Customer Analysis Model for Trusted Online Merchant

Osama Emam
Faculty of computers and Artificial Intelligence, Helwan University, Cairo, Egypt

Arafa Taher
Faculty of computers and Artificial Intelligence, Helwan University, Cairo, Egypt

Hanan Fahmy
Faculty of computers and Artificial Intelligence, Helwan University, Cairo, Egypt

ABSTRACT
Trust is one of the most important factors which influence the behavior of online shopping. Previous researchers indicated several factors that affect online brand shopping. Security, privacy, and trust are among these factors. This paper tries to propose a new model is called Merchant Segmentation Trust (MST) based on the huge amount of data comes from online shopping not on the last-mentioned factors using bigdata techniques. The propose model (MST) consist of four phases, the first phase is processing the data (cleaned, prepared, transformed) to extract the useful transactions attributes like order date, ship date, number of complaints, duration of complaints, net profit, etc. And aggregate all merchant with its transactions in form that can be used by the clustering algorithm. The second phase is to using the bigdata techniques as the huge amount of data comes from the rapid growth of online shopping by using K-means clustering algorithm. The third phase using traditional Recency, Frequency, Monetary (RFM) model to rank the resulting clusters to define the most trustable merchant, finally evaluate the clustering accuracy by using sum square error (SSE) and extract the most trustable merchant by the propose MST model compared against the traditional factors which result from the review or from the survey to ensure that merchant's behavior can rate the trust among other merchants.

Keywords
Bigdata; CRM; K-means clustering algorithm; RFM model; clv rank; sum square error (SSE)

1. INTRODUCTION
One of the most challenges faced by online merchants is to keep up consumers' interest and getting them to reuse their applications, revisit their e-shopping system and ultimately repurchase their products. Online sale systems designers and marketing staff invest great effort to introduce high quality solutions at the technical, informational, and customer service levels, but the performance of those investments, in terms of repeated purchase, requires: 1) that consumers find their shopping experience useful and enjoyable, and 2) that they trust and have a positive attitude towards shopping on the online site of a specific merchant.

The technology acceptance model (TAM) [1,2] discuss that determine using of information technology by the intention to use behavior, which, in turn, is explained by the attitude towards usage and perceived usefulness; attitude is conjointly formed by the beliefs of usefulness and ease of use. All other possible antecedents of intended use, such as technology properties or the individual or the organizational property, are all supposed to be fully mediated by beliefs.

It is very important to build trust with customers so that they will have repeat purchases, and make new customers to old

customers. On the one hand, this highlights the importance of useful activities for companies to improve customers' satisfaction, such as providing high quality goods, comfortable shopping environment and beautiful website, or establishing a good relationship with consumers. On the other hand, it is equally important for the companies to provide constructive information, build trustworthy relationships with consumers, which will support consumers with more information and resources about the goods, and ultimately help them with shopping satisfaction. Therefore, it is critical for online sellers or companies to provide various kinds of activities to enhance customers' trust and satisfaction.

Enterprise Success depends on customers and strategies used to know the behavior patterns of customer. Efficient analysis of customer's data, leads enterprise with new way of approach to develop their business process to meet customer requirements. In recent advanced technologies, the customer's data can be processed and transformed into useful form for business.

CRM is a business tool that is commonly used by many companies to develop their business process and make it success. CRM refers to strategies, practices and tools used to manage and analyze the data and interactions of customers with the aim of improving customer relations.

CRM can record all customer data into single database and distribute it into multiple departments. Providing quick access and manage of customer data to the business user.

This creates powerful analysis with customer actions and pattern with efficient and unique database, which allow and helps companies to make profitable.

In the online world today, web pages have the opportunity to save the attributes of their visitors' software and hardware environment and also some of their behavioral aspects. This data might contain valuable information about them [3].

Most companies have many tangible assets to support their businesses such as knowledge worker and business premise but the greatest asset that a company has is the customers [4], As the greatest asset for a company, customers' satisfaction is the main target for every company to achieve their goal in business. [5] Acknowledged the importance of intangible assets (satisfaction of customer) that needs to be highlighted in predicting future financial performance and not historical accounting measure that is currently used by many companies. Furthermore, every company must gather as much information as possible with meticulous knowledge about the customers' need [3] to make sure that they produce or market the product that customers are looking for. Thus, the ability of any company to understand their customers' needs, preferences, buying behaviors, and price sensitivity is a major advantage against their competitors [6]. It is very important

for every company to retain their loyal customers as [7] stated that a five percent increase in customer retention would result in 35 to 95 percent increment in average customer lifetime value, which is an easy profit for a company compared to investing to find new customers, whose loyalty are uncertain. Customer analysis, an analysis about customer behaviors and activities while transacting, is increasingly important among companies today. As a result, a lot of research on this topic has been done [4], [8], [5]. [10] Defined customer analysis as an analysis of customer activities and behaviors when doing a transaction with an organization or a company.

2. RELATED WORK

Online trust is influenced by many factors. Fig [1], researchers have incorporated factors such as ease of use [15] and usefulness to determine the degree in which these factors can influence the perception of online trust of customers. Researchers also focused on the issue of security and privacy [9,13], and also deployed constructs such as customization of websites, integrity of websites, responsiveness, visual appeal, and entertainment to show attractive design and functions of the websites [14, 15, and 16]. The E-word of mouth in the form of review and comments has also been proposed as one of key factors that influenced the online trust among the social media communities [3].
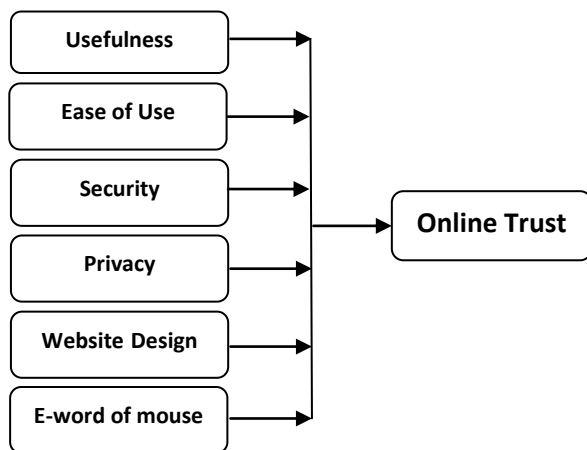

Fig [1] Factors Influencing Trust

In [12] Trust of internet shop-ping testing empirically a model of trust and perceived Risk Highlighted specific factors related to the vendor's perceived integrity, and perceived competence.

In [13] Trust Online brand testing a model of trust in online brand Highlighted security, privacy, brand name, word of mouth, good online experience, and quality of information that influence the online trust of bran significantly.

In [14] Online trust creation Developing and testing a model Highlighted transaction security, consumer data safety, guaranteed return policies, perceived image of website that has significant .Influence on trust creation.

In [15] Trust in ecommerce Vender Developing and testing a model Highlighted customization, brand image, and satisfaction that affected customer trust towards the vendor in m-commerce. Furthermore, the paper mentioned that customization and brand image have equally strong effect on

trust formation. On the other hand, no direct impact on interactivity and responsiveness but satisfaction influenced indirectly on trust towards the vendor.

In [17,18] in this model system quality (including customization, ease of navigation, privacy, and security), information quality (including dynamic content, content personalization and variety of information), and service quality (including responsiveness, empathy, and on-line support capabilities) separately and conjointly affect the consumer's satisfaction, intention, and use which ultimately affect net benefits for the e-merchant.

Another previous researchers also indicated several factors affect online brand trust. These factors are listed and summarized in Table [1]. Basically, the factors can be grouped into Web site factors and consumer factors.

Table [1] Factors Affecting Brand Trust

Reference	Factors Influencing Online Brand Trust
[13]	Web purchase-related factors: security, privacy, brand name, word-of-mouth,
[19]	Security, past experience and third-party recommendations
[20]	Web site factors: navigation, advice, no error, fulfillment, community, privacy/security, trust seals, brand and presentation. Consumer factors: self-confidence/Internet savvy, past behavior, Internet shopping experience, and entertainment experience.
[21]	Trustworthiness of Internet merchant, trustworthiness of Internet shopping medium and contextual factors
[22]	consumer characteristics, product/service characteristics, different markets/cultures/countries, perception of risk and past experience
[23]	Security, merchant legitimacy, fulfillment

3. MERCHANT ANALYSIS MODEL

This section introduces and explain the propose (MST) model Fig [2] for analyzing the merchant transactions based on web and purchasing patterns behavior represented in the Monetary, frequency, recency, time of transaction, number of complaints, time of complaints, state of transactions and duration of reply, to rank the trust of merchant among others.

Data sources that will be used, electronic company that have more than 250,000 transactions to more than 25 merchants Table [2], it is supposed that more recent and frequent user session and more number of clicks reflect how far customers are interesting to know about the company's products and are intended to make future purchase based on the trust. merchants' transaction data are preprocessed to get the most meaningful attributes to be included in K-means clustering algorithm for merchant value segmentation. Then rank the merchants based on the cluster ranking according to the traditional model (RFM) to understand their characteristics.

3.1 Data Pre-processing

Data in transaction data set are in unsuitable format for merchant segmentation. It needs to be cleaned, and extract related attributes affecting trust Table [1]. Then the extracted data will be transformed into an appropriate format for the K-means clustering algorithm. The preprocessing steps of the transaction data are explained as follow:

3.1.1 Transaction Data Pre-processing

In this step, the transaction data set is cleaned and the affecting trust attributes are selected. These attributes are transformed to (Monetary) which mean all the money the merchant get from the customers, (Num of Complaints) which mean number of complaints which the customer makes about the merchant's orders product, (Duration of Complaint) which mean the how time the merchant takes to reply to the customer, (time of session) which mean the duration of session which the customer takes to make order on its website, (Net profit) which mean the net gain of the all merchant, (Delay) which mean the numbers of delay days from the must ship date ,(Frequency) which mean the number of transactions per merchant and finally (Recency) the largest number of days the merchant don't receive orders.

3.1.2 Data Aggregation

In this step, the transactions dataset is containing more than one merchant. Each merchant contains its customer transaction, grouping all merchant transaction depending on the merchant column name and represent each merchant in one record by its attribute's values.

3.1.3 Data Normalization

The used transaction dataset present patterns whose attributes or features values within different dynamic ranges. In this case, features with large values will have a larger influence than those with small values. However, this not necessarily will reflect their importance for defining the clusters. The normalized value of the attributes is calculated based on the Min-Max normalization method [25] equation [1] as follow:

$$v'_i = \frac{v_i - minA}{maxA - minA} (newmax_A - newmin_A) + new_{min_A} (1)$$

3.2 Merchant Segmentation

transaction data set Algorithm [1].

Algorithm [1] Basic K-means algorithm
1: Select *K* points as initial centroids.
2: **repeat**
3: Form *K* clusters by assigning each point to its closest centroid
4: Recompute the centroid of each cluster
5: **until** Centroids don't change

Table [4] shows the five merchant segments resulted from using the transaction dataset attributes, each cluster is shown with the corresponding number of merchants, the average value of Monetary, Frequency, Net profit, Number of - Complaints, Time of session, Duration of Complaints, Delay, Recency attributes.
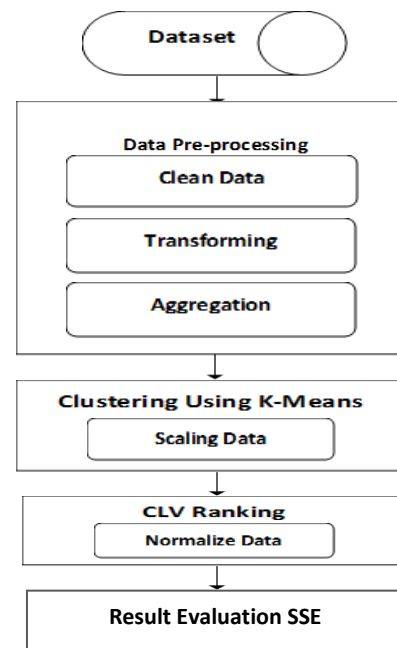

Fig [2] The Propose Model (MST)

Table [2] part of the used dataset

Order#	Order Date	Must ship Date	Ship Date	Amount	Merchant	Terms	Number Of Complaint	Duration Of Complaint	Time Of session	cost
20285021	9/5/2019	9/9/2019	9/9/2019	962	E-bay	CC	0	20.24	22.14	86.00%
20285020	9/5/2019	9/9/2019	9/9/2019	585	Walmart	N30	1	12.58	13.28	81.00%
20285019	9/8/2019	9/10/2019	9/9/2019	672	AAFES	N30	1	12.37	26.87	89.00%
20285018	9/5/2019	9/9/2019	9/9/2019	2844	Newegg	N45	1	27.10	21.00	85.00%
20285015	9/6/2019	9/10/2019	9/9/2019	489	Bluestem	N30	1	8.80	25.40	87.00%
20285009	9/6/2019	9/9/2019	9/9/2019	44.99	Amazon	N30	0	11.29	20.39	87.00%

After the preprocessing step, data is ready and suitable to be used for merchant segmentation into different value segments based on purchasing patterns. K- means clustering algorithm is used for merchant value segmentation using the normalized

| Cluster Number	Centroid
Cluster 1	[0.19949528,0.6534107]
Cluster 2	[0.7672818,0.232735]
Cluster 3	[0.1992047,0.03748469]

Cluster 4	[0.9843321,0.88780648]
Cluster 5	[0.30082299,0.31173377]

Merchants are segmented into five clusters **Fig [3]**, assigning one of the (upward ↑or downward↓) pattern to each one of the attributes. This assignment is based on a comparison between the value of each cluster with overall value of the five clusters.

For all of (Monetary, Frequency, Net profit) attributes if the value of each of them less than overall value of them on all clusters take downward (↓). For all of (Monetary, Frequency, Net profit) when the value of each of them exceed overall value of them on all clusters take upward (↑) Table [4].

For all of (Number of Complaints, Time of session, Duration of Complaints, Delay, Recency) attributes if the value of each of them less than overall value of them on all clusters take upward (↑). For all of (Number of Complaints, Time of session, Duration of Complaints, Delay, Recency) when the value of each of them exceed overall value of them on all clusters take downward (↓) Table [4].

Fig [3] Visualize the clusters using (K-means)
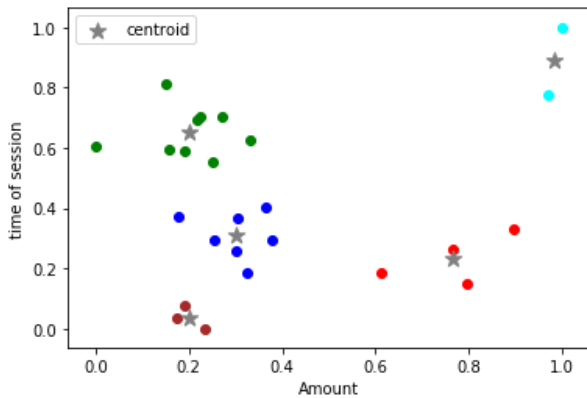
3.3 Ranking

The most powerful and simplest model to perform cluster ranking is the RFM model – Recency, Frequency, and Monetary value [25]. [26] defined RFM as: (1) R (Recency): the period since the last purchase; a lower value corresponds to a higher probability of the customers making a repeat purchase; (2) F (Frequency): number of purchases made within a certain period; higher frequency indicates greater loyalty; (3) M (Monetary): the money spent during a certain period; a higher value indicates that the company should focus more on that customer.

Average CLV value of each cluster can be calculated with equation (2) Table [3]:

$$CLV_{ci} = NR_{ci} * wR_{ci} + NF_{ci} * wF_{ci} + NM_{ci} * wM_{ci} \,(2)$$

NRci refers to normal Recency of cluster ci, WRci is Weighted Recency, NFci is normal Frequency, WFci is

weighted Frequency, NMci is normal Monetary, and WMci is weighted Monetary, considering weight equal to 1 to all columns according to equation (2).

After segmenting merchants, the average CLV is calculated based on it then a CLV rank is assigned to each segment.

C#	Merchants Number	NR	NF	NM	CLV Value
C1	16	0.109925	0.091995	0.767282	0.969202
C2	36	0.195569	0.179772	0.199495	0.574836
C3	28	0.193473	0.274482	0.300823	0.768778
C4	8	0.092650	0.026147	0.984332	1.103129
C5	12	0.095511	0.092081	0.199205	0.386797

Table [3]. Normal amount of RFM parameter in each cluster

3.4 Characteristics of the Clusters

- **Cluster 4** Rank 1 from the Table [4] and Fig [4] that has largest clv, which contains 8% of the merchant is the most profitable, and interested Merchant to the customers and has low frequent transactions with high monetary but spent a lot of time of session to make order with low number of complaints high response to the complaints that make this cluster's merchant more trustable (↑↑↑↓↑↓↓↑).
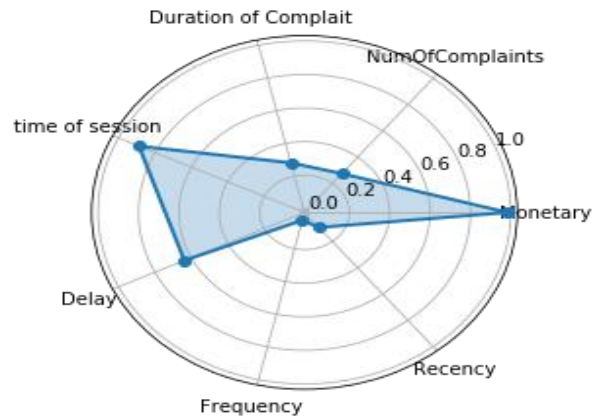

Fig [4] Cluster 4 characteristics

- **Cluster 1** Rank 2 from the Table [4] and Fig [5], which contains 16% of the merchant is more profitable, and has a lot of transactions with high monetary with low of time of session to make the order but high number of complaints with high response to the complaints and has purchased recently and more frequent transactions that make this cluster's merchant is trustable (↑↓↑↑↑↓↓↑).
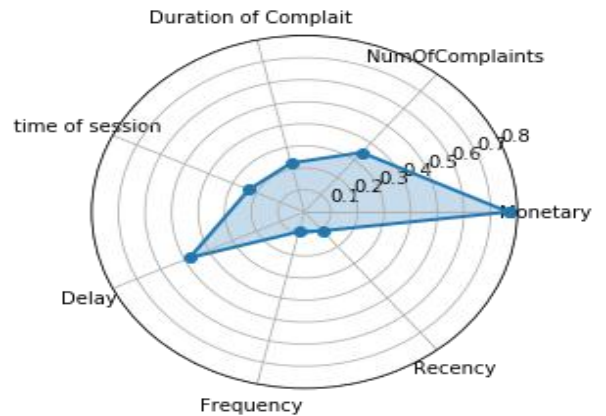

Fig [5] Cluster 1 characteristics

- **Cluster 3** Rank 3 from the Table [4] and Fig [6], which contains 28% of the merchant is low profitability, and

high number of complaints but with high response to the complaints, spent a lot of time of session to make order and more frequent transactions but with it has long time not make transaction between the year that make this cluster's merchant is less trustable (↓↓↑↑↓↑↑↓).
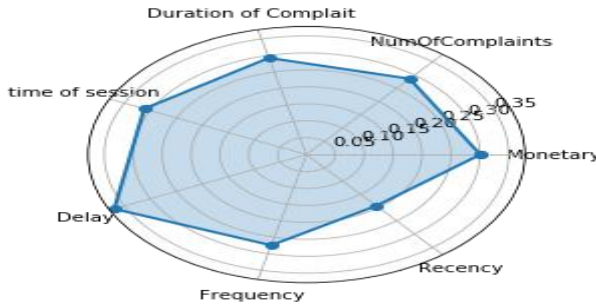

Fig [6] Cluster 3 characteristic

- **Cluster 5** Rank 5 from the Table [4] and Fig [8],which contains 12% of the merchant is low profitability, and has low frequent transactions low monetary and high number of complaints but high response to the complaints and spent a less time of session to make order that make this cluster's merchant less trustable(↓↓↑↑↓↑↓↑).

Table [4] The Characteristics of Five Clusters

C#	Monetary	Complaints Number	Complaint Duration	Time of session	Net profit	Delay	Frequency	Recency	CLV rank	pattern
C1	0.767282	0.345239	0.227598	0.232735	0.772512	0.478696	0.091995	0.109925	2	↑↓↑↑↑↓↓↑
C2	0.199495	0.209744	0.162441	0.653411	0.186903	0.410714	0.179772	0.195569	4	↓↑↑↓↓↑↑↓
C3	0.300823	0.286218	0.292619	0.311734	0.33562	0.369345	0.274482	0.193473	3	↓↓↑↑↓↑↑↓
C4	0.984332	0.274307	0.281224	0.887806	0.803866	0.638119	0.026147	0.09265	1	↑↑↑↓↑↓↓↑
C5	0.199205	0.31225	0.369276	0.037485	0.335773	0.286298	0.092081	0.09551	5	↓↓↑↑↓↑↓↑
Avg	0.490227	0.285552	0.266632	0.424634	0.486935	0.436634	0.132895	0.137426		

- **Cluster 2** Rank 4 from the Table [4] and Fig [7], which contains 36% of the merchant is more profitable, and has a lot of transactions with high monetary but spent a lot of time of session to make order with low number of complaints but high response to the complaints that make this cluster's merchant less trustable (↓↑↑↓↓↑↑↓).
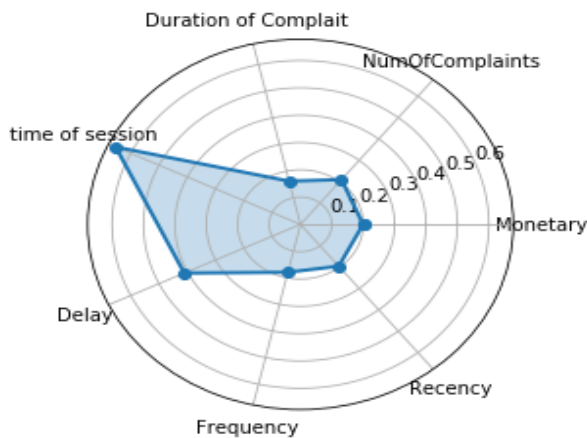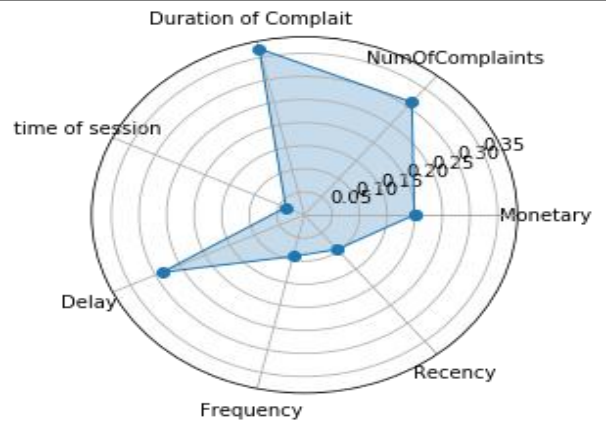

Fig [7] Cluster 2 characteristic

Fig [8] Cluster 5 characteristic

4. EVALUATION

This section explains how to evaluate and test the validity and accuracy of the merchant's segmentation, clustering analysis doesn't have a solid evaluation metric that can used to evaluate the outcome of different clustering algorithms, Moreover, since k-means requires k as an input and doesn't learn it from data.

4.1 Evaluate Clustering Accuracy

The elbow method Fig [9], in which the sum of squares at

each number of clusters is calculated and graphed is the optimal method to determine number of clusters.

As (k) increases, the sum of squared distance tends to zero. The graph Fig [9] shows that k=5 is not a bad choice. Sometimes it's still hard to figure out a good number of clusters to use because the curve is monotonically decreasing and may not show any elbow or has an obvious point where the curve starts flattening out

4.2 Evaluate Rankin

As mentioned before the propose model MST model trying to prove that the merchant's behavior transaction can rate the trust among other merchants as the traditional factors which result from the review or from the survey which mentioned in Fig [1] and Table [1].The dataset has an average of 10-Star rating for each transaction. Aggregate all the reviews to their merchant and normalize the data to be on the same scale with other attributes values. Notice that the highest review assigned to the first ranked cluster Table [5].
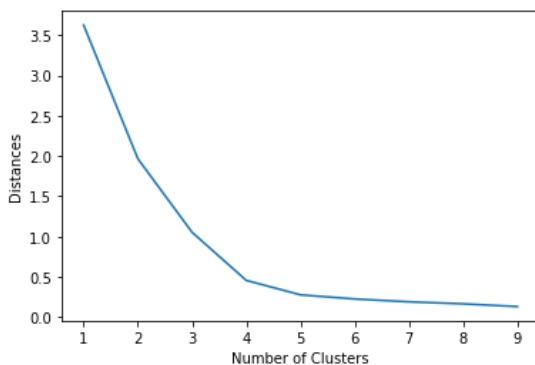

Fig [9] The "Elbow" Method

It can be concluded that reviews can be extract and rank the most trustable merchants from analysis the merchant transaction behavior by adding more attributes which meet the mentioned factored Fig[1], Table[1] and represent some of attributes in the Recency of session, Frequency of session, and Monetary to the traditional RFM attributes to ensure that the high ranking resulting clusters meet the high reviews.

Table [5] Ranking and Reviews Comparison

Cluster #	Percent Of Merchants	CLV Ranking	Reviews Range
C1	16%	0.969202	0.821429
C2	36%	0.574836	0.380952
C3	28%	0.768778	0.612245
C4	8%	1.103129	0.928571
C5	12%	0.386797	0.095238

5. CONCLUSION AND FUTURE WORK

Finally, this study aimed to extract the most trustable merchant. this work presents a new model called Merchant Segmentation Trust (MST) using bigdata techniques to extract and analysis the merchant's behavior and its characteristics and the degree of trust to the customer depending on the merchant transaction's behavior which the company already owned not in the factors or the review of the customers which save the company energy, effort and its capital.

The factors that influencing trust like usefulness factor can measure it by the number of transaction attribute and the (ease of use, website design) factors can be represented into time of session that that customer take to make the order and e-word of mouse factor can be act into the number of complains and the duration the merchants take to reply to the customer complains.

As a future work point, adding more attributes like type of payment, return of product, … etc. Which may enhance the accuracy of merchant's segmentation and dispense with review factors also adding a weight to the used attributes make us know the most influencing factors whether positive or negative.

6. REFERENCES

[1] F.D. Davis, R.P. Bagozzi, and P.R. Warshaw, "UserAcceptance of Computer Technology: A Comparison of Two Theoretical Models", Management Science, 35 (1989), pp.982-1003.

[2] J. Ingham, J. Cadieux, and A.M. Berrada, "E-Shopping Acceptance: A Qualitative and Meta-Analytic Review", Information & Management, 52 (2015), pp. 44-60.

[3] S. Kim, and H. Park. "Effects of various characteristics of social commerce (s-commerce) on consumers' trust and trust performance".International Journal of Information Management, 2013, 33(2), 318-332.

[4] N. C. G. Campbell and M. T. Cunnigham[1983]. Customer analysis for strategy development in industrial markets. Strategic Management Journal, 4:369–380.

[5] C. D. Ittner and D. F. Larcker [1998]. Are nonfinancial measures leading indicators of financial performance? an analysis of customer satisfaction. Journal of Accounting Research, 36:1–35.

[6] S. I. Chowdhury[2009]. A conceptual framework for data mining and knowledge management In H. Rahman, editor, Social and Political Implications of Data Mining: Knowledge Management in E-Government, pages 1–15. Hershey: IGI Global.

[7] F. F. Reichheld and T. Teal. The loyalty effect: the hidden force behind growth, profits, and lasting value. Harvard Business School Press, 2001.

[8] A. G. Parsons[2003]. Assessing the effectiveness of shopping mall promotions: Customer analysis. International Journal of Retail and Distribution Management, 31(2):74–79.

[9] Parvatiyar & Sheth (2001) defined CRM as ''a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers to create superior value for the company and the customer".

[10] V. Rainardi. Building a Data Warehouse with Examples in SQL Server. Apress, 2008.

[11] M. Hajli, "A research framework for social commerce adoption" .Information Management & Computer Security, 2013, 21(3), 144-154.

[12] R. Connolly and F. Bannister "Factors influencing Irish consumers' trust in internet shopping". Management Research News, 2008, 31(5), 339-358.

[13] H. Y. Ha "Factors influencing consumer perceptions of brand trust online". Journal of Product & Brand Management, 2004, 13(5), 329- 342.

[14] S. Sahney, K. Ghosh and A. Shrivastava, "Conceptualizing consumer "trust" in online buying behaviour: an empirical inquiry and model development in Indian context". Journal of Asia Business Studies, 2013, 7(3), 278-298.

[15] Y. S. Yeh, and Y. M. Li, "Building trust in m-commerce: contributions from quality and satisfaction. Online Information Review, 2009, 33(6),1066-1086.

[16] L. C. Harris and M. M Goode "Online services capes, trust, and purchase intentions". Journal of Services Marketing,2010, 24(3), 230-243.

[17] W.H. DeLone and E.R. McLean, "Measuring Ecommerce Success: Applying the DeLone & McLean Information Systems Success Model", International Journal of Electronic Commerce, 9 (2004), pp. 31-47.

[18] 2016 49th Hawaii International Conference on System Sciences "From E-Shopping System Quality to the Consumer's Intention to Return: A Meta-Analytic Study of the Mediation of Attitude, Usefulness, Enjoyment, and Trust".

[19] S. Srinivasan, Role of trust in e-business success,

[20] V. Shankar, F. Sultan, G. L. Urban, and I. Bart, The Role of Trust in Online Customer Support, Working Paper, Sloan School of Management, MIT, Cambridge, MA 02142, 2002.Knowledge Management in E-Government, pages 1–15. Hershey: IGI Global.

[21] M. K. O. Lee, and E. Turban, A Trust Model for Consumer Internet Shopping, International Journal of Electronic Commerce, vol. 6, no. 1, September 2001.

[22] J. Salo, and H. Karjaluoto, A conceptual model of trust in the online environment, Online Information Review, vol. 31, no. 5, pp. 604-621, 2007.

[23] S. Dayal, H. Landesberg, and M. Zeisser, How to Build Trust Online, Marketing Management, Fall, pp. 64- 69, 1999.

[24] Han, J., Kamber, M., & Pei, J., (2011). Data Mining: Concepts and Techniques, third ed. Morgan Kaufmann Publishers, Burlington, Massachusetts

[25] C.H. Cheng and Y.S. Chen, Classifying the segmentation of customer value via RFM model and RS theory, Expert systems with Applications (2009) 4176-4184.

[26] J.R. Bult and T. Wansbeek, Optimal selection for direct mail, Marketing Science 14 (1995) 378-39.

Information Management & Computer Security, vol. 12, no. 1, pp. 66-72, 2004.