Product Recommendation System using Associative Rule Mining

Dhrumil Makwana
University of Colorado Boulder
Boulder Colorado USA
dhrumil.makwana@colorado.edu

Tejas Kaushik
University of Colorado Boulder
Boulder Colorado USA
tejas.kaushik@colorado.edu

Abhinav Venkatesh
University of Colorado Boulder
Boulder Colorado USA
abhinav.venkatesh@colorado.edu

PROBLEM STATEMENT

Through this project we intend to identify customer purchasing patterns by analyzing the items purchased in sequence to determine cross sell. For this project we will consider data from multiple platform store to get relevant interesting patterns for that store platform

By Identifying this interesting pattern, we can increase the sales by optimizing the product placement, offer special deals and creating product bundle to encourage further sales of these combination. Popularly used in Amazon, Walmart, Target, and many more

KEYWORDS

Decision making, data mining, association rule, market basket analysis

1 INTRODUCTION

The process of obtaining information from vast amounts of data is known as data mining. Even a phrase as simple as "knowledge mining" could not fully capture the significance of sifting through a lot of data. Mining is the process of extracting a few precise nuggets from a large quantity of raw material. Many other terms, including knowledge mining from databases, knowledge extraction, data/pattern analysis, and data archaeology, have meanings that are somewhat like or dissimilar from data mining.

Simply put, knowledge discovery from databases requires data mining as a necessary step. The steps involved in knowledge discovery as a process are listed below in an iterative sequence:

I) Data cleaning (to eliminate ambiguous and erratic data).

II) Data integration is the process of combining data from various sources.

iii) Data Selection (where data from the database that are pertinent to the analysis task are retrieved).

iv) Data transformation (where data are transformed or consolidated into forms suitable for mining, for example, using summarization or aggregation operations.

v) Data mining (a crucial procedure that employs clever techniques to extract data patterns).

vi) Pattern evolution (to find the patterns that best represent knowledge based on some interesting measures).

vii) Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

For discovering associations between a sizable collection of data elements, association rule mining (ARM) is used. Numerous industries are getting concerned about mining association rules from their databases because to the vast amount of data contained in databases. Examples include catalog design, cross-marketing, and other corporate decision-making processes that can benefit from the discovery of intriguing association patterns between vast quantities of business transaction data. Market basket analysis is a classic illustration of association rule mining.

By finding relationships between different things that customers put in their shopping baskets, this approach analyzes the purchasing behaviors of consumers. By learning which products customers regularly buy with one another, businesses can use the identification of such links to broaden their marketing strategy.

Market basket analysis (MBA) is a data mining method for identifying relationships across datasets. These associations may be represented by association

rules. The official formulation of the problem is as follows: Let I be a collection of things (i1, i2, ..., i nth). Let C be a collection of transactions that results in R . An identifier called RID is used to specifically identify each transaction.

The approach can be expressed as follows: if there are subsets of product items, A and B, then an association rule is in the shape of A*B. It indicates that if a consumer buys A, he or she will also buy B. Support and confidence are two metrics that show how certain newly found association rules. Confidence measures the accuracy of rules.

For Example, if a customer buys diapers they will also tend to buy beer at the same time is represented in association rule.
Diaper - Beer Support = 25% and confidence = 90%

If association rules meet both a type of equation minimum support criterion and a minimum confidence threshold that can be specified by users or domain consultants, they are deemed to be useful. Figure 1 displays an example of a market basket analysis. This is the ideal illustration of association rule mining. This market basket analysis system will assist managers in understanding the groups of goods that customers are most likely to buy. The consumer transaction data from all retail shops may be used in this analysis. Their planning for marketing or advertising strategies will be guided by these findings. Managers can recommend novel arrangements for store layouts, for instance, with the aid of market basket analysis. Things that are frequently bought together can be put close together based on this data to further encourage the selling of such items as a set.

2 LITERATURE SURVEY

The application of market basket analysis can be traced back to [1] where the problem of mining association rules was first explored. With a large dataset of customer transactions wherein each transaction contained items purchased by a customer in one visit to the store, they were able to present an algorithm to identify association rules between items in the dataset. The goal was to identify rules such as "90 percent of transactions that purchase bread and butter also purchase milk." They presented the importance of finding rules with minimum support and confidence. This was further expanded upon by [2] where Apriori algorithm was formally presented. This was done to help provide a fast algorithm to identify relevant rules by taking decisions based on prior information which was already available. The application of market basket analysis can be applied to the real world as showcased by [3]. In this paper, the customer behavior is analyzed to help companies gain a competitive edge by arranging their products in the supermarket in a way that would lead to an increase in their profits. Apart from the application as presented by [3], data mining techniques also have a variety of applications in different domains. [4] presents an excellent analysis of the application of such techniques in the fields of marketing, web analysis and finance. The effectiveness of different algorithms for associative rule mining is compared in [5] where matrix Apriori and FP growth is compared in terms of performance. This case study compares the performance on datasets having different characteristics and tries to identify the underlying causes for difference in performance. They found matrix Apriori to perform better than FP-Growth in terms of performance for threshold values below 10% but building the matrix imposed a higher cost which would later help by finding the itemsets at a faster rate.

3 PROPOSED WORKS

The Data Cleaning / Preprocessing

Remove irrelevant Data like Country, Quantity, Customer ID, Price, Stock Code etc. These parameters won't affect the result of the knowledge extracted after data mining. In case of Country the dataset is skewed towards 1 country and there are about 32k outlier which would also be removed to get higher accuracy. There will also be missing data or unspecified data which will be removed as after removing irrelevant data the dataset will have Invoice Number/Bill Number, Item Name, Date in which missing cells cannot be replaced with average. As in this project we have 2 dataset which will be merged there might be some redundant data which has to be

handled. Stock Code is removed as in this dataset the basket is created using Bill Number and item names are listed in description, also in 2nd dataset there is not parameter which is like Stock Code.

Data Transformation / Integration

The dataset contains rows of items bought under common Invoice Number or Bill Number. This fact could be used to build the basket. In Integration, there are 2 Dataset with different parameter name but similar data format which will be handled before merge like Item name and description are similar parameters for different dataset.

Design

In this project, Apriori Algorithm and ARM will be used for and working principle is given below.

Working Principle

Step 1: Accept the minimum support as minsup and minimum confidence as minconf.

Step 2: Determine the support count for all the item as s.

step 3: Select the frequent items. (Items with s >= minsup)

step 4: The set candidate k-items is generated by 1-extension of the large (k-1) itemsets generated in step3.

step 5: Support for the candidate k-itemsets is generated by a pass over the database.

step 6: Itemset that do not have minsup are discarded and the remaining itemsets are called large k-itemsets

step 7: The process is repeated until no larger item.

step 8: The interesting rules are determined based on the minimum confidence.

Pseudo Code apriori

Join Step: To generate Ck join Lk-1 with itself.

Prune Step: Any (k-1) itemset which is not frequent cannot be a subset of frequent k-1 itemset.

k-itemset – An itemset with k items

L_k – Set of large itemsets having k items. Every member Of this set has two parts:-
1.Itemset 2. Support count

C_k – Set of candidate itemsets having k items.Every Member of this set has two parts:-
1.Itemset 2.Support count

1. L1={large 1-itemsets};
2. for (k = 2; Lk-1 ≠ Ø; k++) do begin
3. Ck = apriori-gen(Lk-1);
4. For all transactions t ϵ D do begin
5. Ct = subset(Ck, t);
6. for all candidates c ϵ Ct do
7. c.count++;
8. end
9. Lk ={c ϵ Ck | c.count ≥ minsup}
10. end
11. Answer = Uk Lk;

4 DATA SET

Market basket analysis

URL:
https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis

Market Basket Dataset contains transaction from 2010 to 2011. In this dataset we will be considering Invoice Number, Item Name and Date

Online retail UCI

URL:
https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

This data set contains all the transactions occurring

for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers. The important attributes used for market basket analysis include the invoice number i.e., the unique 6-digit integral number uniquely assigned to each transaction, description i.e., product name and Invoice date and time.

5 EVALUATION METHODS

In this evaluation scheme we split the dataset with 80:20 ratio into training and testing dataset. Then we compare the output association rules with the association rules generated from the test dataset to compute precision, recall and accuracy of the model. We also compute lift, support and confidence of every association rule generated as output. We also vary the number of output association rules and compare it with coverage over the dataset i.e., percentage of items in the dataset that model can recommend and personalization i.e., negative cosine similarity between itemsets.

6 MILESTONES

1). Data Cleaning and Preprocessing. Date :- 07/21/22
2). Data Integration. Date:- 07/22/22
3). Data Transformation. Date:- 07/27/22
4). ARM. Date:- 08/06/22

6.1 MILESTONES COMPLETED

1). Data Cleaning and Preprocessing :- In this part, the dataset is cleaned with respect to rows with outliers, Any relevant cell with Nan and all the irrelevant columns.

2). Data Transformation:- In this part, the data is modified to be feed into the analysis process. For this we combined the items which were under 1 common Bill Number or Invoice Number. The Item Names are in 1 string comma separated. The total dataset after merging the relevant rows in 1 common are 63k data entries.

6.2 MILESTONES TODO

1). Data Integration:- We have 2 datasets, and we are yet to safely combine both datasets.

2). ARM:- Associative rule mining, this step will be started once the Integration of datasets is done.

7 TOOLS

1). Python
2). Pandas
3). NumPy
4). Matplotlib
5). Sci-kit Learn

8 REFERENCES

[1] R Agrawal, T Imielinski, A Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D C 1993 pp 207 216

[2] R Agrawal, R Srikant, Fast algorithms for mining association rules, Proceedings of the 20 th VLDB Conference, Santiago, Chile, 1994 pp 478 499

[3] Raorane AA, Kulkarni RV, Jitkar BD Association Rule Extracting Knowledge Using Market Basket Analysis Research Journal of Recent Sciences 2012 1 2 19 27

[4] I Bose, R K Mahapatra, Business data mining a machine learning perspective, Information and Management 39 2001 211 225

[5] Yıldız B. and Ergenç B., (Turkey) in Comparison of Two Association Rule Mining Algorithms without Candidate Generation, International Journal of Computing, and ICT Research, 674(131), 450-457 (2010)

9 RESULTS SO FAR

1). Data Cleaning and preprocessing:

We divided data cleaning in 3 parts.

Step 1:- Removing rows which does not have United Kingdom in country column(As the dataset is skewed towards UK).

Results of Step 1:-

- Before this step the dataset had [522064 rows x 7 columns] (dataset 1), [1067371 rows x 8 columns](dataset 2)
- After performing this step, the result has [487622 rows x 7 columns](dataset 1), [981330 rows x 8 columns](dataset 2).
- After this step combine 120483 rows were removed.

Step 2:- Removing any rows in which relevant parameters have Nan or empty cells.

Results of Step 2:-

- Before this step the dataset had [487622 rows x 7 columns] (dataset 1), [981330 rows x 8 columns](dataset 2)
- After performing this step, the result has [486167 rows x 7 columns]dataset 1), [976948 rows x 8 columns] (dataset 2).
- After this step combine 5837 rows were removed.

Step 3:- Removing the irrelevant columns and reducing the dimension of the dataset.

Results of Step 3:-

- Before this step the dataset had 7 columns in dataset 1 and 8 columns in dataset 2.
- After this step the dataset has 3 columns in both datasets.

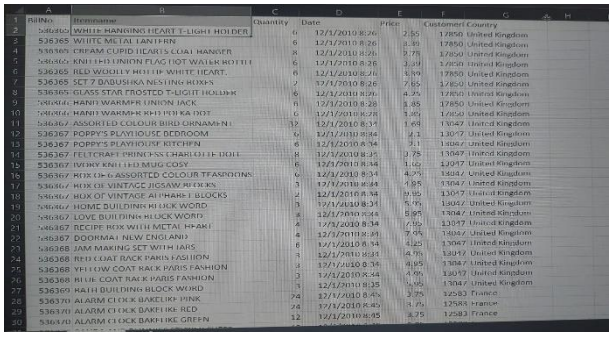

Fig 1.1 :- Dataset 1 Before

Fig 1.2 :- Dataset 1 After

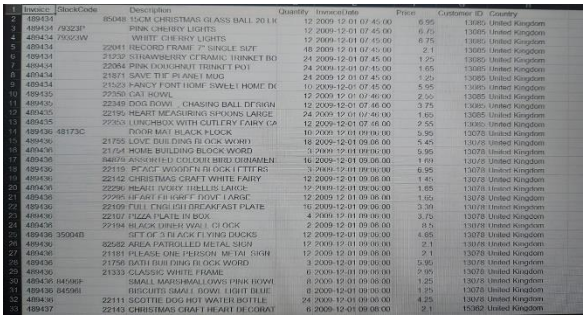

Fig 2.1 :- Dataset 2 Before

Fig 2.2 :- Dataset 2 After

2). Data Transformation:-

In data transformation we are merging the rows which have same Bill Number or Invoice Number in each dataset. The Items are divided and kept in separate rows and after transformation the Item names are merged under 1 string comma separated.

In the resulting dataset the bill number for dataset 1 and Invoice number in dataset 2 are static and the Item name, description and transaction date is merged under 1 string and comma separated.

The result has 63k data entries combined after data transformation.

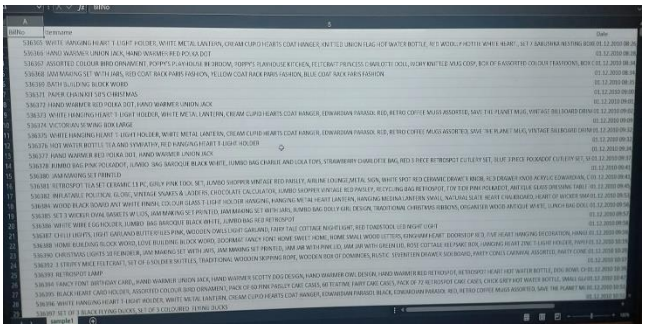

Fig 3 :- Dataset 1 After Transformation

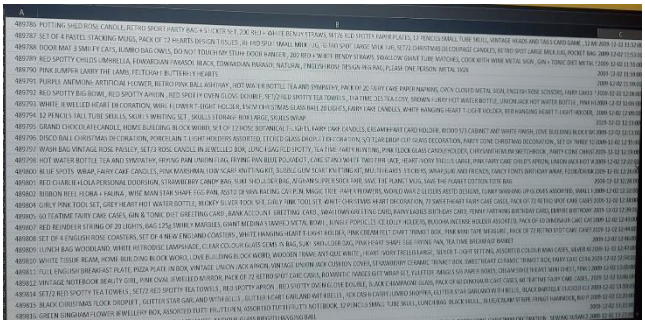

Fig 4 :- Dataset 2 After Transformation