MARKET BASKET ANALYSIS FOR A SUPERMARKET

Conference Paper · November 2022

CITATIONS
0

READS
2,938

3 authors, including:

Cecil Ignatius Hermina
Bannari Amman Institute of Technology
4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

MARKET BASKET ANALYSIS FOR A SUPERMARKET

[1]Cecil Ignatius Hermina, [2]Aishwaryalakshmi B & [3]Dr.B.Gopalakrishnan

[1]Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Tamilnadu. Email: cecilignatiushermina.ec19@bitsathy.ac.in
[2]Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Tamilnadu. Email: aishwaryalakshmi.ec19@bitsarhy.ac.in
[3]Department of Information Technology, Bannari Amman Institute of Technology, Tamilnadu. Email: gopalakrishnanb@bitsathy.ac.in

ABSTRACT:

The Market Basket Analysis (MBA) method of data mining looks for a collection of items that frequently occur together in a large dataset or database. This technology is used in various industries like retail to promote cross selling and to help in product placement and in fraud detection and other uses. Understanding consumer purchasing trends and preferences has been made easier with the use of this technology. The use of it in international corporations is widespread. Along with the advancement of technology, current business trends have undergone a significant transformation. This technique analyses the purchasing habits of customers by determining the relationships that exist between the various items that are placed in the customers' shopping baskets. Because of shifts in the expectations of customers, it is essential for businesses to improve the precision of their operations. In this particular scenario, we examine a neighbourhood grocery store to investigate and contrast the Apriori and FP Growth algorithms' respective run times. In this paper we also find the combination of products that are sold together.

Keywords: *FP Growth, Apriori, Market basket Analysis.*

1. INTRODUCTION

Retail business is one of the industries with high level competition. The growth of the business is decided by the response speed and the ability to understand customers' behaviour [1]. The main challenge for this industry is customer data collection. Market basket analysis is an acquisition of analytical techniques focused to find relations and relations among products based on the customer's purchase [2]. Market basket analysis helps to analyse the probability of a customer purchasing different products together. Analysing purchase patterns can help improve marketing methods to improve sales and increase profits. Most of the unused data are stored in the archive to be used as sales report [3].

In the retail industry, the utilisation of data mining techniques is essential to the accomplishment of company goals. Data mining techniques help uncover patterns of consumer spending by disentangling the connections and associations that exist between various products [4]. Association rules are applied in order to determine frequent item sets based on support and confidence thresholds that are defined at different levels. The term "frequent itemset" refers to a group of items with minimal support [5]. The number of transactions for an itemset within the data set is known as support. The degree of each revealed pattern's certainty serves as a measure of confidence.

2. LITERATURE SURVEY

Data mining is the process of discovering useful information hidden within large amounts of data. The association between different goods in a dataset can be discovered using the data mining technique known as market basket analysis [6]. It is widely used to mine transactions or basket data, especially in retail. Using this technology has made it simpler to understand consumer buying trends and preferences [7]. Market basket analysis aids in selecting discounts and sales promotion strategies for distinct consumer segments [8]. Factors like confidence, support, and lift are employed to determine the association or relationship between the items. Support displays the frequency with which a given itemset appears in all transactions. The degree of confidence indicates the possibility that if item A is purchased, item B will follow [9-10]. The lift between the antecedent and consequent is the correlation. ECLAT algorithm is a more efficient type pf Apriori algorithm. This is because ECLAT algorithm works vertically while Apriori works horizontally and this makes ECLAT faster [11]. FP Growth does not need creating candidate sets and therefore FP Growth is 5 times faster.

2.1. Association Rule Mining

One of the most used data mining approaches is the association rule. Finding relationships between the objects in a data set is helpful. This is accomplished by the analysis of recurring patterns and numerous measures, including support, confidence, life, leverage, and conviction. Support provides a summary of the frequency of a group of things that appear in all transactions. We are supported by the percentage of all transactions in which the itemset appears [12].

$$\text{supp}(X{-}{>}Y) = \frac{(\text{Transactions containing both X and Y})}{(\text{Total No. of transactions})}$$

An item set is referred to as a "frequent itemset" when the support is greater than the required minimum support threshold. Low support rule is disregarded when seen from a business standpoint.
Confidence is the probability that when item X is purchased, item Y will follow.

$$\text{confidence}(X{-}{>}Y) = \frac{\text{support}(X{-}{>}Y)}{\text{support}(X)}$$

When X is present in the shopping cart, the likelihood that Y is also present increases in comparison to when X is absent. This can be determined through the use of the lift.

$$\text{Lift}(X{-}{>}Y) = \text{confidence}(X{-}{>}Y)/\ \text{support}(X)$$

Given that A and C are not connected in any way, the disparity in the observed and expected frequencies of X and Y occurring together is referred to as the "leverage," and it is characterised by the term "leverage."

$$\text{Leverage}(X{-}{>}Y) = \text{support}(X{-}{>}Y) - \text{support}(X){*}\text{support}(Y)$$

Conviction can be defined as the frequency with which X takes place without Y divided by the frequency with which the rule makes an incorrect prediction (i.e. the frequency with which the rule makes an incorrect prediction). If X and Y were truly independent of one another, the ratio of the frequency of incorrect predictions to the probability of X occurring without Y should be equal to one (1).

$$\text{Conviction}(X \rightarrow Y) = \text{support}(Y)/\text{confidence}(X \rightarrow Y)$$

2.2. Apriori Algorithm

Apriori algorithm is utilized for mining regular item sets in a data set. This is a classical technique in a data mining. This algorithm is formulated to work on data sets containing large number of exchanges/ transactions. Apriori algorithm is a well-known rule to inspect relations among items or variables in a large data set[13].
Apriori algorithm uses 2 principles.
1. All subsets of a frequent itemset must be frequent.
2. For any infrequent itemset, all its supersets must be infrequent too.

2.3. ECLAT Algorithm

Even in this day and age, the ECLAT algorithm is still a common option when considering association rule algorithms (which stands for Equivalence Class Clustering and bottom-up Lattice Traversal). It is more efficient than the Apriori algorithm, which is a direct comparison for this method. By carrying out its operations in a horizontal fashion, the Apriori algorithm is able to simulate the breadth-first search that occurs in a graph. The ECLAT algorithm, on the other hand, takes a vertical approach by simulating the depth-first search that occurs in a graph. Because of this, ECLAT is preferable to Apriori in terms of accuracy while also being superior in terms of speed. You will not obtain any new subsets that are not already present in the prefix tree when you use ECLAT.

In order to accomplish this, it computes the support value by finding the intersections of the transaction sets. The ECLAT is superior to other tests of its kind because, unlike those other tests, it does not require the candidate to search the database for evidence but rather uses k+1 item sets.

As it uses a Depth-First Search methodology, ECLAT is significantly more efficient than the Apriori algorithm and needs a great deal less memory to function. A cursory examination of the information is all that is required to carry out computations while being provided with individualised assistance.

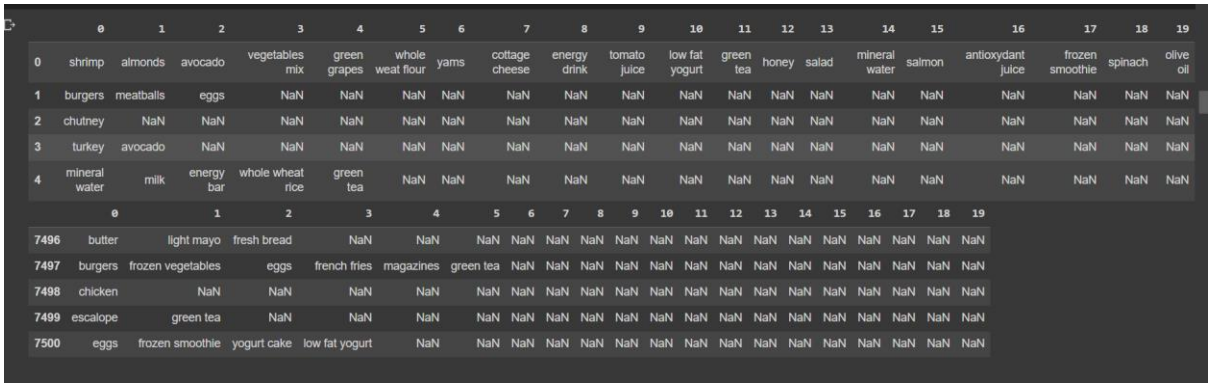
2.4. FP Growth

FP Growth is an improvement that can be made to the Apriori method. The database is portrayed by FP Growth in the form of a tree known as an FP tree, which is also known as a Frequent Pattern Tree. The relationship between the sets of items is maintained by the structured tree. The structure of the tree is determined by the initial item sets contained within the database. Each node in the FP tree, which is used to mine the pattern that occurs the most frequently, is a representation of one of the item sets. The nodes further down the tree represent the itemsets, while the node at the very top of the tree represents null. The connection that exists between the tree's root node and its child nodes (itemsets) and their respective neighbours is maintained while the tree is being constructed.

3. RESEARCH METHODS

Association rules are found to be functional if minimum support threshold and minimum confidence threshold meet the threshold set by the user or consultant. Market basket analysis rules can be written as if **{A} then {B} i.e. {A} => {B}**.

The store data that we have collected is of the size 7501 x 20. Each transaction in the dataset is the combination of products bought together by a customer. Figure 1 shows the head and tail of the dataset.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole weat flour	yams	cottage cheese	energy drink	tomato juice	low fat yogurt	green tea	honey	salad	mineral water	salmon	antioxydant juice	frozen smoothie	spinach	olive oil
1	burgers	meatballs	eggs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	chutney	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	turkey	avocado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	mineral water	milk	energy bar	whole wheat rice	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
7496	butter	light mayo	fresh bread	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7497	burgers	frozen vegetables	eggs	french fries	magazines	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7498	chicken	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7499	escalope	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7500	eggs	frozen smoothie	yogurt cake	low fat yogurt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1. Head and tail of the dataset

This data needs a lot of pre-processing. TransactionEncoder() of mlxtend. preprocessing does the pre-processing for us. TransactionEncoder() helps find the different products in transaction and give each transaction a list that contains a binary array where True represents a purchased product. The pre-processed data is visualised as bar charts and each column represent distinct item. Figure 2 shows sales of unique items.

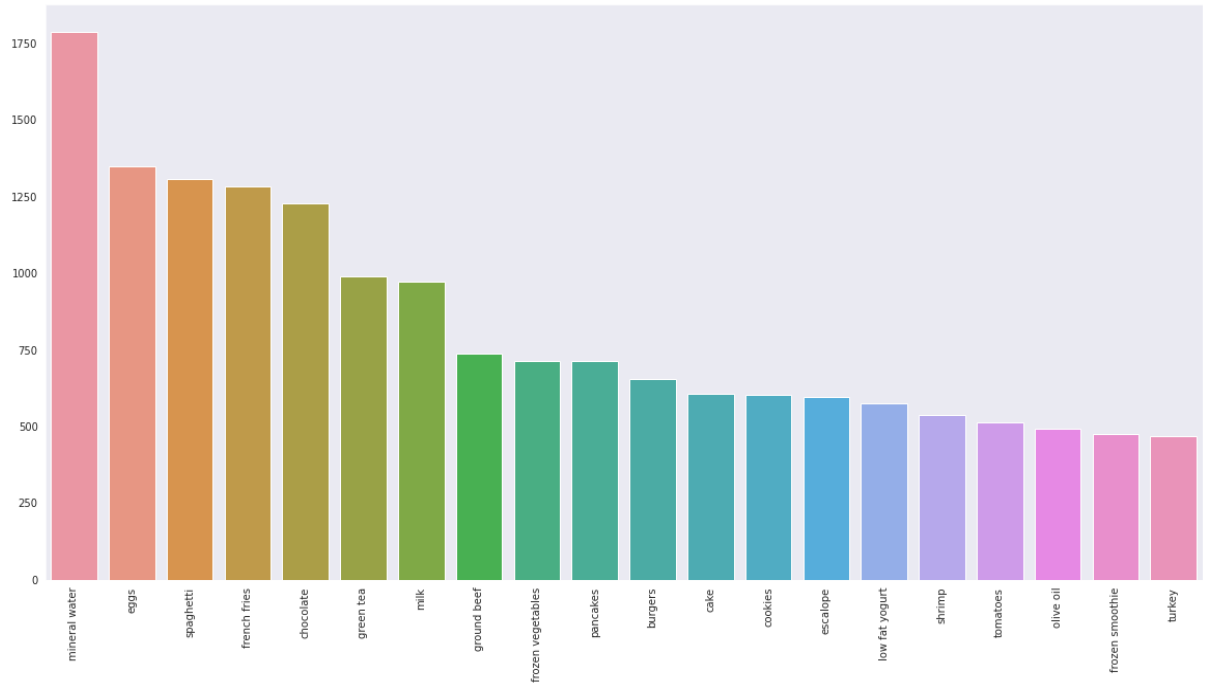


Figure 2. Sales of unique items

The most popular item in the store, as shown in figure 2 above, is mineral water. Rule generation involves two steps in the process. The first rule is to create a list of repeating items, and the second is to create a list of rules that are acceptable given the observed items. One approach is to look at all potential subsets of the item set under consideration, look at the item sets' support values, and then only consider the item sets with values higher than the minimal threshold support value. The apriori algorithm is a support measure that opposes monotonicity. As a result, the search space is smaller, which speeds up the construction of repeating item sets. Apriori cuts back the supersets of an itemset that does not meet the condition of minimum threshold value.

'Bottom up' approach is the process of expanding recurrent itemset individually.

1. Create itemset that meets the minimum threshold that has only one item L_1. Use self-join and create all workable combinations of L_1 and the outcome is L_2.
2. At each step we create candidate item sets and for every candidate we examine the entire database to identify support and separate candidates that do not meet the minimum threshold value.
3. In the same way we create L_k from L_{k-1} until self-join is not applicable.

The number of association rules that can be used if n elements are in a set is 3<supn- 2n+1 + 1. To generate all these rules requires relatively large number of steps. Apriori breaks down this process. From all the workable rules we recognize the ones they are above the minimum confidence level. Confidence of rules from the same itemset obey the anti-monotone property.

Conf (A, B, C -> D) >= Conf (B, C-> A, D) >= Conf (C-> A, B, D)

Apriori requires numerous candidate item sets creation to examine the support of every itemset created and this is computationally expensive. This limitation can be overcome by FP Growth which is an advancement of Apriori algorithm. A pattern is created without candidate generation. FP tree is a database representation in form of a tree. The association between item sets is preserved in the tree. The database is unusable due to one common item. The term "pattern fragment" refers to the broken or fractured portion. In contrast to the Apriori algorithm, the itemsets of the fragmented patterns are examined, which decreases the temporal complexity.

The FP tree is built using the itemsets that are located at the very beginning of the database. Every node in the FP tree represents a different itemset in some way. The value null is represented at the root node, while the itemsets are represented at the child nodes in the hierarchy. During the process of tree formation, the connection between the nodes is maintained.

Steps in FP Growth to mine frequent pattern.

1. Scan the database as in the Apriori.
2. To build the FP tree from the tree's root. Root is represented by Null.
3. Scan the database to study trransactions. The itemset with maximum transactions is placed at the top. The item sets are arranged in descneding order of number of transactions.
4. Common itemsets are linked to the nodes of another itemset.
5. The count of the common node and new node is increased by 1 as the nodes are constructed and joined.

6. node is analysed frist along the links. The travesed path is called conditional pattern base.
7. Tree construction for conditional FP. The conditional FP tree is made up of the itemsets that satisfy the threshold support..
8. FP with conditions Tree produces regular patterns.

Creation of candidate sets is not required by FP Growth and that is why FP Growth is faster than Apriori algorithm.

4. RESULTS AND DISCUSSION

From Figure 2 we can conclude that mineral water must be in stock at all times.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(chocolate)	(mineral water)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357
1	(mineral water)	(chocolate)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256
2	(spaghetti)	(mineral water)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314
3	(mineral water)	(spaghetti)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008

Figure 3. Apriori algorithm with lift greater than 1.3

Figure 3 shows the results of Apriori algorithm for which the lift is greater than 1.3. From this we can conclude that 22% of transactions containing mineral water contains chocolate.

32% of transactions containing chocolate contains mineral water. From the comparison of lift, leverage and conviction of {spaghetti and mineral water} and {chocolate and mineral water} we observe that the chances of transaction of {spaghetti and mineral water} is more than {chocolate and mineral water}.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(eggs)	(mineral water)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815
1	(mineral water)	(eggs)	0.238368	0.179709	0.050927	0.213647	1.188845	0.008090	1.043158
2	(spaghetti)	(mineral water)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314
3	(mineral water)	(spaghetti)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008
4	(chocolate)	(mineral water)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357
5	(mineral water)	(chocolate)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256

Figure 4. FP Growth algorithm with lift greater than 1.3

Figure 4 shows the results of FP Growth for which the lift is greater than 1.3. From the result we observe that spaghetti and mineral water are most likely to occur together.

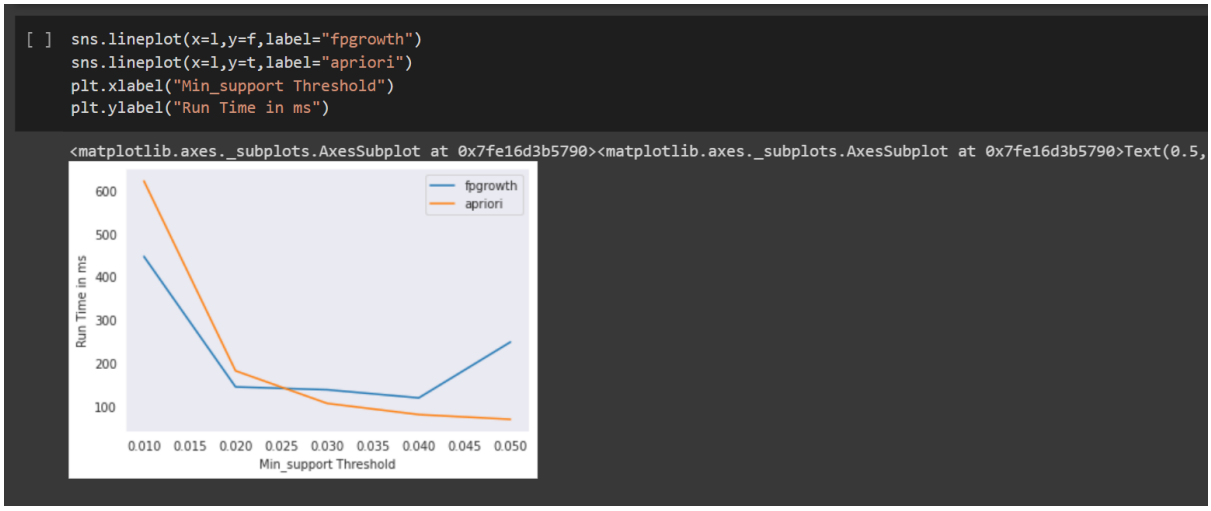


Figure 5. Run time comparison – Apriori and FP Growth

From Figure 5 we get the insights about the run time comparison between Apriori and FP Growth. Apriori algorithm requires creation of candidate sets, therefore it is much slower than FP Growth. FP Growth is five times faster.

5. CONCLUSION

Based on the findings, the FP Growth algorithm can be considered not only more sophisticated than the Apriori algorithm, but also significantly quicker. This data shows that mineral water is the most commonly purchased product. Thus, there must always be a supply of mineral water for purchase.

REFERENCES

1. Raorane, A. A., R. V. Kulkarni, and B. D. Jitkar. "Association rule–extracting knowledge using market basket analysis." Research Journal of Recent Sciences ISSN 2277 (2012): 2502.

2. Zheliznyak, Iryna, Zoriana Rybchak, and Iryna Zavuschak. "Analysis of clustering algorithms." In Advances in Intelligent Systems and Computing, pp. 305-314. Springer, Cham, 2017.

3. Rettinger, Achim, Uta Lösch, Volker Tresp, Claudia d'Amato, and Nicola Fanizzi. "Mining the semantic web." Data Mining and Knowledge Discovery 24, no. 3 (2012): 613-662.

4. Aguinis, Herman, Lura E. Forcum, and Harry Joo. "Using market basket analysis in management research." Journal of Management 39, no. 7 (2013): 1799-1824.

5. Kumar, K. Suresh, T. Ananth Kumar, S. Sundaresan, and V. Kishore Kumar. "Green IoT for Sustainable Growth and Energy Management in Smart Cities." In Handbook of Green Engineering Technologies for Sustainable Smart Cities, pp. 155-172. CRC Press, 2021.

6. Aguinis, Herman, Lura E. Forcum, and Harry Joo. "Using market basket analysis in management research." Journal of Management 39, no. 7 (2013): 1799-1824.

7. Solnet, David, Yasemin Boztug, and Sara Dolnicar. "An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue." International Journal of Hospitality Management 56 (2016): 119-125.

8. Wong, Jehn-Yih, Huei-Ju Chen, Pi-Heng Chung, and Nai-Ching Kao. "Identifying valuable travelers and their next foreign destination by the application of data mining techniques." Asia Pacific Journal of Tourism Research 11, no. 4 (2006): 355-373.

9. Miguéis, Vera L., Ana S. Camanho, and João Falcão e Cunha. "Customer data mining for lifestyle segmentation." Expert Systems with Applications 39, no. 10 (2012): 9359-9366.

10. Usharani, S., P. Manju Bala, T. Ananth Kumar, R. Rajmohan, and M. Pavithra. "Smart Energy Management Techniques in Industries 5.0." Hybrid Intelligent Approaches for Smart Energy: Practical Applications (2022): 225-252.

11. Zhao, Qiankun, and Sourav S. Bhowmick. "Association rule mining: A survey." Nanyang Technological University, Singapore 135 (2003).

12. Birant, Derya. "Data mining using RFM analysis." In Knowledge-oriented applications in data mining. IntechOpen, 2011.

13. Badhe, Vivek, Ramjeevan Singh Thakur, and G. S. Thakur. "Profit Pattern Mining Using Soft Computing for Decision Making: Pattern Mining Using Vague Set and Genetic Algorithm." In Pattern and Data Analysis in Healthcare Settings, pp. 213-239. IGI Global, 2017.