Journal of Science, Engineering and Technology, Vol. 6 (1), March 2019: pages 51-59

CONCEPTUALISATION OF MARKET SEGMENTATION AND PATTERNS FOR PRE-CHRISTMAS SALES IN AN ONLINE RETAIL STORE

Anthony O. Otiko[1], John A. Odey[2], Gabriel A. Inyang[1]

Email: otikotony@gmail.com; otikotony@crutech.edu.ng; johnodey@unical.edu.ng; gain140270@gmail.com

[1]Department of Computer Science, Cross River University of Technology, Calabar, Nigeria

[2]Department of Computer Science, University of Calabar, Calabar, Nigeria

ABSTRACT

In the last 25 years, digital marketing has become a key component in retail business. There has been a considerable growth in the number of online retail stores and online sales. This paper creates a market segmentation for an online retail store, using Association Rule Mining and Clustering. The segmentation provides information on the clusters of buying patterns for Pre-Christmas sales. Analysis is done using SAS and R mining tools.

Keywords: Online, Retail, Segmentation, Association, Patterns, Strategies, Mining, Clusters

1.0 INTRODUCTION

In the last two decades, there is substantial increase in digital marketing as it has become a key component in retail business. This has led to a considerable growth in number of online retail stores and sales. With this, further research is required to provide adequate marketing strategies to boast sales which this paper attempts to offer.

Smith (1956) first introduced the concept of market segmentation and defined it as a "process of subdividing a market into distinct subsets of customers that behave in the same way or have similar needs. Each subset may conceivably be chosen as a market target to be reached with a distinctive marketing strategy" (Doyle, 2011). We are interested in finding the buying patterns of people from online retail stores. Some discussions are presented in Singh et al (2014).

Our main objective in this project is to find out the buying patterns of the customers, that is, if they buy certain product, how likely are they to buy another particular product. To investigate such relations for the transactions data, several data mining techniques have been used in the literature (Brusco et al, 2003, Ho et al, 2012 etc). We will employ association rule mining and clustering to a large transaction data of interest and find out patterns in the baskets of individual customers.

2.0 SEARCH STRATEGY

A search for the term "retail transaction data" in the web using Google was done. The largest repository of online machine learning databases are available at: https://archive.ics.uci.edu/ml/datasets.html (Lichman, 2013). In this database, the researcher looked for data sets which are suitable for clustering or classification analysis.

3.0 PREPARATION OF DATASET AND IMPLEMENTATION

The research used the online retail data from Chen et al (2012). It is available from https://archive.ics.uci.edu/ml/datasets/Online+Retail. This data set contains all transactions from 1st December, 2010 to 9th December 2011 for a UK-based online retail company. There are in total 541909 transactions. To reduce the data set, we deleted all cancelled transactions and restricted our data to all transactions between 1st December, 2010 and 23rd December, 2010, that is to the pre-Christmas period. In the reduced data set there are 41753 transactions.

The attributes in the data set are as follows:

- Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.
- Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- Unit Price: Unit price. Numeric, Product price per unit in sterling.
- Customer ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

For the purpose of this analysis, Invoice No, Invoice Date and Country were ignored. Since Stock Code and Description are essentially identifier of the same product, we used only one of them in our analysis.

There are no missing values, so we did not employ any method for missing values. The data set was available in that repository as a Microsoft Excel file. However, we converted it to a comma-separated text file (.csv) for ease of importing into SAS and R.

Table 1: Online Retail Store Data (First few lines)

Invoice No	**Stock Code**	**Description**	**Quantity**	**Invoice Date**	**Unit Price**	**Customer ID**	**Country**
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17850	United Kingdom

3.1 DATA MINING USING SAS ENTERPRISE MINER

To begin with, the researcher imported the data set in SAS and set different variables as follows:

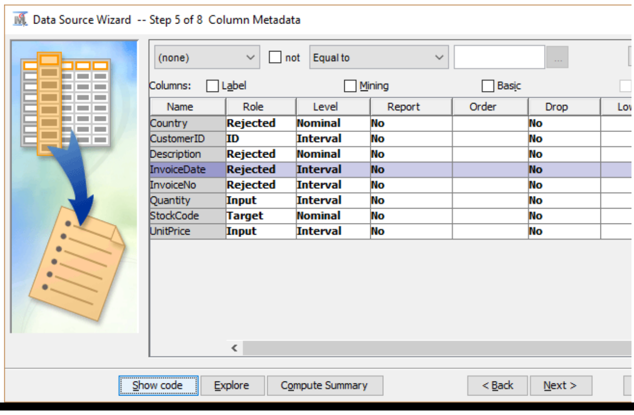

Figure 1: Setting up variables in SAS for Association Rule Mining

3.2 ASSOCIATION RULE MINING

After selecting the data source and setting variables as above, the following diagram for association rule mining in Enterprise Miner were created. The role of the data was changed to Transaction data to make it amenable for Association rules.

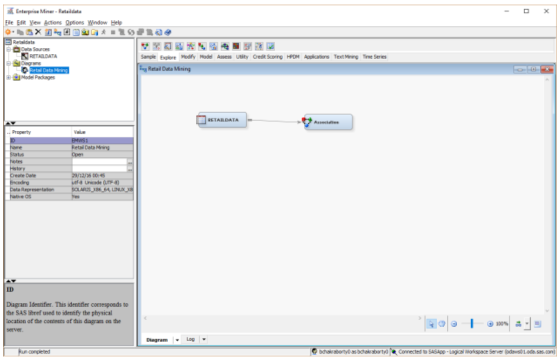

Figure 2: The Diagram in Enterprise Miner for Association Rule Mining

After running the association rules, the results in Tale 2 were obtained:

Table 2: 4-way Association Rules

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
4	0.79	100	0.79	126.57	7	21245 & 20675 ==> 21244 & 20674
4	0.79	100	0.79	126.57	7	21244 & 20674 ==> 21245 & 20675
4	0.79	100	0.79	126.57	7	22727 & 22192 ==> 22726 & 22193
4	0.79	100	0.79	126.57	7	22726 & 22193 ==> 22727 & 22192
4	0.9	100	0.9	110.75	8	21671 & 21669 ==> 21670 & 21668
4	0.9	100	0.9	110.75	8	21670 & 21668 ==> 21671 & 21669
4	0.9	100	0.79	110.75	7	22727 & 22193 ==> 22726 & 22192
4	0.9	100	0.79	110.75	7	22866 & 22534 ==> 22865 & 22530
4	0.9	100	0.79	110.75	7	22632 & 22534 ==> 22865 & 22530
4	0.9	100	0.79	110.75	7	22866 & 22534 ==> 22865 & 22531
4	0.9	100	0.79	110.75	7	22866 & 22530 ==> 22865 & 22531
4	0.9	100	0.79	110.75	7	22632 & 22534 ==> 22865 & 22531
4	0.9	100	0.79	110.75	7	22632 & 22530 ==> 22865 & 22531
4	0.9	100	0.79	110.75	7	22866 & 22530 ==> 22865 & 22534
4	0.9	100	0.79	110.75	7	22632 & 22530 ==> 22865 & 22534
4	0.79	87.5	0.79	110.75	7	22865 & 22534 ==> 22632 & 22530
4	0.79	87.5	0.79	110.75	7	22865 & 22531 ==> 22632 & 22530
4	0.79	87.5	0.79	110.75	7	22865 & 22531 ==> 22632 & 22534
4	0.79	87.5	0.79	110.75	7	22865 & 22530 ==> 22632 & 22534
4	0.79	87.5	0.79	110.75	7	22726 & 22192 ==> 22727 & 22193
4	0.79	87.5	0.79	110.75	7	22865 & 22534 ==> 22866 & 22530
4	0.79	87.5	0.79	110.75	7	22865 & 22531 ==> 22866 & 22530
4	0.79	87.5	0.79	110.75	7	22865 & 22531 ==> 22866 & 22534
4	0.79	87.5	0.79	110.75	7	22865 & 22530 ==> 22866 & 22534

It is observed that for the first 15 relations, the confidence is 100%, which indicate that whenever a customer bought that combinations of items on the left hand side, he/she went to buy the combination of objects on the right. The lift gives strength of the association and all of these association rules have very high lift values.

3.3 CLUSTERING

The role of the data was changed to raw for clustering and the following diagrams were created for clustering with Unit Price and Quantity as input variables to create the segmentation.

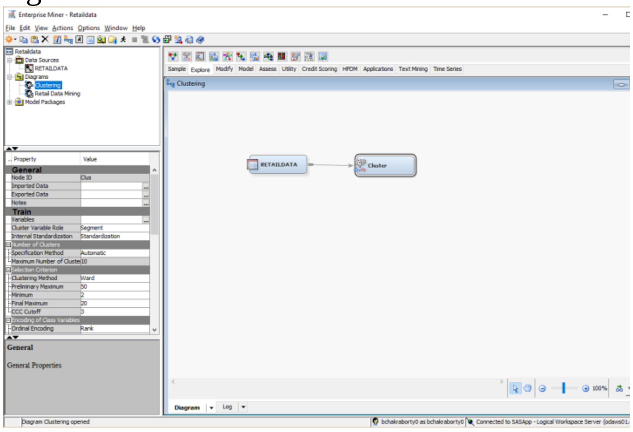

Figure 3: The Diagram in Enterprise Miner for Clustering

After running clustering, the researcher obtained the following segment plot

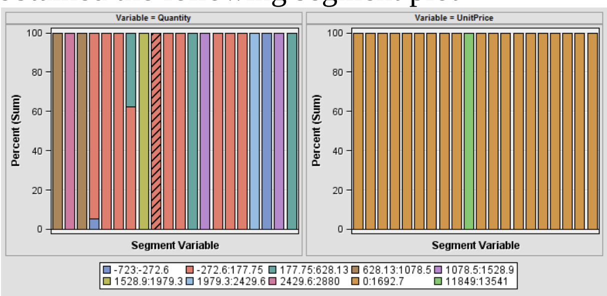

Figure 4: Contribution of the variables Quantity and UnitPrice in different Segments

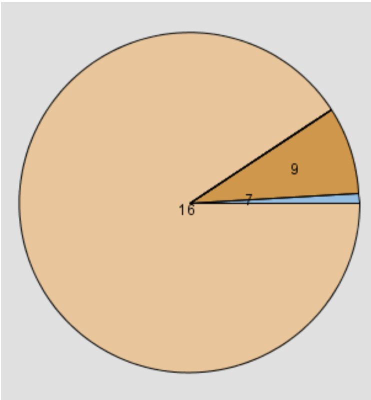

Figure 5: Comparison of Segment sizes

This shows that segment 16 is the largest and two segments 9 and 7 are small but significant. The segment plot shows the percent variations of the variables quantity and UnitPrice in these two segments.

3.4 IMPLEMENTATION IN R

For association rule mining, the package a rules were applied. The data is in a "single" item format that is each line contains a single item and several lines are there for a single transaction with a transaction id. Here we use CustomerId as the Id for creating the baskets and we need to read the data into R and make it a transaction data suitable for using with the functions in arules.

Parameter specification:

Confidence minval smax arem aval originalSupport maxtime support minlen maxlen target
0.8 0.1 1 none FALSE TRUE 5 0.015 1 10 rules ext FALSE

Algorithmic control:

filter tree heap memopt load sort verbose
0.1 TRUE TRUE FALSE TRUE 2 TRUE
Absolute minimum support count: 13
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[2411 item(s), 885 transaction(s)] done [0.00s].
sorting and recoding items ... [529 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [90 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].

The researcher used the minimum value of the support to be 0.015 after trying out several values of the parameter support to obtain a reasonable set of rules. The rules were sorted in a decreasing order of the "lift" parameter and inspect the first 15 rules in the following:

```
     lhs                    rhs      support    confidence lift
[1]  {22300}               => {22301}  0.01807910 0.9411765  39.66387
[2]  {22450}               => {22451}  0.01694915 0.8823529  35.49465
[3]  {22562}               => {22563}  0.01581921 0.8235294  33.12834
[4]  {84997A,84997B,84997D} => {84997C} 0.01581921 0.9333333  33.04000
[5]  {22593}               => {22595}  0.01581921 1.0000000  32.77778
[6]  {84997B,84997D}        => {84997C} 0.01807910 0.8888889  31.46667
[7]  {84997A,84997D}        => {84997C} 0.01694915 0.8823529  31.23529
[8]  {84997B,84997C,84997D} => {84997A} 0.01581921 0.8750000  30.97500
[9]  {84997A,84997C}        => {84997B} 0.01807910 0.9411765  30.84967
[10] {84997A,84997C,84997D} => {84997B} 0.01581921 0.9333333  30.59259
[11] {84997A,84997B}        => {84997C} 0.01807910 0.8421053  29.81053
[12] {84997B,84997D}        => {84997A} 0.01694915 0.8333333  29.50000
[13] {84997C,84997D}        => {84997A} 0.01694915 0.8333333  29.50000
[14] {84997C,84997D}        => {84997B} 0.01807910 0.8888889  29.13580
[15] {22961,22962}          => {22963}  0.01694915 0.8823529  28.92157
```

Table 3: Association Rules from R. Sorted by lift and only the first 15 are presented

Note that, the apriori function in R always creates rules with only one item on the right hand side, unlike SAS. These rules indicate that the customers who bought the item 22300 is highly likely to buy the item 22301 with lift 39.66. Similarly, for 4-way associations, customers who bought 84997A, 84997B and 84997D are likely to buy 84997D. Now note that, the 4th rule, 8th rule and the10th rule are essentially the same. So there is need to prune the rules with identical structures.

```
lhs                    rhs      support    confidence lift
[1]  {22300}               => {22301}  0.01807910 0.9411765  39.66387
[2]  {22450}               => {22451}  0.01694915 0.8823529  35.49465
[3]  {22562}               => {22563}  0.01581921 0.8235294  33.12834
[4]    {84997A,84997B,84997D}  =>  {84997C}  0.01581921  0.9333333
33.04000
[5]  {22593}               => {22595}  0.01581921 1.0000000  32.77778
[6]  {84997B,84997D}        => {84997C} 0.01807910 0.8888889  31.46667
[7]  {84997A,84997D}        => {84997C} 0.01694915 0.8823529  31.23529
[8]  {84997A,84997C}        => {84997B} 0.01807910 0.9411765  30.84967
[9]  {84997B,84997D}        => {84997A} 0.01694915 0.8333333  29.50000
[10] {22961,22962}          => {22963}  0.01694915 0.8823529  28.92157
[11] {20967,20970}          => {20969}  0.01581921 0.8235294  28.03167
[12] {84997C}               => {84997B} 0.02259887 0.8000000  26.22222
[13] {20969}                => {20967}  0.02372881 0.8076923  23.05831
[14] {20725,22662}          => {22382}  0.01581921 0.9333333  20.65000
[15] {22383,22662}          => {22382}  0.01581921 0.8750000  19.35938
```

Table 4: Pruned Association Rules. Sorted by lift and only the first 15 are presented

It was observed that there are strong associations in the buying patterns for items 84997A, 84997B, 84997C, 84997D. Other than that, it was also observe that customers who bought 22300 is likely to buy 22301 and who bought 22450 is likely to buy 22451.

Note that, the set of association rules obtained in R are very different from those obtained in SAS. This is due to the fact that they use different algorithms for association rule mining.

Next, we visualize the association rules, using the arules Viz package.

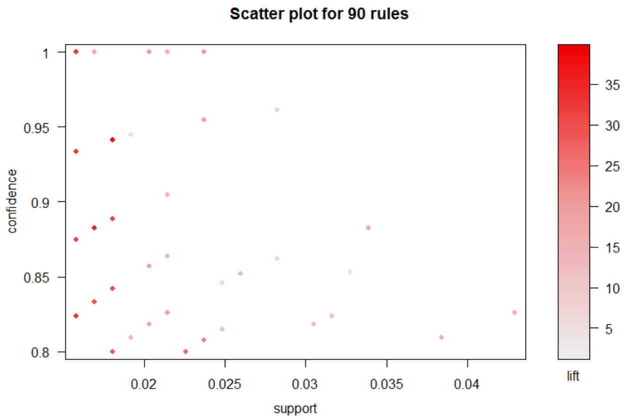

Figure 6: Scatter plot of confidence against support by lift of all rules obtained in R

This plot provides information on confidence, support and lift for the rules generated. We may also like to see the relation using graphs.

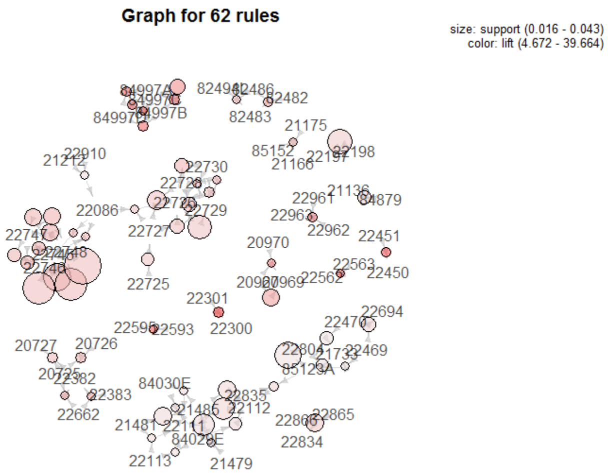

Figure 7: Graph showing all association rules after pruning

It is observed from this plot that the items 84997A, B, C, D are close together with smaller support but high lift values. Items 22747, 22748, 22749 are also associated with high support, but low lift values.

The above graph is easier to visualize if we restrict to first 15 association rules sorted according to lifts.

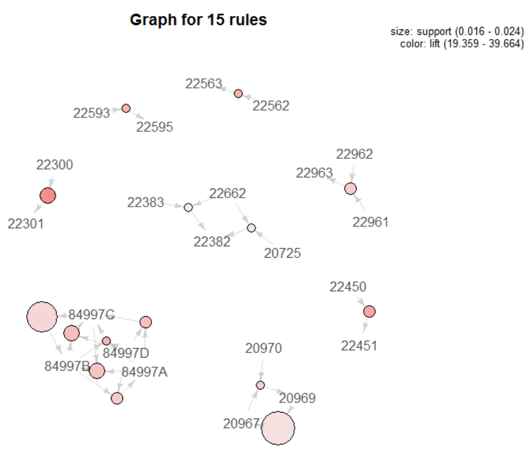

Figure 8: The graph showing top 15 pruned association rule sorted by lift.

The following parallel coordinate plot also helps us to visualise the associations.

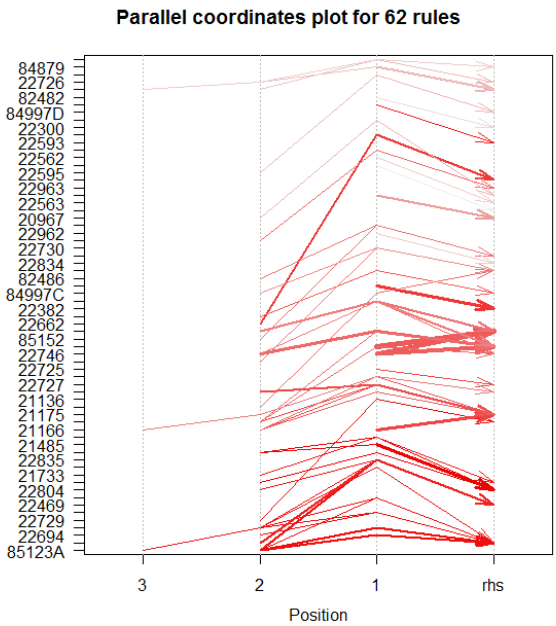

Figure 9: Parallel coordinates plot for all pruned association rules

Similarly for the first 15 association rules sorted by lifts:

```
>plot(rules.pruned[1:15], method="paracoord", control=list(reorder=TRUE))
```

Now the researcher performed clustering of the items based on how many times they appeared in customers' baskets using R. Firstly, a dissimilarity matrix was created using the co-occurrences of items in each customer's basket in a pairwise manner. A hierarchical clustering was performed with complete linkage using the function hclust and a dendrogram with 5 clusters were ploted.

```
retail.ctree
   1    2   3  4  5
2457  236  98 10 10
```

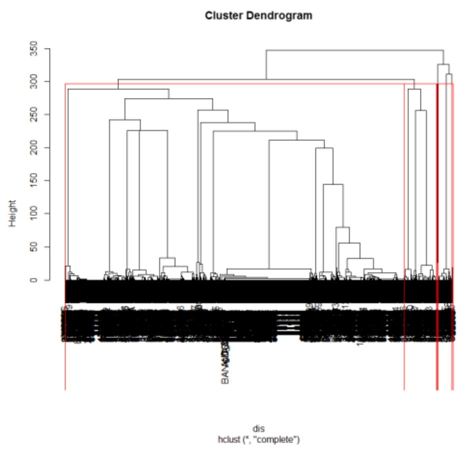

Figure 10: Dendrogram showing the clusters obtained from clustering the customer's baskets.

It is seen that there are mainly 3 clusters 2457 items, which are most likely the lowly sold items and two other clusters of sizes 236 and 98. The smaller clusters of size 10 might be interesting as they may contain the items which appear in most of the customers' baskets. The items in these two clusters are:

```
> names (retail.ctree) [retail.ctree==5]

 [1] "21209"  "22111"  "22214"  "22492"
"22633"  "22745"  "79321"  "84898F"
"85035B"
[10] "85180B"
> names (retail.ctree)[retail.ctree==4]
 [1] "21034"  "21630"  "22055"  "22338"
"22911"  "22962"  "84415B"  "85025C"
"85062"
[10] "85123A"
```

4.0 RESULTS, ANALYSIS AND DISCUSSION

The top 3 association rules obtained in SAS (Table 2) suggest that the items {21244, 21245, 20674, 20675} are associated, {22726, 22727, 22192, 22193} are associated and {21668, 21668, 21670, 21671} are associated. It is observed that, products with consecutive stock codes are associated. This can be explained because they may be very similar products which are related products.

The top association rules obtained in R (Table 4 and Figure 8) suggest that the items {84997A, 84997B, 84997C, 84997D} are associated, {22300, 22301} are associated, {22450, 22451} are associated, {20969, 20970, 20971} are associated. Again we observe similar consecutive stock codes.

Clustering in R produces two interesting clusters of 10 items each: {"21209" "22111" "22214" "22492" "22633" "22745" "79321" "84898F" "85035B"} and {"21034" "21630" "22055" "22338" "22911" "22962" "84415B" "85025C" "85062"}. These items are showing some buying patterns in customers' baskets. These clustering is obtained by creating a co-occurrence of items in customers' baskets.

5.0 CONCLUSION

In this paper, association rule mining and clustering were used to find out segmentation or clusters and buying patterns for the Pre-Christmas sale of an online store. We observe that there are some strong association in buying patterns of certain products and we can have clusters of items showing closeness in buying patterns. Depending on the online store, they may use this information in promoting and/or improving their marketing strategies.

It should be noted here that R considers only association rules with only one item on the right hand side, whereas SAS has no such limitation. For this reason, the rules obtained using SAS and R are quite different. However, the rules in both cases are easily interpretable.

REFERENCES

Brusco, M. J., Cradit J. D. and Tashchian A. (2003) Multicriterion clusterwise regression for joint segmentation settings: An application to customer value, Journal of Marketing Research, pp. 225–234.

Daqing C., Sai L. S., and Kun G. (2012). 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. 197-208.

Doyle C. (2011) A dictionary of marketing. Oxford University Press.

Ho G.T, Ip W., Lee C., and Mou W. (2012) Customer grouping for better resources allocation using GA based clustering technique, Expert Systems with Applications, vol. 39, no. 2, pp. 1979–1987.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Singh A. K, Rumantir G, South A., and Bethwaite A. (2014). Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14). ACM, New York, NY, USA, Article 16, 7 pages. DOI=http://dx.doi.org/10.1145/2640087.2644161