Central branch will apply apriori algorithm on the received daily sales report. Apriori algorithm generates categories of product according to association rules generated.

Central branch generates the output file for each branch on its daily sales report. The output file is sent to the each branch accordingly.

The output file contains the various categories of products. The employees will arrange products in the mall according to received Output file.

<table>
<thead>
<tr>
<th>Item</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread</td>
<td>4</td>
</tr>
<tr>
<td>Cookies</td>
<td>2</td>
</tr>
<tr>
<td>Milk</td>
<td>4</td>
</tr>
<tr>
<td>Juice</td>
<td>3</td>
</tr>
<tr>
<td>Sugar</td>
<td>4</td>
</tr>
<tr>
<td>Coffee</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item set</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Bread, Milk}</td>
<td>3</td>
</tr>
<tr>
<td>{Bread, Juice}</td>
<td>2</td>
</tr>
<tr>
<td>{Bread, sugar }</td>
<td>3</td>
</tr>
<tr>
<td>{Milk, Juice}</td>
<td>2</td>
</tr>
<tr>
<td>{Milk, Sugar }</td>
<td>3</td>
</tr>
<tr>
<td>{Juice, Sugar }</td>
<td>3</td>
</tr>
<tr>
<td>{Bread, Milk, Sugar }</td>
<td>3</td>
</tr>
</tbody>
</table>

**Minimum Support: 3**

**Support (A, B)=(no. of occurrences of A U B in transaction /no. of transactions)*100**

**Figure: Example of Apriori.**
V. CONCLUSION

The main approach of privacy preservation when doing association rule mining, construction of a system for data mining, by using Secure computation and TEA encryption technology is carried out. It avoids data leakage which cause by data sharing. The knowledge hiding, using ISL achieve sensitive rules hiding, and present an optimization method to get a better result.

REFERENCES


