Instacart Market Basket Analysis Using Social Network

by

Xiao Han & Patrick (Xianzhang) Wang

For the course ISOM-673 Social Network Analytics

Table of Contents

[Abstract 3](#_Toc90673415)

[Background 4](#_Toc90673416)

[Data 6](#_Toc90673417)

[Analysis Approach 8](#_Toc90673418)

[Implementation and Results 11](#_Toc90673419)

[Products Analysis 11](#_Toc90673420)

[Department Analysis 15](#_Toc90673421)

[Advanced Analysis of Association Rules 18](#_Toc90673422)

[Grouped Matrix Analysis 18](#_Toc90673423)

[Coefficient Analysis 20](#_Toc90673424)

[Conclusion 23](#_Toc90673425)

[Extension 24](#_Toc90673426)

[Appendix 26](#_Toc90673427)

[Reference 27](#_Toc90673428)

# Abstract

In this project, we are aiming to use market basket data from Instacart to identify the products that most frequently bought together. Using association rules data mining algorithms (Apriori) and social network analytics tool, we analyzed a vast amount of real-world retail data and uncovered some meaningful customer shopping patterns to help retailers such as Instacart to increase sales and maximize profit.

The Instacart dataset of this project includes transaction data with 131,209 unique orders, 39,199 unique products and 21 unique departments. Using R programming, we applied packages such as ‘arules’ and ‘igraph’ to analyze the data. For the analysis, we analyzed both the product data and department data to gain insights on the specific product level and general department level. The result shows clear core-periphery structure with cores as *bananas* in the product analysis. In the department analysis, we are also able to find departments that are worth the attention for cross-selling, such as *bakery* department. We also conducted some advanced analysis such as Grouped Matrix Analysis and Coefficient analysis. Using k-means clustering method and the Jaccard distance, the analysis shows the products and departments groups with relatively higher lift values. The correlation analysis result also indicates that customers tend to make similar transactions following the association rule.

This project utilizes the Apriori Algorithm and introduces the usage and implementation of this algorithm with clear examples and applications. The project output and results, with interactive tables and social network graphs, are easy to comprehend by different stakeholders and applicable to different business context. As future steps, other algorithms such as FP-Growth could be explored.

# **Background**

Understanding the shopping patterns of customers includes figuring out the multiple product combinations purchased together from the same department or from multiple departments. With the easily obtained purchase details (POS machine receipt, online order details, and so on), retail companies are able to utilize the transaction data in order to adjust the location of products in the aisle or plan marketing activities. Online retailers can also use the data to provide “frequently bought together” information or “recommended for you based on your previous orders” information to improve upselling and cross-selling.

It is not a difficult problem to find the product combinations that are frequently bought together because theoretically we can just simply calculate and list all the possible product combinations. However, given the enormous number of products in a retail store or on an online retailing site, this easy problem can be computationally impossible to solve, or at least take a long period of time. According to Bloomberg (2020), Instacart processed over $35 billion in sales in 2020 and had millions of new customers and booming delivery during pandemic (Kang J., 2020). If we want to consider all the possible combinations of items even regardless of volume, for instance, the number of items equivalent to n, the number of combinations we need to take into account N (Appendix A) would be:

With exponential growth and time complexity being the biggest challenges, however, retail stores need to make decisions fast to uncover the changing customer needs, adjust the product price, and make weekly marketing or promotion advertisements. For online retail sites like Instacart, real-time decisions are especially essential to provide customers with purchasing suggestions based on previous orders.

In order to improve the calculation speed and ensure accuracy, marketing analysts have utilized many algorithms such as Apriori and FP-Growth. The utilization of these kinds of techniques are in the category of association rules data mining and unsupervised model. This technique is popular and universally applied in market basket data analysis.

More importantly, the algorithm helps Instacart, and retailers find out which product combination is more significant and stronger than others. With the significance value as the edge weight, we can make simpler and much more meaningful social network graphs to showcase the purchasing relationships among the products. The overview of products will reveal hidden relationships and provide quantitative evidence of relationships.

The result of our analysis will be association rules that help us generate conclusions that if the customer buys one or some items, what the probability is that the customer also buys some other products. As the results of our analysis, these association rules are easy to interpret for different stakeholders and to implement in different business context.

# **Data**

The data source of the project is the Instacart Market Basket Analysis Competition from Kaggle, which is designed for prediction competition where winners generally used supervised models, such as XGBoost and Light GBM. This dataset is provided by Instacart and contains 1,384,617 records for each product in one order, and other variables such as whether it was reordered, the sequence of products added to the order and the department it belongs to. Table 1 shows the first 5 records of the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| department\_id | 4 | 4 | 4 | 4 | 15 |
| aisle\_id | 24 | 24 | 83 | 83 | 95 |
| product\_id | 13176 | 47209 | 10246 | 49683 | 43633 |
| order\_id | 1 | 1 | 1 | 1 | 1 |
| add\_to\_cart\_order | 6 | 7 | 3 | 4 | 5 |
| reordered | 0 | 0 | 0 | 0 | 1 |
| product\_name | Bag of Organic Bananas | Organic Hass Avocado | Organic Celery Hearts | Cucumber Kirby | Lightly Smoked Sardines in Olive Oil |
| aisle | fresh fruits | fresh fruits | fresh vegetables | fresh vegetables | canned meat seafood |
| department | 4 | 4 | 4 | 4 | 15 |

Table 1: First 5 records of the Instacart dataset

We transferred the format to be a sparse matrix of items and orders in R as transactions for processing. The transaction data include 131,209 unique orders, 39,199 unique products and 21 unique departments. Figure 1 shows the distribution of the number of items purchased in one order. Most of the orders have 5 to 10 items. Figure 2 shows the most frequently bought items. The most frequently bought item is Banana.

Chart, histogram

Description automatically generated

Figure 1: Distribution of the number of items purchased in one order

Chart

Description automatically generated

Figure 2: Most frequently bought items

# **Analysis Approach**

In this project, we mainly used the Apriori algorithm and social network tools to uncover and interpret the product combinations. R is our primary programming language. The model output is showcased using the html engine featuring the interactive tables and interactive social network graphs.

Apriori is an algorithm proposed by Agrawal and Srikant in 1994 to discover the frequent item sets from transactional data. The name of the algorithm, *Apriori*, is based on the method of this algorithm. This algorithm uses the *priori knowledge* of product combinations. If one itemset is infrequent, all its supersets are not frequent and should be pruned, as we can see in the following figure for illustration.

Diagram

Description automatically generated

Figure 3: Apriori algorithm example

There are some key concepts in the Apriori Algorithm and its implementation. In the following example, the concepts are introduced using simulated retail data.

Five orders are simulated below with the products purchased in this order.

|  |  |
| --- | --- |
| Order 1 | Milk, Apple |
| Order 2 | Apple, Diaper, Beef, Banana |
| Order 3 | Beer, Diaper, Beef, Orange |
| Order 4 | Apple, Beer, Diaper, Beef |
| Order 5 | Apple, Beer, Diaper, Orange |

Table 3: Simulated orders and products purchased in this order

An association rule is defined as the directed relationship between one or multiple products and another product. It is an implication expression of the form X => Y, where X and Y are disjoint itemsets. For example, the rule {diaper} => {Beer} in this simulated order table indicated the directed relationship that if a customer purchases diaper, the customer is likely to also buy beer. The relationship is not limited to a one-to-one relationship. An association rule could also be defined as {Apple, Diaper} => {Beer}, indicating that if a customer purchases apple and diaper, it is likely that the customer also purchased beer. These rules are the basis for our further analysis. Breaking down the whole product purchasing network into different directed edges would also help us to make decisions based on the product relationships.

The evaluation of an association rule could be evaluated by some measures such as support, confidence, and lift.

* Support: The fraction of transactions that contain both X and Y. Support determines how often a rule is applicable to a given data set (Tan et al., 2005).
* Confidence: Confidence determines how frequently items in Y appear in transactions that contain X (Tan et al., 2005). The higher the confidence, the more likely it is for Y to be present in transactions that contain X.
* Lift (aka. interest ratio): Measures the ratio of confidence against its baseline, the support of Y (Tan et al., 2005). It can identify the condition where some products are popular enough to occur in the order and the recommendation is inessential (Brin et al., 1997). Lift greater than one implies sales improvement of recommendations.

In the example in Table 3, the association rule ({Diaper} => {Beer}) has a support of 3/5 =0.6 since it occurs in 60% of all transactions (3 out of 5 transactions). The confidence of this association rule has a confidence of (3/5) / (4/5) = 0.75, meaning that the probability that {Diaper} is in the order, and {Beer} is also in the order is 75%. The lift of the rule is (3/5) / (4/5) \* (3/5) = 1.25 > 1, which means a 25% improvement in the baseline sale of beer made by the association rule.

# **Implementation and Results**

For the analysis and implementation, we analyzed both the product data and department data to gain insights on the specific product level and general department level.

## **Products Analysis**

Using the arules package in R, we are able to get the association rules using the Instacart Data. Two parameters are used as thresholds to evaluate and limit the association rules as output: minimum support and minimum confidence.

To choose the optimal minimum support and minimum confidence, we looped different support and confidence values to generate the according number of association rules and the average lift. The table output below is sorted by the average lift, reflecting the average improvement of association rules in sale.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Support | Confidence | Number of Rules | Average Lift |
| 1 | 0.001 | 0.40 | 59 | 7.469197 |
| 2 | 0.001 | 0.35 | 107 | 7.255647 |
| 3 | 0.001 | 0.30 | 205 | 6.909723 |
| 4 | 0.001 | 0.25 | 425 | 6.508450 |
| 5 | 0.001 | 0.45 | 24 | 5.501728 |
| 6 | 0.002 | 0.50 | 1 | 4.387230 |
| 7 | 0.003 | 0.40 | 1 | 4.387230 |
| 8 | 0.003 | 0.45 | 1 | 4.387230 |
| 9 | 0.003 | 0.50 | 1 | 4.387230 |
| 10 | 0.004 | 0.40 | 1 | 4.387230 |

Table 4: Parameter tuning output (top 10)

For the model results, we expect the number of rules to be around 100, so we have enough rules but not too many to gain insights and take business actions. We chose the support value as 0.001, meaning that the fraction of transactions that contain both X and Y is at least 0.1%. Because we have 130,329 transactions in total, support value being 0.001 means that the selected rules must appear at least 130 times. We chose the confidence value as 0.35, meaning that when the customer purchases product X, the possibility that the customer also purchases product B is at least 35%.

The algorithm output is attached below, showing the number of total transactions, the minimum support count and the rules selected.

Text

Description automatically generated

Figure 4: Apriori output in R

We made the table output interactive and sorted the data based on confidence. In the table, LHS and RHS represent itemset X on the left-hand side and itemset Y on the right-hand side in the rule {X} => {Y} respectively. The measure *coverage* is the support of LHS.

We can find that, from the top 5 rules, the RHS are all *Bag of Organic Bananas* or *Bananas*, showing that when the customer purchases the products on left-hand side, the possibility that the customer also purchases bananas is quite high.

Table

Description automatically generated

Table 5: The interactive table of product association rules

The interactive table result can also be used to search for a specific product. For instance, when filtering *lime* on the right-hand side, we can find out the products on the left-hand side that will help the sale of *lime* by comparing the measures of the rules.

The social network graph will help us have a better overview of these association rules and see if there is any core-periphery structure in this network. The figure X below clearly shows some cores that could be observed.

Scatter chart

Description automatically generated

Figure 5: Product network generated by package ‘igraph’

To see each point and the edge, we utilized the ‘arules’ package, and created an interactive graph for easier zooming and highlighting, where the blue rectangles represent products and red circles represent rules. In the graph, directed blue edges start from the LHS rectangles, transit at the circle nodes, and point to the RHS rectangles with the color changing to red. For example, a product with high indegree will receive more concentrated red arrows in the graph of the network. In addition, the dark red color of the circle means quite high lift value of the association rule.

The same core-periphery structure could also be observed in Figure X. One core that is evident is *Bag of Organic Bananas,* which corresponds to the result we get from the table. The result from ‘igraph’ package also shows that *Bag of Organic Bananas* has the largest indegree centrality.

Chart

Description automatically generatedChart, diagram

Description automatically generated

Figure 6: Product network

## **Department Analysis**

Another goal of our project is to recommend new products despite the lack of historical order information. To achieve this target, we made an association rule table, a department network and a grouped matrix based on Apriori algorithm with the threshold of support = 0.1, confidence = 0.4 after a similar parameter tuning process.

Table

Description automatically generated

Table 6: The interactive table of product association rules ordered by descending confidence

Table

Description automatically generated

Table 7: the interactive table of product association rules ordered by descending lift

From the table, we can observe that: 1. The rules where RHS is *produce* have the highest confidence (around 0.9) with lift greater than one, which means the high effectiveness of *produce* recommendation on the website of many departments; 2. The rules where RHS is *bakery* have the highest lift (around 1.5), indicating the significant improvement of recommendation on bakery items.

The corresponding business insights are: 1. The designer can put *produce* department on the banner after other departments on user interface; 2. It is reliable to recommend new *bakery* products when customers visit the items of the departments that are the LHS of the strong rules where *bakery* is the RHS.

Chart

Description automatically generatedChart

Description automatically generated

Figure 7: The department network and the node *bakery*

The table output is visualized in the network, where the core-periphery structure has a few cores, such as *dairy eggs*, *produce* and *frozen,* the peripheral departments include *meat & seafood*, *canned goods* and *breakfast*, etc. and the nodes that represent rules where *bakery* is RHS are in dark red which means high lift.

# **Advanced Analysis of Association Rules**

## **Grouped Matrix Analysis**

A grouped matrix is to cluster the numerous association rules and generate more insights. It clusters the items on the LHS using k-means clustering method and the Jaccard distance as the input of the model which describes the similarity between the lift values of rules containing different LHS and the same RHS itemsets, based on the idea that ‘antecedents that are statistically dependent on the same consequents are similar and thus can be grouped together’ (Hahsler and Chelluboina, 2011).

Chart, scatter chart

Description automatically generated

Figure 8: The grouped matrix of department association rules

In the grouped matrix, the nodes are in the sequence of descending median lift value of the group from top to bottom and from left to right. Therefore, we focus on red and purple nodes on the top-left which have highest lift values.

The bright red nodes on the row of *bakery* means the high lift of similar rules where *bakery* is the RHS. The nodes on the column, additionally, indicate how frequent and effective the group of the association rules are. For instance, similar rules containing *deli* and *beverages* as the LHS (the seventh column) cover all the seven departments on the RHS in the figure and have high lift values. It implies that *deli* and *beverages,* thoughthey may not be the LHS of one rule, can lead to improvement in various product sales similarly and suggests that Instacart try recommending products of a new department when customers have visited *deli* or *beverages* despite the lack of order history of the new department.

Chart

Description automatically generated

Figure 9: The grouped matrix of product association rules

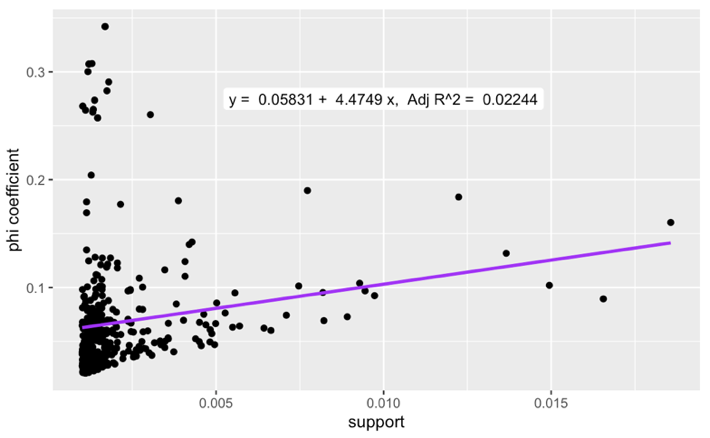
In the resulting visualization of grouped matrix of products, we can find that, compared to the department matrix, most of the product groups have fewer RHS. Many itemset will have a promotion effect on bananas, however, with slight promotion effect. Nevertheless, the Yogurt products have stronger lifts on other kinds of yogurt products, showing that even though purchasing Yogurt does not necessarily lead to the purchase of many other products, it has strong effect on the purchase of other Yogurt products. In the business context, the various kinds of Yogurt that are most frequently brought together can be good recommendations and alternative products when one kind of Yogurt is out of stock.

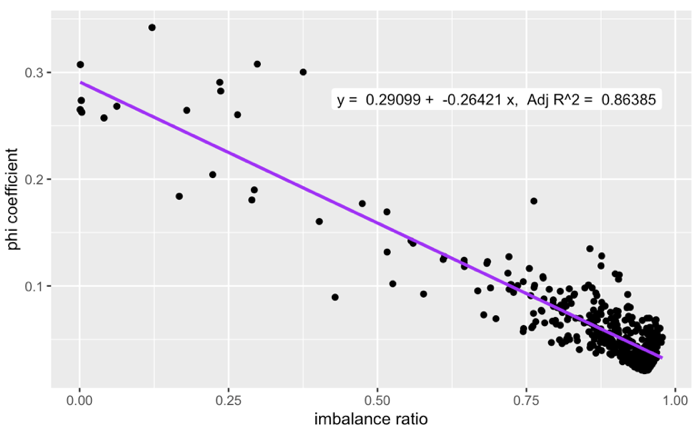
## **Coefficient Analysis**

A strong rule as defined will provide suggestions on recommendations, while there are many measures to evaluate the rule from several aspects. One way to understand the strong rule better is to know the relationship between the measures and the strength of the rule.

Among multiple measures, φ-coefficient represents the value of strength of the rule. To generate some insights on what results in strong rules, we used the product data for the correlation, and ran a linear regression model between a few measures and φ-coefficient respectively, including support, imbalance ratio and Jaccard similarity.

For a rule {X} => {Y}, 1. , the strength of the rule (Tan et al., 2004); 2. , the frequency of the rule in the dataset; 3. , the degree of the skewness of the rule, which evaluates how similar the conditional probabilities of X and Y are, defined by Wu et al. (2010); 4. , the similarity between the transactions containing both X and Y (Tan et al., 2004).





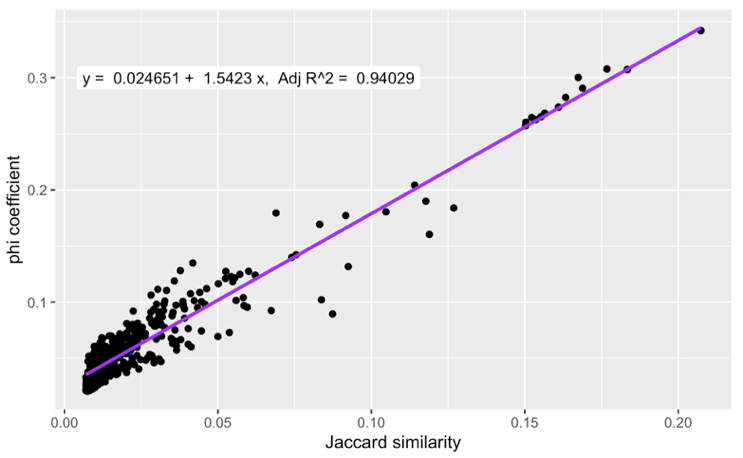


Figure 10: φ-coefficient by support, imbalance ratio and Jaccard similarity

In the figure, we can observe that: 1) the correlation between φ-coefficient and support is unreliable with an adjusted R-squared = 0.02244; 2) the correlation between φ-coefficient and imbalance ratio is negative with an adjusted R-squared = 0.86385; the correlation between an adjusted R-squared = 0.02244 and Jaccard similarity is positive with an adjusted R-squared = 0.94029 and a relatively high coefficient.

Our finding is that a lower degree of skewness of the rule (similar conditional probabilities of X and Y, which are and , as well as a higher similarity between the transactions including the rule, is correlated with a higher strength of the rule. A rule that frequently occurs in the order data, however, is not necessarily to be a strong rule.

The result aligns with the Apriori algorithm where support is only the first step and frequent itemsets should have high confidence to be strong rules. It also indicates that customers tend to make similar transactions following the association rule.

# **Conclusion**

Using R programming, we applied Apriori Algorithm to analyze the Instacart transaction data. From the analysis, we are able to gain business insights into the product level and the department level. The network analysis shows clear core-periphery structure. Some products such as *banana* and some departments such as *bakery*, as cores, are worth the attention to making pricing plans or marketing plans to increase the sales of other products. We also provided multiple tools and frameworks to display and further analyze the data.

With more and complete data, the result of analysis could be implemented in multiple areas for Instacart to grow the business, such as marketing and pricing. As an essential way to increase the cross selling and up selling, the information of “frequently bought together” or “recommended to you based on your purchase” could be obtained from the Apriori Algorithm dynamically with the data of each customer’s purchase information. Especially for online retailers such as Instacart, the algorithm-based recommendation system is easy to be applied to the daily online shopping experience and provides personalized recommendations. Also, the information could be used for marketing purposes and propose personalized discounts.

As an important metric of our analysis, lift is a valuable metric for adjusting the price of the products. Based on the algorithm, we can identify the products pairs which one of them would increase the sales of others. Combined with other information, it is possible to lower the price of the item which will lead to the purchase of other items.

# **Extension**

Our Instacart project has successfully used Apriori algorithm, networks, and advanced analysis to find out association rules, make recommendations and provide business insights. However, we can also extend the analysis to other topics, including model selection, big data scenarios, consumer behavior and association rules of other objects.

Apart from Apriori algorithm, some other models are designed to generate association rules between products for recommendations, for instance, FP-Growth (Han, et al., 2000) and ECLAT (Ogihara, et al., 1997). We can compare the processing time using different sets of parameters, although we are unable to make comparisons between the results of these unsupervised learning methods.

The dataset we used contains only 1,384,617 observations, while we have discussed the time complexity of the problem and the large dataset Instacart and other corporates have. Considering the high demand for updating the model and handling big data, we can extend the analysis using R, Python, and relational databases to the analysis using PySpark and data warehouse, making our model more applicable.

With more data, traditional retailers and online retailers can gain more business insights from the data. 1. With customer id, the purchase behaviors of customers in a specific region could be analyzed using social network analysis and Apriori algorithm. For retailers, the analysis result will be helpful for the supply chain management and area-based marketing. 2. Item location on an aisle: Using image recognition, the data of item location on an aisle is relatively easy to obtain. We are able to combine the sales data and item location data to have more specific information of association rules and use the association rule results to arrange the items on an aisle. 3. Date and time of purchase: Some items have more sales in a specific period of time of a day or a year. With date and time information, the calculation of association rules would be more specific and accurate.

We can try association rules mining and build the network of not only products, but also texts and images, based on the data mining techniques designed for these types of data. This challenging task is worth trying after we learn more about special data preprocessing methods, advanced frameworks, and the application of association network.

# **Appendix**

**The computation of the number of possible item combinations**

Let *n* denote the number of items, and *N* denote the number of combinations.

We can regard one combination as itemset X and itemset Y in one set, which takes two steps: 1. Obtain X including *i* items, from the *n* items; 2. Obtain Y including *j* items, from the *n-i* items.

1. The number of combinations of the first step, *N1*, is:

It means we take *i* items from *N* items, excluding the condition where we take *n* items, because we should guarantee there is at least one item in Y.

2. The number of combinations of the first step for each i, *N2*, is:

3. Now we can compute *N = N1 \* N2*.

After simplification,

# **Reference**

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (pp. 255-264).

Bloomberg L.P. (2020). Instacart’s Frantic Dash from Grocery App to Essential Service. Retrieved from <https://www.bloomberg.com/news/features/2020-05-06/instacart-was-overwhelmed-by-coronavirus-overnight>

Hahsler, M., & Chelluboina, S. (2011). Visualizing association rules: Introduction to the R-extension package arulesViz. *R project module*, (6), 223-238.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, *29*(2), 1-12.

Instacart (2017). Instacart Market Basket Analysis Competition. *Kaggle*. Retrieved from <https://www.kaggle.com/c/instacart-market-basket-analysis/overview>

Kang J. (2020). Instacart Looked Like a Savior. Now Stores Aren’t So Sure. *Wall Street Journal.* Retrieved from <https://www.wsj.com/articles/instacart-looked-like-a-savior-now-stores-arent-so-sure-11609151401>

Ogihara, Z. P., Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. In *In 3rd Intl. Conf.* on *Knowledge Discovery and Data Mining*.

Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293-313.

Tan, P. N., Steinbach, M., Kumar, V. (2005). Introduction to Data Mining. Addison Wesley. ISBN: 0321321367.

Wu, T., Chen, Y., & Han, J. (2010). Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, *21*(3), 371-397.