Market Segmentation of EFTPOS Retailers

Ashishkumar Singh **Grace Rumantir**
Faculty of Information Technology Caulfield
Princes Highway, Caulfield East
Victoria 3145
{ashish.singh, grace.rumantir}@monash.edu

Annie South
Australia New Zealand Bank
Level 9, 833 Collins Street,
Docklands, Victoria 3008
annie.south@anz.com

Abstract

Almost all the papers on market segmentation modeling using retail transaction data reported in the literatures deal with finding groupings of customers. This paper proposes the application of clustering techniques on finding groupings of retailers who use the Electronic Funds Transfer at Point Of Sale (EFTPOS) facilities of a major bank in Australia in their businesses. The RFM (Recency, Frequency, Monetary) analysis on each retailer is used to reduce the large data set of customer purchases through the EFTPOS network into attributes that may explain the business activities of the retailers. Preliminary results show that groupings of retailers with distinct combinations of RFM values can be established which encourage further modeling using other business and demographic attributes on bigger data set over a longer period of time.

Keywords: Market Segmentation, Clustering, RFM Analysis, EFTPOS.

1 Introduction

Electronic Funds Transfer at Point Of Sale (EFTPOS) is one of the leading methods of payment at checkouts or Point Of Sale (POS). Payments on EFTPOS terminals are done using debit and credit cards, the most common non-cash payment methods. The banking sector gains profits by the renting out EFTPOS machines to retailers through set up fee, periodic service fee and transaction fee on each purchase put through the EFTPOS machine. Banks also profit through the availability of interest free fund en-route to the designated retailer's bank account after being debited from the payee's account and the availability of the deposited fund into the retailer's account itself. In order for the banking sector to develop this part of the business further, it can make use of market segmentation modeling to gain better understanding on the business behaviours of the retailers using the EFTPOS facilities.

The concept of market segmentation was first introduced in Smith (1956) and defined as the "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). The heart of any good market segmentation tool is its ability to analyse, understand and draw good market segments based on customer's purchasing behaviours. Whilst market segmentation has been extensively applied on transaction data to find insights into customer purchasing behaviours on the market, very limited work has been done to find in-sights into retailer business activities. This paper reports on the use of clustering techniques on EFTPOS transaction data to identify segments of retailers who have certain common characteristics.

The selection of attributes plays an important role in good clustering analysis. This paper uses the RFM (Recency, Frequency, Monetary) analysis, popular in marketing, to reduce the data set for the clustering experiments.

The paper is organized as follows: Section 2 gives a summary of the review of the literatures on market segmentation papers for the past 10 years; Section 3 explains the data reduction/transformation using the RFM Analysis; Section 4 briefly explains the clustering techniques used; Section 5 outlines and discusses the results of the experiments; Section 6 concludes the paper and explains our plans for the future.

2 Related Work

Table 1 summaries our review of the literatures on related work in market segmentation with respect to the attribute selection techniques and clustering techniques employed. All of the papers reviewed report the use of transaction data as input and segmentation experiments on customers. Only one paper i.e. (Bizhani & Tarokh 2011), reports segmentation experiments on retailers using EFTPOS data. This is the only work on EFTPOS data we have found in the literatures. The work reported in Bizhani & Tarokh (2011) is on a much smaller data set and considers individual EFTPOS machines as "retailers". The data set we have been acquiring for our work falls in the category of Big Data with each merchant/retailer having multiple EFTPOS machines. Our experience in acquiring, secured-storing and processing the commercial in confidence EFTPOS data has been reported in Singh, A., et al. (2014).

As shown in **Table 1** socio-demographic analysis and RFM analysis are the two most popular attribute selection techniques for market segmentation. For example, in Y. Kim et al. (2005) and J. H. Lee & Park (2005), attributes derived from socio-demographic characteristics (e.g.

Copyright (c) 2014, Australian Computer Society, Inc. This paper appeared at Australasian Data Mining Conference (AusDM 2014), Brisbane, 27-28 November 2014. Conferences in Research and Practice in Information Technology, Vol. 158. Richi Nayak, Xue Li, Lin Liu, Kok-Leong Ong, Yanchang Zhao, Paul Kennedy Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

average age of residents, proportion of residents in high status and others) of customer have been reported as effective in customer segmentation. Whereas in Olson et al. (2009), attributes based on RFM comparatively yield better clustering results and can effectively handle a large degree of multi-dimensional data. With respect to clustering algorithms used, K-means and hierarchical clustering are the most popular techniques for market segmentation.

Table 1. Summary of review of the literatures on market segmentation in the past 10 year

Related Work	Attribute Selection Techniques	Clustering Techniques			
	Socio-demographic Analysis	RFM Analysis	K-means	Hierarchical clustering	Other
(Y.-S. Chen et al. 2012) (Lefait & Kechadi 2010)		X	X		
(Hsieh 2004)		X			X Neural Network
(Bizhani & Tarokh 2011)		X	X		X Unsupervised Learning Vector Quantization
(D. Chen et al. 2012)		X	X		X Decision Tree
(Olson et al. 2009)		X			X
(Namvar et al. 2010)	X	X	X		
(Y. Kim et al. 2005) (J. H. Lee & Park 2005)	X				X Neural Networks
(Dennis et al. 2003)	X				X
(Ho et al. 2012)			X Genetic Algorithms		
(Salvador & Chan 2004)			X	X	
(D. Gaur & S. Gaur 2013)			X	X	X
(Zakrzewska & Murlewski 2005)			X	X	X Density based Clustering
(Alam et al. 2010) (Yoon et al. 2013)				X	
(Li et al. 2009)				X Chameleon	
(Suib & Deris 2008)				X Hierarchical Pattern Based Clustering	

3 RFM Analysis

EFTPOS transaction data are being collected from one of the four major banks in Australia. This paper reports on the preliminary stage of our project where we use data for a total period of 18 days, starting from 19-September-2013 to 07-October-2013. Each transaction record has 55 attributes.

The 18 daily data files contain approximately 77.5 million transaction records from over 1 million unique retailers. This high volume of data makes even the basis operations, such as finding the total monetary amount of each retailer in the data set, very time consuming and resource intensive. To overcome this Big Data problem, a hash table for the retailer ids is used.

Normalisation is very important in this clustering project as clustering algorithms use various distance measures between attribute values. An attribute with values covering a large range, like retailer's Monetary values, may dominate over another attribute with smaller range of values, like retailer's Recency and Frequency values. To alleviate this problem, in this project, the min-max normalization technique is used where the values of all three attributes are normalised into a similar range between 0 and 100.

The RFM (Recency, Frequency, Monetary) analysis was proposed in Hughes (2006). This paper proposes the use of these three attributes to group retailers exhibiting similar business activities:

Recency - the Recency value for a retailer is the time interval between a global datum and his/her latest transaction. A retailer with a smaller Recency value is seen as more current in his/her business activities than a retailer with a bigger Recency value. The global datum is chosen as midnight after the last transaction day in the data set (i.e. the midnight of 8th October 2013 (00:00:000000 in HH:MM:SSSSSS format). Hence, all the Recency values are calculated from midnight of 8th October 2013 backwards.

Frequency - the total number of transactions put through all of the EFTPOS terminals belonging to a retailer forms the Frequency value of the retailer.

Monetary value - the total amount of transactions put through all of the EFTPOS terminals belonging a retailer forms the Monetary value of the retailer.

Figure 1 shows the histogram of the distribution of the Recency, Frequency and Monetary values of all of the retailers in the data set. The histogram shows that the Frequency and Monetary values are skewed to the lower end of the values.

The histogram in **Figure 1** is open ended at the top end because there are small numbers of very large values of the three attributes in the data set. This is shown more clearly in the histograms in **Figure 2** where the horizontal axis of each histogram is intentionally "squeezed" in the middle to show these extreme values and also because there are very few data in this range (no data for Frequency and Monetary values).

Figure 2 also shows the correlations amongst the three attributes. There is positive correlation between Frequency and Monetary values and no correlation between Recency with the other two attributes. This implies that retailers that generate a lot of transactions tend to accumulate large Monetary values within the observation time period.

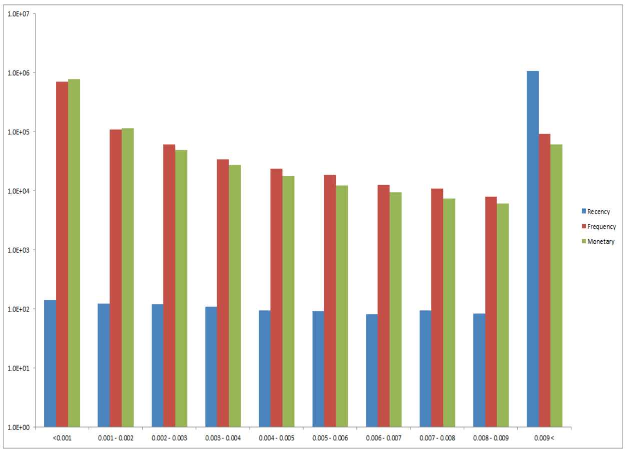


Figure 1. The distribution of Recency, Frequency and Monetary values of all retailers

4 Cluster Analysis

Being the most popular methods used in the literatures for market segmentation, we use K-means and Agglomerative Hierarchical Clustering (AHC) in this preliminary work. Clustering analysis on a large data set is time consuming and processor intensive. with 16 cores and 32 GB RAM is employed. For this work, parallelisation in the form of multi-threading using Intel Xeon and AMD Opteron CPUs of various clock speeds. Our experience in acquiring, secured-storing and processing the commercial in confindence EFTPOS data is reported in Ashish, et al. (2014).

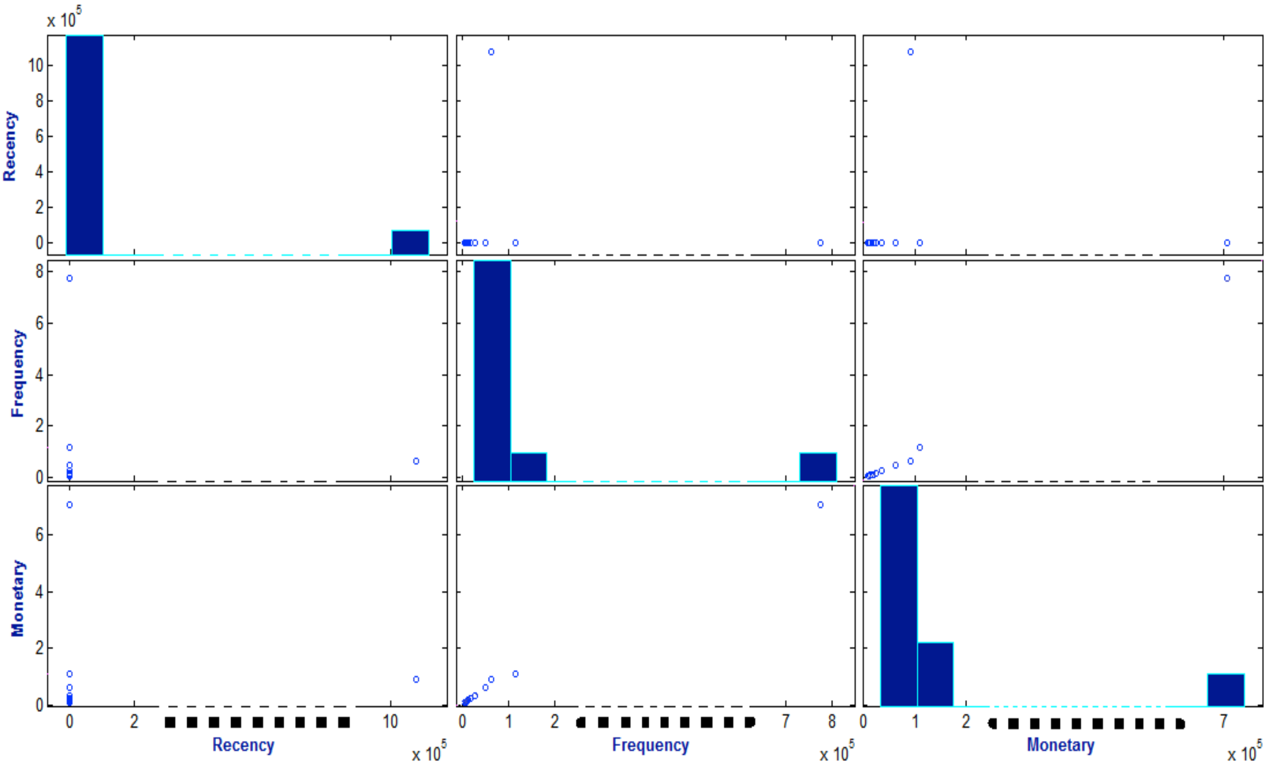


Figure 2. The correlations amongst the three attributes. The middle of each of the x-axis are deliberately truncated to show the small number of extreme values at the upper end, made possible as there are no data Frequency and Monetary values and very few Recency values in the middle range

For the K-means clustering, the calculation of the Euclidean Distance between each data point to the centroid of a cluster is done in parallel over subsets of the data set. For the AHC, the calculation of the Ward Minimum Variance to evaluate the inter-cluster distance measure (starting with clusters each with 1 member data point) and the creation of the clusters through agglomerative process are done on 60% of the data set with the remaining 40% are allocated to the clusters based on Euclidean distance.

In this project, we create clustering models with 2 to 25 clusters using both K-means and AHC.

Both sets of clustering models are then tested based on Dunn's Index (Dunn† 1974) and Davis-Bouldin Index (Davies & Bouldin 1979) to find the number of clusters which result in high intra-cluster similarity and high inter-cluster dissimilarity.

Dunns Index is given as:

$$D = \min_{1 \le i \le c} \left\{ \min_{1 \le j \le c, j \ne i} \left\{ \frac{\delta (C_i , C_j)}{\max_{1 \le k \le c} \{\Delta(C_k)\}} \right\} \right\}$$

where: $\delta (C_i , C_j)$ is the inter-cluster distance between clusters C_i and C_j; $\Delta(C_k)$ is the intra-cluster distance of cluster C_k; c is the number of clusters.

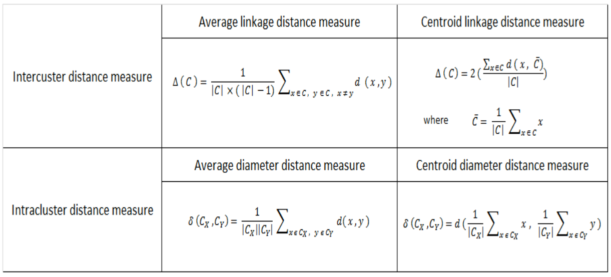
Also Davis-Bouldin Index is given as:

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{i \ne j} \left\{ \frac{\Delta (C_i) + \Delta (C_j)}{\delta (C_i , C_j)} \right\}$$

where: $\delta (C_i , C_j)$ is the inter-cluster distance between two clusters C_i and C_j; $\Delta(C_i)$ and $\Delta(C_j)$ are the intra-cluster distances of clusters C_i and C_j respectively; c is the number of clusters.

Two inter-cluster distance measures and two intra-cluster distance measures are used to evaluate these two indices as outlined in **Table 2**.

Table 2. Two pairs of inter-cluster and intra-cluster distance measures to be used in Dunn's and Davies Bouldin Indices.



	Average linkage distance measure	Centroid linkage distance measure
Intercuster distance measure	$\Delta(C) = \frac{1}{\lvert C\rvert \times (\lvert C\rvert - 1)} \sum_{x \in C,\ y \in C,\ x \ne y} d(x,y)$	$\Delta(C) = 2(\frac{\sum_{x \in C} d(x, \bar{C})}{\lvert C\rvert})$ where $\bar{C} = \frac{1}{\lvert C\rvert} \sum_{x \in C} x$
	Average diameter distance measure	Centroid diameter distance measure
Intracluster distance measure	$\delta(C_X, C_Y) = \frac{1}{\lvert C_X\rvert \lvert C_Y\rvert} \sum_{x \in C_X,\ y \in C_Y} d(x,y)$	$\delta(C_X, C_Y) = d(\frac{1}{\lvert C_X\rvert} \sum_{x \in C_X} x,\ \frac{1}{\lvert C_Y\rvert} \sum_{x \in C_Y} y)$

5 Results and Discussion

The Dunns and Davies–Bouldin indices for each of the 24 clustering models (with 2 to 25 clusters respectively) created using each of the two clustering techniques are plotted in **Figure 3** and **Figure 4** respectively. Comparing the results between the K-means and AHC models, we can see that the AHC models with 16 and 17 clusters have higher Dunn's indices than any of the K-means models. The K-means models with 13 clusters or more tend have higher Davies–Bouldin indices than the AHC models with the same number of clusters. It is then decided to analyse the AHC model with 19 clusters with a view that manual merging of clusters might be warranted post-analysis.

Table 3 shows the AHC clusters, each with the centroid represented by the average RFM values. Each of the R, F, M values in the data set is divided into 5 quantiles, with labels 1 to 5. Based on the centroid values, each cluster is assigned a 3 digit label which values depend on which range each of its R, F, M values falls in.

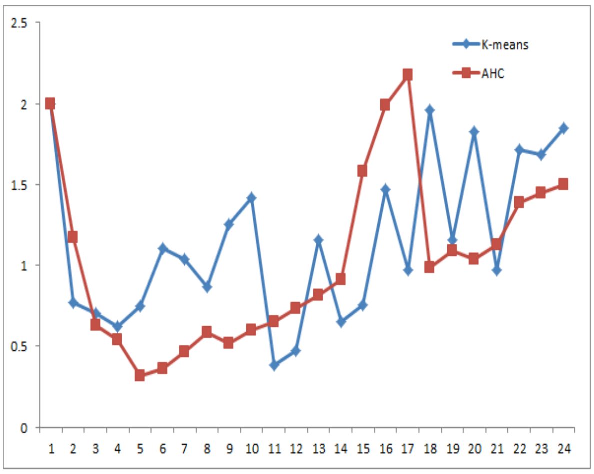


Figure 3. A plot of Dunns indices for each of the number of clusters (x-axis) in the models created using K-means and Agglomerative Hierarchical Clustering techniques

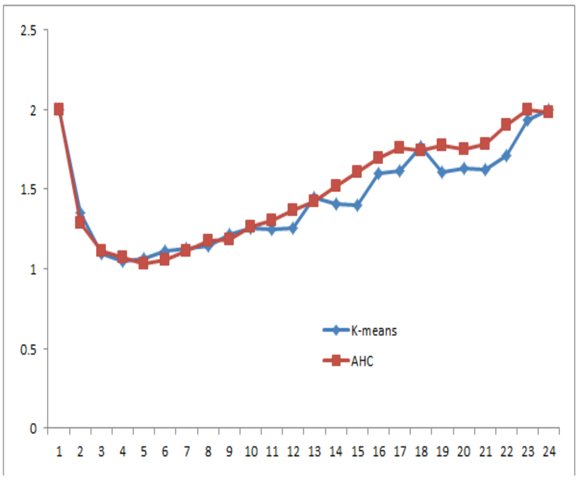


Figure 4. A plot of Davies–Bouldin indices for each of the number of clusters (x-axis) in the models created using K-means and Agglomerative Hierarchical Clustering techniques

Low value in Recency suggests the retailer has recent transactions. Hence, low value in Recency and very high values in both Frequency and Monetary suggest the feature of a retailer with active, successful business. In **Table 3**, retailers in Clusters 1, 2, 18 (constitute 34.23%

of the total retailers) exhibit such characteristics. The bank may wish to provide incentives for these retailers to maintain their excellent business performances.

Retailers characterized by moderately high Recency value with average levels of Frequency and Monetary values are seen in Clusters 5, 6, 13, 14, 15. These retailers have businesses with moderate level of activities but have recently been inactive for some time. The bank may want to engage these retailers into more active participation in generating EFTPOS transactions as this will consequently boost their Frequency and Monetary values.

Table 3. Clustering Results

Cluster #	%	Average Recency	Average Frequency	Average Monetary Value	R,F,M labels (1 to 5 quantiles)
1	**10.30**	**2.368704**	**0.004604**	**0.003177**	**255**
2	**15.27**	**1.018574**	**0.038917**	**0.017611**	**155**
3	2.41	67.028300	0.000117	0.000333	523
4	2.30	71.897224	0.000105	0.000313	523
5	1.72	47.020206	0.000343	0.000331	433
6	3.20	42.105650	0.000293	0.000681	434
7	1.31	82.475670	0.000085	0.000163	522
8	2.14	76.985930	0.000071	0.000287	523
9	2.11	97.306885	0.000025	0.000220	523
10	2.32	92.311350	0.000067	0.000303	523
11	3.25	57.261173	0.000191	0.000473	523
12	2.79	62.190346	0.000169	0.000406	523
13	5.12	27.091051	0.000393	0.000818	434
14	4.46	32.134422	0.000318	0.000703	434
15	7.36	21.853570	0.000626	0.001149	434
16	11.78	17.150494	0.001816	0.002548	345
17	2.39	52.247010	0.000147	0.000263	523
18	**8.66**	**6.925114**	**0.011165**	**0.004900**	**255**
19	11.12	12.214928	0.002484	0.002045	245

Retailers characterized by very high Recency value with average levels of Frequency and Monetary values are seen in Clusters 3, 4, 7 to 12 and 17. These retailers have been inactive in the EFTPOS network for a long period of time and/or have low level of business activities overall. They constitute 21% of the EFTPOS market for the bank. The bank can see this type of retailers as growth area in the EFTPOS business sector. It however needs to determine if it is cost effective to launch aggressive marketing strategy to help improve the performance of these retailers. If it is decided that no aggressive campaigning is to be done for these retailers, due to the size of this market, some kind of marketing campaign will nevertheless still be required to maintain them.

6 Conclusion and Future Work

This paper proposes the use of clustering techniques to group retailers on an EFTPOS network based on the similarities in their business activities as characterized by how recent their business activities are, how frequent they conduct their business on the EFTPOS network and how much money their business activities have generated over a period of time. The preliminary results show that there are distinct combinations of RFM values of retailers in the clusters that may give the bank indications of the different marketing strategies that can be applied to each of the retailer types.

Further analysis into each cluster to find out more on the characteristics of the business and background of the retailers will help the bank in fine tuning their target marketing strategy for each retailer type. The next step of this project will be broken down into 3 categories. First, in observing if there are latent variables in the data set that may influence the variations in the volume of transactions in different days and possibly different periods in a day. Second, in trying out other clustering techniques to see of better quality clusters can be formed. Third, in building classification or causal models to find explanatory rules on the characteristics of each cluster using other attributes in the data set (e.g. the line of business a retailer is in, the business premise, etc.) and in exogenous variables like socio-demographic, advertising, social media data.

The data set used in this preliminary work is just from the first 2 weeks of our EFTPOS data extraction. We have since collected a few months of EFTPOS transaction data which will allow us to conduct more extensive Big Data analysis with more convincing results as a consequence. This will also open up new research avenues into the kinds of suitable big data computing techniques for market segmentation projects involving MapReduce/Hadoop system that we have put in place for this project.

7 Acknowledgments

The authors wish to thank Jon Scheele and Jo Spencer for their help in data acquisition and Blair Bethwaite for his help in setting up the secure computing environment to enable the execution of this commercial in confidence Big Data project.

8 References

Alam, S. et al., 2010. Particle swarm optimization based hierarchical agglomerative clustering. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. pp. 64–68.

Bizhani, M. & Tarokh, M.J., 2011. Behavioral rules of bank's point-of-sale for segments description and scoring prediction. *Int. J. Industrial Eng. Comput*, 2, pp.337–350.

Chen, D., Sain, S.L. & Guo, K., 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing \& Customer Strategy Management*, 19(3), pp.197–208.

Chen, Y.-S. et al., 2012. Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment. *Computers in Biology and Medicine*, 42(2), pp.213–221.

Davies, D.L. & Bouldin, D.W., 1979. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2), pp.224–227.

Dennis, C. et al., 2003. Market segmentation and customer knowledge for shopping centers. In *Information Technology Interfaces, 2003. ITI 2003. Proceedings of the 25th International Conference on*. pp. 417–424.

Doyle, C., 2011. *A dictionary of marketing*, Oxford University Press.

Dunn†, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), pp.95–104.

Gaur, D. & Gaur, S., 2013. Comprehensive Analysis of Data Clustering Algorithms. In *Future Information Communication Technology and Applications*. Springer, pp. 753–762.

Ho, G.T. et al., 2012. Customer grouping for better resources allocation using GA based clustering technique. *Expert Systems with Applications*, 39(2), pp.1979–1987.

Hsieh, N.-C., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, 27(4), pp.623–633.

Hughes, A.M., 2006. *Strategic database marketing*, McGraw-Hill.

Kim, Y. et al., 2005. Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, 51(2), pp.264–276.

Lee, J.H. & Park, S.C., 2005. Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications*, 29(1), pp.145–152.

Lefait, G. & Kechadi, T., 2010. Customer Segmentation Architecture Based on Clustering Techniques. In *Digital Society, 2010. ICDS'10. Fourth International Conference on*. pp. 243–248.

Li, J., Wang, K. & Xu, L., 2009. Chameleon based on clustering feature tree and its application in customer segmentation. *Annals of Operations Research*, 168(1), pp.225–245.

Namvar, M., Gholamian, M.R. & KhakAbi, S., 2010. A two phase clustering method for intelligent customer segmentation. In *Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on*. pp. 215–219.

Olson, D.L. et al., 2009. Comparison of customer response models. *Service Business*, 3(2), pp.117–130.

Salvador, S. & Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. pp. 576–584.

Singh, A., et al., Clustering Experiments on Big Transaction Data for Market Segmentation, in the Proceedings of the BigDataScience '14 Conference, August 04 - 07 2014, Beijing, China, ACM 978-1-4503-2891-3/14/08, http://dx.doi.org/10.1145/2640087.2644161 (in press)

Smith, W.R., 1956. Product differentiation and market segmentation as alternative marketing strategies. *The Journal of Marketing*, 21(1), pp.3–8.

Suib, D.S. & Deris, M.M., 2008. An efficient hierarchical clustering model for grouping web transactions. *International Journal of Business Intelligence and Data Mining*, 3(2), pp.147–157.

Yoon, S.-H. et al., 2013. A data partitioning approach for hierarchical clustering. In *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*. p. 72.

Zakrzewska, D. & Murlewski, J., 2005. Clustering algorithms for bank customer segmentation. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*. pp. 197–202.