

A Comprehensive Market Basket Analysis Method using Apriori Algorithm

Himanshu Singh¹, Nikhil Shelke², Aniket Bavaskar³, Shradha Nikam⁴, Prof. Pradip Shewale⁵, Prof. Deepa Mahajan⁶

¹⁻⁶Department Of Computer Engineering, Dr. D. Y. Patil Institute Of Technology, Pimpri, Pune

Abstract - Since the introduction of electronic sale, retailers have had at into their inventory a vast amount of data. The challenge has been how to utilize that data to produce business inference. Most retailers have already figured out a way to understand the basics of the business: what are they selling, how many units are moving and the sales amount. However, few have deployed enough model to analyze the information at lowest level of granularity: the market transaction. The main reason for this is, perhaps, the pre notion that looking at data at this level of granularity is very much expensive and has limited business authority. This article will explore value of market basket analysis through real scenarios, outlining along the way why the users don't need a strong statistics background to understand it and benefit from it.

Market basket analysis, is the process of analyzing transaction-level data to drive business value. At this level, the information is very useful as it provides the business users with direct visibility into the market of each of the customers who shopped at their store. The data becomes a gateway into the events as they happened, understanding not only the quantity of the items that were purchased in that particular basket.

Key Words: Graphology, Market Basket Analysis, Machine Learning, Apriori Algorithm, Jupiter Notebook

1. INTRODUCTION

Apriori is an concept for frequent item set association rule learning over relational databases. It proceeds by processing the frequent individual items in the database and extending them to larger item sets as long as those item sets appear frequently in database. The frequent item determined by Apriori can be used to determine association rules which determine general phenomenon in the database given or loaded: this has applications in fields such as market basket analysis.

The Apriori algorithm was introduced by Agrawal and Srikant in 1994. Apriori is designed to operate on

databases which contains transactions (for example, lists of items bought by customers, or details of a website frequently visited or IP addresses searches regularly). Other algorithms are dedicated for designing and determining association rules in data having no transactions (or having no timestamps). Each transaction is seen as set of items (an itemset).

Apriori uses an approach called bottom up approach, where frequent subsets are extended one item at one time known as candidate generation, and other groups of candidates are tested against the data. The algorithm terminates the code when no further successful extensions found in the dataset.

1.1 Motivation

The human behavior is predictable for many reason and their buying habits don't change overnight which led to buying same combination of things all time, but we are unable or very hard to identify these pattern by over self so we take a machine learning algorithm to do it.

1.2 Proposed Framework

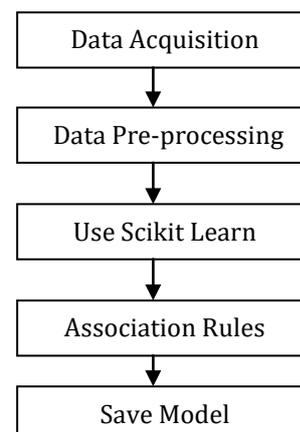
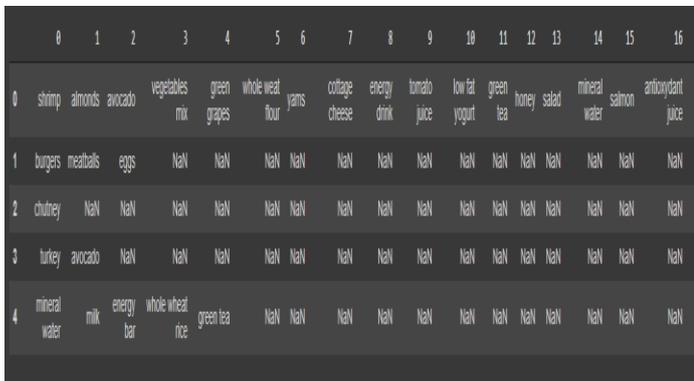


Fig. 1 Proposed Framework

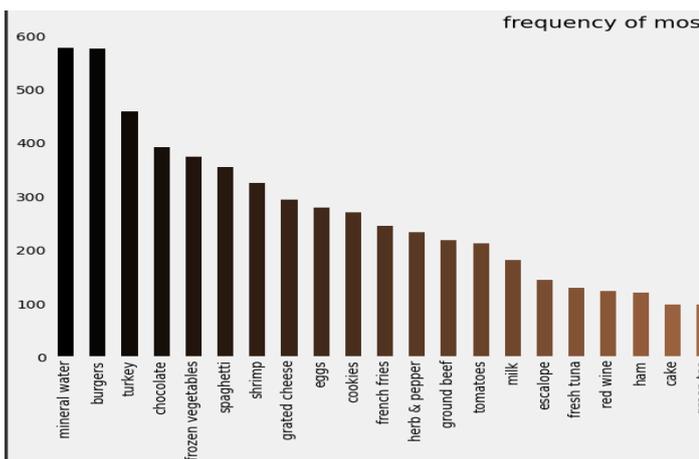
2. IMPLEMENTATION AND RESULTS

Data Acquisition – Load the data of Market Receipt in csv or tsv file into the google colab.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole wheat flour	yams	cottage cheese	energy drink	tomato juice	low fat yogurt	green tea	honey	salad	mineral water	salmon	antioxidant juice
1	burgers	meatballs	eggs	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
3	turkey	avocado	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

Data Pre-processing – Load the data and fill the missing and non-accepted values from the dataset and visualise it.



Apply Apriori – Below are the support vectors for the 2 items which are more than 50 per cent related, means there are 52 per cent chance that if someone buy water then he/she also buy chocolate etc.

	support	itemsets	length
24	0.052660	(mineral water, chocolate)	2
25	0.059725	(mineral water, spaghetti)	2
26	0.050927	(mineral water, eggs)	2

3. CONCLUSIONS

We are implementing the application in which, the input will be the Market receipt to the application, and the that will be forwarded to the system for pre-processing. The dataset variables from the document are broken up into the predefined features. This process is continuous which is determining the relative positions of these features and comparing them with the database of feature-graphs goes on until a match is obtained. The output will be the predicted behaviour at market of that person.

REFERENCES

- [1] Tan, Pang-Ning; Kumar, Vipin; Srivastava, Jaideep (2004). "Selecting the right objective measure for association analysis". *Information Systems*. 29 : 293–313
- [2] Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF).
- [3] Zaki, Mohammed J. (2001); SPADE: An Efficient Algorithm for Mining Frequent Sequences, *Machine Learning Journal*, 42, pp. 31–60.
- [4] 2-Aurélien-Géron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O'Reilly-Media-2019.
- [5] Gionis, Aristides; Mannila, Heikki; Mielikäinen, Taneli; Tsaparas, Panayiotis (2007). "Assessing data mining results via swap randomization". *ACM Transactions on Knowledge Discovery from Data*.
- [6] Menzies, T.; Ying Hu (2003). "Computing practices - Data mining for very busy people". *Computer*
- [7] Liu, Jinze; Paulsen, Susan; Sun, Xing; Wang, Wei; Nobel, Andrew; Prins, Jan (2006). "Mining Approximate Frequent Itemsets in the Presence of Noise: Algorithm and Analysis". *Proceedings of the 2006 SIAM International Conference on Data Mining*. pp. 407–418.
- [8] Zaki, Mohammed J. (2001); SPADE: An Efficient Algorithm for Mining Frequent Sequences, *Machine Learning Journal*, 42, pp. 31–60
- [9] Zimek, Arthur; Assent, Ira; Vreeken, Jilles (2014). *Frequent Pattern Mining*. pp. 403–423.
- [10] King, R. D.; Srinivasan, A.; Dehaspe, L. (Feb 2001). "Warmr: a data mining tool for chemical data". *J Comput Aided Mol Des*. 15 (2): 173–81.