Application of an Apriori algorithm for analyzing consumer purchasing patterns

MORAN BUCHNIK[1]

[1] Department of Management, Bar Ilan University, Ramat Gan, 52900, Israel

Abstract

Wholesale companies collect a large amount of data on their customers but is rarely utilized. Today, there are tools that can enable exploitation, such as data mining for strategic decisions, data market analysis and the use of an Apriori algorithm, which I will expand on in this article.[5]

The purpose of the research is to create a shopping basket that suits the customer and increase the profits of the wholesale company. The transaction data of suppliers provided a lot of information regarding purchases of products that are related, that is, if the customer purchases a certain product, chances are that he will be interested in purchasing another product that will be related to the product he has already purchased.

We discovered that by using the Apriori algorithm, we can give the customer the most attractive shopping baskets and increase the supermarket's sales, advertise campaigns and retain existing customers.

Introduction

Frequent itemset mining leads to the discovery of associations and correlations between items in huge transactional or relational datasets. With vast amounts of data continuously being collected and stored, many industries are becoming interested in mining such kinds of patterns from their databases. The disclosure of “Correlation Relationships” among huge amounts of transaction records can help in many decision-making processes.

A popular example of frequent itemset mining is Market Basket Analysis. This process identifies customer buying habits by finding associations between the different items that customers place in their shopping baskets.

Among the algorithms that were proposed to solve this problem, there is the Apriori algorithm. This algorithm iterates over all possible sets of items up to a certain size and finds correlations between them. It helps to find frequent itemsets in transactions and identifies association rules between these items.

In this paper, I show the strength of Apriori by analyzing it results on the Online Retail II data set [3]. This dataset contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The database contains 4374 customers from all over the world, 3952 from the UK.

Research Methodology

Data Mining

Data mining or often referred to as knowledge discovery in databases (KDD) is a process that includes gathering, using historical data to find order, patterns or relationships in large data.

This data mining output can be used to help make future decisions. The development of KDD has caused the use of pattern recognition to diminish because it has become a part of data mining. [1]

Association rule learning

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. In any given transaction with a variety of items, association rules are meant to discover the rules that determine how or why certain items are connected. [4]

There are two elements of these rules:
Antecedent (IF): This is an item/group of items that are typically found in the Itemsets or Datasets.
Consequent (THEN): This comes along as an item with an Antecedent/group of Antecedents. [8]

In order to select interesting rules from the entire set of possible rules, constraints are used on a variety of significant and interesting measurements. The most well-known constraints are minimal threshold values in support and confidence

Support Combined percentage of the two items: for identifying the combination of the item which is fulfill the minimum requirement of support value. Support value of an item is achieved by using the following formula:

Support = $P(A \cap B) = \frac{number\ of\ transactions\ containing\ A\ and\ B}{total\ number\ of\ transactions}$

Confidence is the percentage of all transactions satisfying X that also satisfy Y.

The confidence value of an association rule, often denoted as X ⇒Y, is the ratio of transactions containing both X and Y to the total amount of X values present, where X is the antecedent and Y is the consequent. [6]

Confidence $(X \Rightarrow Y) = P(X|Y) = \frac{Support\ (X \cap Y)}{Support(X)} = \frac{number\ of\ transactions\ containing\ A\ and\ B}{total\ number\ of\ transactions}$

The lift of a rule is defined as:

Lift $(X \Rightarrow Y) = \frac{Support\ (X \cap Y)}{Support(X) \times Support(Y)}$

If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

Apriori Algorithm

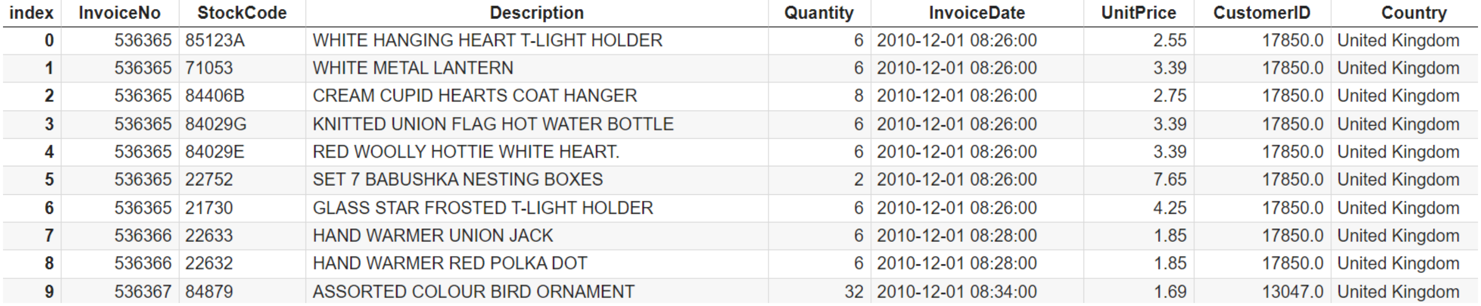
Rakesh Agrawal and Ramakrishnan Srikant originally introduced the method in 1994. The Apriori algorithm identifies the common set of items or elements in a transaction database and sets association rules between the items. The method uses a "bottom-up" strategy, in which frequent subgroups are expanded one item at a time (creating candidates), and groups of candidates are tested innately. When no more successful rules can be obtained from the data, the algorithm stops. [2]

Discussion

The database

High-Frequency Pattern Analysis this stage looks for item combinations that meet the minimum requirements of the support value in the database.

In this study, I will use at Apriori algorithm in Phyton.

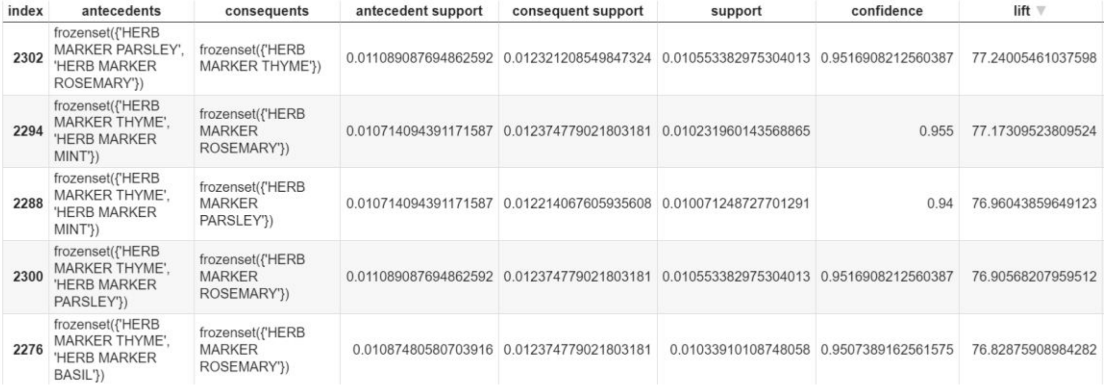


index	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

Filter by support & confidence

From the output below, when the rule has more than 1% support and more than 90% confidence, I can see that the level is very high for spices such as thyme, parsley, rosemary, and thyme.

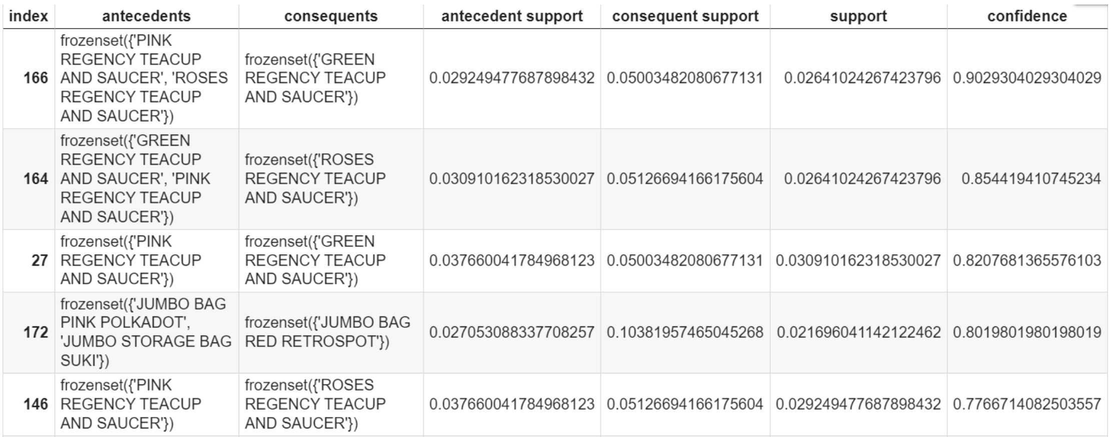
I think the reason for this is that a customer buys spices, he will buy a few different brackets to enhance the flavor. In addition, the British usually make their own spice mixture called Mixed Spice which is probably made with the spices that customers buy at the supermarket



index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift ▼
2302	frozenset({'HERB MARKER PARSLEY', 'HERB MARKER ROSEMARY'})	frozenset({'HERB MARKER THYME'})	0.011089087694862592	0.012321208549847324	0.010553382975304013	0.9516908212560387	77.24005461037598
2294	frozenset({'HERB MARKER THYME', 'HERB MARKER MINT'})	frozenset({'HERB MARKER ROSEMARY'})	0.010714094391171587	0.012374779021803181	0.010231960143568865	0.955	77.17309523809524
2288	frozenset({'HERB MARKER THYME', 'HERB MARKER MINT'})	frozenset({'HERB MARKER PARSLEY'})	0.010714094391171587	0.012214067605935608	0.010071248727701291	0.94	76.96043859649123
2300	frozenset({'HERB MARKER THYME', 'HERB MARKER PARSLEY'})	frozenset({'HERB MARKER ROSEMARY'})	0.011089087694862592	0.012374779021803181	0.010553382975304013	0.9516908212560387	76.90568207959512
2276	frozenset({'HERB MARKER THYME', 'HERB MARKER BASIL'})	frozenset({'HERB MARKER ROSEMARY'})	0.01087480580703916	0.012374779021803181	0.01033910108748058	0.9507389162561575	76.82875908984282

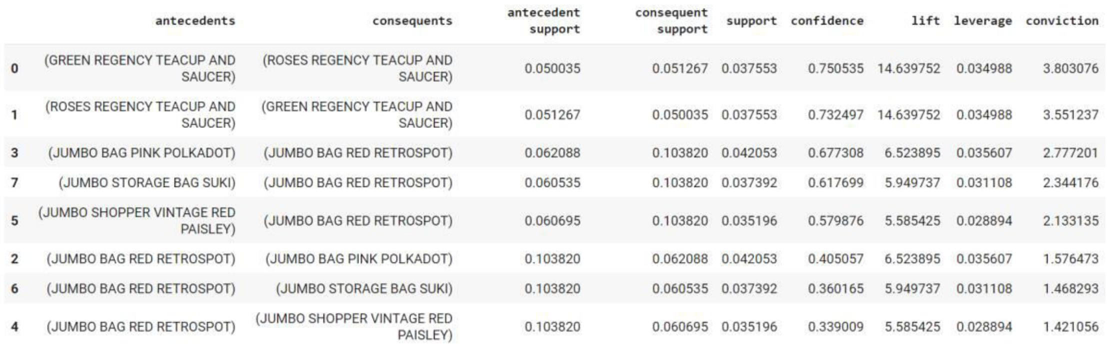
In this study, the rule has more than 2% support and more than 75% confidence, we can see the rules for British transactions are analyzed a little deeper, it is seen that the British people buy different colored tea-plates together. A reason behind this may be because typically the

British enjoy tea very much and often collect different colored tea-plates for different occasions.



index	antecedents	consequents	antecedent support	consequent support	support	confidence
166	frozenset({'PINK REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER'})	frozenset({'GREEN REGENCY TEACUP AND SAUCER'})	0.029249477687898432	0.05003482080677131	0.02641024267423796	0.9029304029304029
164	frozenset({'GREEN REGENCY TEACUP AND SAUCER', 'PINK REGENCY TEACUP AND SAUCER'})	frozenset({'ROSES REGENCY TEACUP AND SAUCER'})	0.030910162318530027	0.05126694166175604	0.02641024267423796	0.854419410745234
27	frozenset({'PINK REGENCY TEACUP AND SAUCER'})	frozenset({'GREEN REGENCY TEACUP AND SAUCER'})	0.037660041784968123	0.05003482080677131	0.030910162318530027	0.8207681365576103
172	frozenset({'JUMBO BAG PINK POLKADOT', 'JUMBO STORAGE BAG SUKI'})	frozenset({'JUMBO BAG RED RETROSPOT'})	0.027053088337708257	0.10381957465045268	0.021696041142122462	0.8019801980198019
146	frozenset({'PINK REGENCY TEACUP AND SAUCER'})	frozenset({'ROSES REGENCY TEACUP AND SAUCER'})	0.037660041784968123	0.05126694166175604	0.029249477687898432	0.7766714082503557

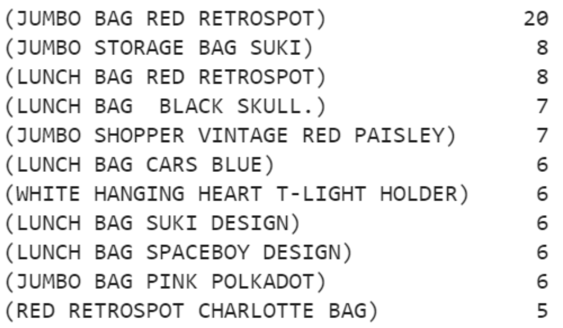
From the output below, in the rule has more than 3.5% support, it can be seen that British people are buying different storage bags in the same order.
This is probably because the British are environmental people and protect the environment and therefore buy reusable bags.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.050035	0.051267	0.037553	0.750535	14.639752	0.034988	3.803076
1	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.051267	0.050035	0.037553	0.732497	14.639752	0.034988	3.551237
3	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.062088	0.103820	0.042053	0.677308	6.523895	0.035607	2.777201
7	(JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.060535	0.103820	0.037392	0.617699	5.949737	0.031108	2.344176
5	(JUMBO SHOPPER VINTAGE RED PAISLEY)	(JUMBO BAG RED RETROSPOT)	0.060695	0.103820	0.035196	0.579876	5.585425	0.028894	2.133135
2	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.103820	0.062088	0.042053	0.405057	6.523895	0.035607	1.576473
6	(JUMBO BAG RED RETROSPOT)	(JUMBO STORAGE BAG SUKI)	0.103820	0.060535	0.037392	0.360165	5.949737	0.031108	1.468293
4	(JUMBO BAG RED RETROSPOT)	(JUMBO SHOPPER VINTAGE RED PAISLEY)	0.103820	0.060695	0.035196	0.339009	5.585425	0.028894	1.421056

<u>Consequents analysis</u>

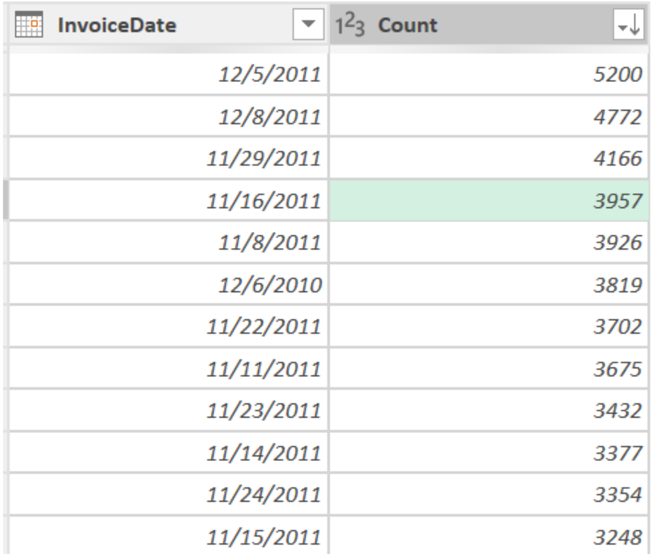
Similar results can be achieved if we take the sum of all the consequents details, you can see that out of all the baskets that customers bought there are 20 baskets that the product jumbo red retro spot bag was a by-product, meaning customers who bought certain products also bought this product, but you can see that 9 out of 10 by-products are reusable bags in different designs. The same conclusion can be drawn that the British take care of the environment and do not use disposable bags



```
(JUMBO BAG RED RETROSPOT)              20
(JUMBO STORAGE BAG SUKI)                8
(LUNCH BAG RED RETROSPOT)               8
(LUNCH BAG  BLACK SKULL.)               7
(JUMBO SHOPPER VINTAGE RED PAISLEY)     7
(LUNCH BAG CARS BLUE)                   6
(WHITE HANGING HEART T-LIGHT HOLDER)    6
(LUNCH BAG SUKI DESIGN)                 6
(LUNCH BAG SPACEBOY DESIGN)             6
(JUMBO BAG PINK POLKADOT)               6
(RED RETROSPOT CHARLOTTE BAG)           5
```

Analyzing the dates

I noticed that most of the big purchases were in November and December and the reason for this is probably the Christmas holiday that many customers purchase gifts for their loved ones



InvoiceDate	Count
12/5/2011	5200
12/8/2011	4772
11/29/2011	4166
11/16/2011	3957
11/8/2011	3926
12/6/2010	3819
11/22/2011	3702
11/11/2011	3675
11/23/2011	3432
11/14/2011	3377
11/24/2011	3354
11/15/2011	3248

Conclusion

1. Change in the value of support has significantly changed the number of rules.

   A support value of 1% produced a rule of 3760 rules.
   A support value of 2% produced a rule of 180 rules.
   A support value of 3.5% produced a rule of only 8 rules.

2. The advantages of the test results for the wholesale company:

   a. Bags can be sold to customers many times in almost every purchase.

   b. Most customers who buy a cup of tea in a certain color will also buy the extra color so we will consider increasing the number of colors of the cups so that the customers will buy more.

   c. Customers purchase different types of spices with every purchase so women on the same pages add more types of spices so that customers will purchase more types.

   d. In the months of December and December, the number of customers increased significantly due to Christmas, so we will make sure that the amount of stock on the shelves is greater to make promotions on products that are consequent.

Reference

1. Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994.
2. AL-MAOLEGI, Mohammed; ARKOK, Bassam. An improved Apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948*, 2014.
3. Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197â€“208, 2012 (Published online before print: 27 August 2012. doi: 10.1057/dbm.2012.17).
4. Goh, Dion H., and Rebecca P. Ang. "An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents." *Behavior Research Methods* 39.2 (2007): 259-266.
5. Hossain, Maliha, AHM Sarowar Sattar, and Mahit Kumar Paul. "Market basket analysis using apriori and FP growth algorithm." *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019.
6. Panjaitan, Suprianto, et al. "Implementation of apriori algorithm for analysis of consumer purchase patterns." *Journal of Physics: Conference Series*. Vol. 1255. No. 1. IOP Publishing, 2019.
7. Pathan, V., and Patil Palande Shende. "A study on Market Basket Analysis and Association Mining." *Proceeding Natl. Conf. Mach. Learn*. 2019.
8. Vaithiyanathan, V., et al. "Improved apriori algorithm based on selection criterion." *2012 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2012.
9. Shaoqian, Yu. "A kind of improved algorithm for weighted Apriori and application to Data Mining." *2010 5th International Conference on Computer Science & Education*. IEEE, 2010.