Applying Data Visualization and Mining Techniques to Improvise Luxury Goods Promotion through E-Retail

Roma Chauhan
Institute of Management Education
G.T. Road, Sahibabad
Ghaziabad

Ritu Chauhan
Jamia Hamdard
Hamdard University
New Delhi

Alok Goel
Indian Institute of Technology
Roorkee
Uttarakhand

ABSTRACT

Data mining has been used by researchers and key organizations for variety of purpose to improve the business intelligence processes. The research paper explains the use of data mining to promote selling of luxury goods with in e-retailing value chain in order to facilitate process of data analysis and decision making. We have analyzed how launching a product in e-retail domain requires sufficient detailed results taken from effective data mining technique. In order to determine the regions with closed and heavily populated wealth clusters, clustering technique was applied. To achieve the desired objective first we have obtained the database of most affluent people across the globe and further applied clustering and data visualization technique to do some predictive analysis which can further help in focusing product marketing in e-retailing. The result achieved describes region fragmentation on basis of wealth and accordingly the e-retail concept can be implemented for those specific regions. This paper illustrates cutting edge advantages of product selling through e-retail and the challenges encountered to implement data mining to achieve the desired objective.

General Terms

Data mining, k-means, clustering, WEKA.

Keywords

Data Mining, e-commerce, business intelligence, WEKA, data visualization, clustering, K-Means.

1. INTRODUCTION

Due to intensive advent of technology the way market has approached customers they have undergone dynamic change, to bring the luxury shopping experience at par the retailers need to invest intelligently. The luxury retailers should provides better services, so they could reach mass audience. Luxury producers are laying down enormous efforts to improve prospects by using data mining techniques. Their main objective is to maintain physical stores and improve e-retail portals across the globe. According to research survey there are enormous efforts being put to increase luxury goods sales across the globe. Luxury retailers have ensured to improve the overall shopping experience by improving brand and product selection, store atmosphere, and customer services.

The consumption pattern of luxury goods by the customer within the country is depended on number of factors such as: liberalization of economic policies, buying habits of the younger generation, financial independence at a young age and increase in media exposure to the people. The tastes and preferences of the current generation are changing rapidly; changing consumption patterns trigger changes in shopping styles of customers and also the factors that drive people into stores [1]. The eager customer does not mind paying extra for better facilities and ambience. The market for luxury goods is surely climbing at an astonishing rate as compared to what it was a decade ago which was almost negligible. The paper illustrates how data mining techniques can be further utilized for the purpose of goods promotion and summarizes core issues of innovation in e-retail to transform it into digital luxury.

The challenge here is to provide goods to the high end customers across the globe including regions such as Asia, Europe, United States, Middle East and others countries through web portals. We are using data mining techniques to understand billionaire's segregation in these geographic regions. A retailer on large scale collects huge volume of data. Before actually an e-retailer starts promoting or selling the product online, the brand marketing strategy needs to be utilized. To achieve the desired goal we in our paper has recommended data mining practice in order to analyze the huge database with e-retailers and how can they further target the geographical regions for product promotion and sale. In this paper, we would be trying to answer critical questions on how e-retail: (1) Can gain competitive edge over the existing retailers and (2) advantages of e-retail over conventional retail.

The rest of the paper is organized as follows. Section 2 briefly gives background explaining drift in retail business from traditional to online. In Section 3, we describe the literature review. Section 4 explains the role of data mining and the challenges encountered and in Section 5 we discuss clustering using K-Means. Section 6 explains how tool WEKA is used for clustering. Finally conclusion is covered in Section 7.

2. E-RETAIL: THE CHANGING DYNAMICS

The customer buying patterns also depends on change in marketing strategies of companies with change in consumer buying behavior. With change in consumer buying behavior the companies also made necessary changes in their marketing strategies [2] opinioned that the customer relationship management unites the potential of marketing strategies and IT to create profitable, long-term relationships with customers and helps in enhancing the opportunities to use data and information to both understand customers and co-create value with them. The companies recognize that customer relationships are the underlying tool for building customer value, and they are finally realizing that growing customer value is the key to increasing enterprise value [16]. The companies tend to design their goods on the basis of market segmentation so that they have goods to

suit every pocket and requirement. The primary reason for shift of interest of customer from physical stores to online retail stores is due to:

- Internet accessibility across the globe has become inexpensive and easy.
- Today's consumer focuses more on technology and credit purchase.

With the advancement in the current internet technology and convenience driven retail environment, e-commerce poses an indispensable complimentary distribution channel for offline retailers. Retailers seeking new forms of differentiation are turning to e-retail to tap the market constituting of millions of people that make up the online consumer population. More interestingly, Internet penetration in several countries is growing at an alarmingly high rate. E-retail also influences offline store visits.

The demand for global luxury online sales is on the increase. Recent reports indicate that the wealthy are almost all online and are pleased with making online purchases. In most developed economies, Internet penetration is as high as 95% and the ratio of wealthy people who have bought goods is also significantly very high. This has given rise to the question of selling luxury goods online.

Website and e-store design seek to achieve more than basic, functional requirements such as providing a conducive and pleasant shopping experience. E-retail involves a constant flow of innovative means of differentiation that will meet the expectations of the online consumer in order to generate more online traffic and maintain customer loyalty. The challenge of selling luxury goods online is enormous. The luxury shopping experience is different from the conventional shopping experience. Selling luxury goods online is a challenging task and the customer purchasing the good would like to touch, feel or smell the product before actually paying for it. This often requires physical store presence, which is absent in the online virtual environment.

The major drawback with the Internet is that it also lacks the exclusive and prestigious locations where the luxury stores are situated. Therefore the question of creating a prestigious online atmosphere, replacing the human senses in the virtual environment and matching 'high class' with the 'mass class' of the Internet world is a challenge to be achieved. Online luxury consumers have high expectations and this includes their belief that although the luxury e-boutique is available to the masses, it should be designed to feel right to only a select few. Fortunately, with feasible strategies, this could be possible. Strategic luxury e-retail design and planning involves the utilization of certain key elements to transfer the 'looks and feel' of sensory goods and the prestigious atmosphere to the e-boutique virtual environment.

3. LITERATURE REVIEW

Data is growing tremendously in gigabytes or terabytes. To deal with massive explosion of data we require proper analyses techniques, so the hidden information in large databases can be strategically analyzed. But the question is how enormous collection of data can be used to draw meaningful conclusions about a particular domain? The solution is data mining which is basically a technique to transfer these data into relevant information. Data mining also known as knowledge discovery in data base, is the process to extract the relevant information from the database. It commonly deals in a wide range of profiling practices such as marketing, Surveillance fraud detection, human factor related issue and scientific discovery. Over the last decade, data mining technology has made impact in many different domains. It is not only used by business organization to collect the business information but it is used in the science and technology to extract information from the enormous data set generated by modern research methods and observations. It is estimated that the digital universe consumed approximately 281 exabyte in 2007, and it is projected to be 10 times that size by 2011.(One exabyte is ~1018 bytes or 1,000,000 terabytes) [3].
Data mining is defined as a knowledge discovery process which helps in discovery of hidden patterns within the database. It is opinioned that the Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of statistics, machine learning and database management systems [4]. Data mining as a knowledge discovery technology has matured and is being adopted by many businesses to enhance their efficiency; technique involves scientific, precise and systematic analysis of data. The technique also helps in future prediction which facilitates process of decision making. Data Mining is a process for sifting through lots of data to find information useful for decision making [5]. The development and the validation of a decision tree, which aims to discriminate between good and bad accounts of the customers of a particular retailer based on a sample of orders placed between certain periods of time was described by [15]. The mining would be able to answer how much sales company is likely to do next year and why?

4. DATA MINING CYCLE AND CHALLENGES

Reflecting this conceptualization of data mining, some observers consider data mining as extended one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation [6]. Data mining (knowledge discovery from data) is Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Figure 1 explains the complete working cycle of data mining.

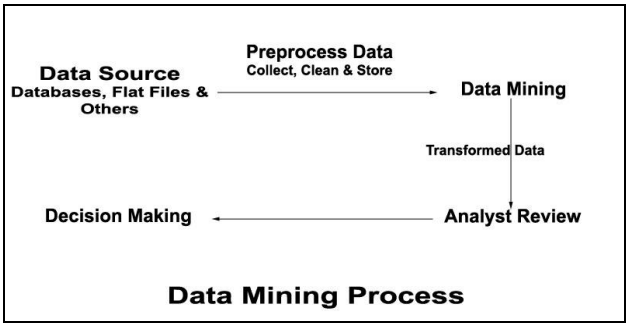

Figure 1 Data Mining Process Description

The initial step in data mining process is collection of data from multiple data sources. The data is further cleaned and stored for further processing. The data mining engine transforms the data into productive knowledge which can be further utilized by analyst in process of decision making.

The research paper explains several challenging problems based on our experiments and analysis in implementing mining techniques to promote goods within e-retail chain. The acceptability of the software was primarily for CSV or ARFF file format. The data was complex to be analyzed. The other factors that lead to complexity are such as:

- Scalability: Very high data volumes
- Complex, structured, semi-structured, and unstructured data
- Data extraction, cleaning, and consolidation from many sources
- Integrate data warehousing, on-line analytical processing (OLAP), and data mining.
- Integrate into complete solutions such as: use results of analysis and mining for decision making, e.g., marketing campaigns, adapting business processes, supply chain optimization

5. TECHNIQUE USED

Clustering can be used as a stand alone tool for analyzing data and may also be used as a pre processing step in other data mining algorithms. Clustering methods can be classified in a number of ways [7]. In clustering of data items present in data sets are first grouped according to similarity between them that can be also based on geographic regions. For doing clustering we have made use of simple K-Means algorithm. Clustering algorithm provides information about data set in a summarized form. It is basically an unsupervised form of learning. The basic objective behind clustering is grouping of similar data sets together. A single cluster will contain items those are similar to each other. The objective of clustering algorithm is to maximize intra cluster similarity and minimize inter cluster similarity. The objects within data sets are close to each other if they belong to same cluster. The Euclidean formula is used for calculating distance.

The primary reason to use K-Means clustering algorithm for analysis is because it provides faster computation in comparison to hierarchical clustering. It produces neat and clean clusters which make the analysis process precise. K-means is an exclusive clustering algorithm it generates a specific number of disjoint, non-hierarchical clusters, this type of algorithm data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. For example the separation of points is achieved by a straight line on a bi-dimensional plane.

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The following steps are involved in k – means algorithm

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids assume k is clusters
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The simplified clustering algorithm is explained in Figure 2. Simply put, k-Means Clustering is an algorithm among several that attempts to find groups in the data [9]. In the new initialization method, the clustering algorithm will only be performed for several iterations during each run. After each run, initial points, which can be used to form the cluster with good structural similarity, are chosen and their distance is checked against that of all points already selected in the initialization array. If the minimum distance of new points is greater than the specified distance, these points will be added to the initialization array [8].

```
Input:
   D = {t_1, t_2, ..., t_n}   // Set of elements
   A      // Adjacency matrix showing distance between elements.
   k      // Number of desired clusters.
Output:
   K      // Set of clusters.
K-Means Algorithm:
   assign initial values for means m_1, m_2, ..., m_k ;
   repeat
      assign each item t_i to the cluster which has the closest mean ;
      calculate new mean for each cluster;
   until convergence criteria is met;
```

Figure 2 K-Means Clustering Algorithm

6. K-MEANS CLUSTERING USING WEKA

WEKA[1] is open source software developed using JAVA language is a collection of algorithms for data mining. It is business intelligence software which provides clustering and data visualization tools. The tools provides user friendly interface for use and allows data visualization that make visualizing and understanding data set easy.

Clustering is also used to group customers into different types for efficient marketing [10]. We would be working on dataset Billionaires 1992, dataset which is available for free use. Fortune magazine publishes the list of billionaires annually. The 1992 list included 233 individuals or families. Their wealth, age and geographic location (Asia, Europe, Middle East, United States or

[1] http://www.cs.waikato.ac.nz/ml/weka/

Other) is reported covering number of 233 cases. In this information age the size of data that an organization holds is not the major concern but to what extent that huge data is utilized is the top priority of companies these days. Data mining is computer based analysis method since human analyst cannot all alone process the enormous volume of data present in enterprise

The technique of applying data mining to business processes has advanced due to availability of open source software in the market. In this paper we will be using WEKA (Waikato Environment for Knowledge Analysis) by University of Waikato. We will be using .arff file format which is one of the file formats supported by the tool. WEKA contains "clusters" for finding groups of similar instances in a dataset. Implemented schemes within the software are: k Means, EM, Cobweb, X means, Farthest First. Clusters can be visualized and compared to "true" clusters (if given). The ARFF File Format is the specific file format used by the tool. The following is the format of when implemented through WEKA. For the present data set we are using following as variable names for the tool [11]:

1. wealth: Wealth of family or individual in billions of dollars

2. age: Age in years (for families it is the maximum age of family members)

3. region: Region of the World (Asia, Europe, Middle East, United States and Other)

6.1 Experimental Results

The K-means algorithm requires three user-specified parameters: number of clusters *K*, cluster initialization, and distance metric. The most critical choice is *K*. The result after applying K-Means using WEKA on the dataset is summarized in Figure 4 and the clusters are obtained.

As from the result (Figure 3) we notices that distance between two points (U, E) (A, O) and (M, O) are also less. Therefore, they can be considered as one cluster for attributes wealth and age. The clustering centroids are obtained as (2.6, 64), (2.5, 50) and (2.7, 71). The last two clusters are of United States and Europe which indicates that on an average both the regions have maximum number of wealth in billion of dollars.

```
Scheme:       weka.clusterers.SimpleKMeans
-N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     relation
Instances:    233
Attributes:   3
              wealth
              age
Ignored:
              region
Test mode:    Classes to clusters evaluation on training data
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 18
Within cluster sum of squared errors: 3.9139074509677974
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute   Full Data       0       1
                (233)     (84)    (149)
==========================================
wealth        2.6815    2.5274    2.7685
age          64.0311   50.6667   71.5654


Clustered Instances

0      84 ( 36%)
1     149 ( 64%)

Class attribute: region
Classes to Clusters:

  0  1  <-- assigned to cluster
 13 25 | A
 30 50 | E
  8 14 | M
  9 20 | O
 24 40 | U
Cluster 0 <-- U
Cluster 1 <-- E
```

Figure 3 Clustering using K-Means

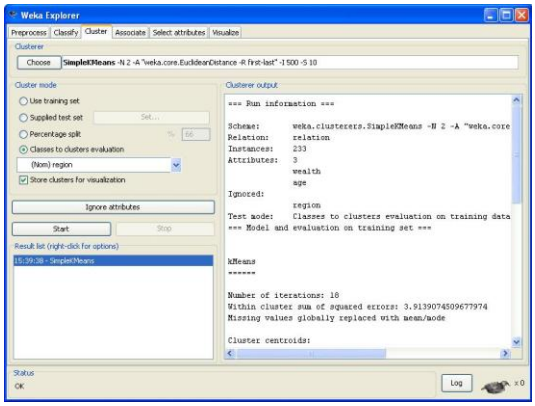

Figure 4 Clustering Screen

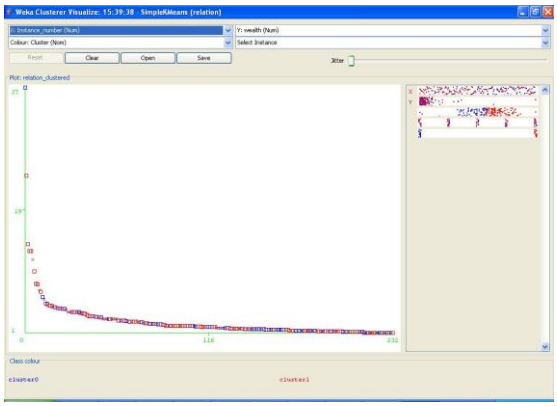

Figure 5 Cluster Visualized

The Figure shows the clusters formed and clusters visualized as in Figure5. The numbers of instance taken into account were 233 with three attributes. The mean value was 2.682 and standard deviation of 3.319. The result was broadly divided into two clusters, cluster 0 and 1. Using WEKA incorrectly clustered instances: 159.0 68.2403 %

Data visualization is one of the most required methods that influence the performance of the clustering algorithm. If the representation of the clustered datasets is good, the clusters are likely to be compact in structure and isolated and even a simple clustering algorithm such as K-Means will find them. Figure 5 shows representation of dataset where K-Means actually fails to partition the dataset into natural obvious clusters. Therefore, visually categorizing the datasets into clusters appears to be a tedious task. [12] Used the minimum message length (MML) criteria [13][14] in conjunction with the Gaussian mixture model (GMM) to estimate *K*. Their approach starts with a targeting large number of clusters, and gradually merges the clusters leading to decrease in distance. The mentioned technique can be used to get better clusters. Using mathematical calculations we have tried to mathematically club together closest of points to form a cluster.

7. CONCLUSION

Although challenging, but it is definitely possible to reach the desired consumer of the luxurious and prestigious store atmosphere of luxury brands through the Internet virtual environment and it is possible to sell luxury fashion goods online. The appropriate result shows that which regions across the globe have majority of wealth so that they can be targeted as proper destinations to market initially. The elements of e-retail success are also depended upon increasing online traffic, website stickiness and online sales turnover. Also clustering results as in Figure 4 and Figure 5 shows that cluster of one region is close to other and therefore the clusters can be clubbed together within a group, not producing much of disparity among the classified regions. The typical challenge to apply any of the mining techniques such like clustering primarily in advance a user should be able o determine that whether the dataset has the capacity to undertake and show the clustering results.

8. REFERENCES

[1] Kaur, P. and Singh, R. (2007), 'Uncovering retail shopping motives of Indian youth', Young Consumers: Insight and Ideas for Responsible Marketers, Vol8, No.2, and pp.128-138.

[2] Sangle, P.S. and Verma, S. (2008) 'Analysing the adoption of Customer Relationship Management in Indian service sector: an empirical study', International Journal Electronic Customer Relationship Management, Vol. 2, No.1, pp.85-99.

[3] Gantz, John F. 2008 (March). The diverse and exploding digitaluniverse. Available online at: http://www.emc.com/collateral/analyst-reports/diverseexploding-digital-universe.pdf.

[4] Feelders, A., Daniels, H. and Holsheimer, M. (2000), Methodological and Practical Aspects of Data. Mining', Information and Management, Vol. 37, Issue 5, pp.271-281.

[5] Noonan, J. (2000), 'Data Mining Strategies', DM Review

[6] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques (New York: Morgan Kaufmann Publishers, 2001), p. 7.

[7] Bradley, P.S., Fayyad, U.M., Reina, C.A., "Scaling EM Clustering to Large Databases", Microsoft Research Technical Report 98-35, 1998.

[8] Zhong Wei, et al. "Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property." IEEE Transactions on Nanobioscience. Vol. 4. No. 3. September 2005. pp. 255-265.

[9] Alpaydin, Ethem. Introduction To Machine Learning. Cambridge, Massachusetts: MIT Press. 2004.

[10] Arabie, P., and Hubert, L. 1994. Advanced methods in marketing research. Oxford: Blackwell. Chap. Cluster Analysis in Marketing Research, pages 160– 189.

[11] Fortune, September 7, 1992. "The Billionaires." pp. 98- 138

[12] Figueiredo, Mario, and Jain, A K. 2002. Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3), 381–396.

[13] Wallace, C.S, and Boulton, D. M. 1968. An information measure for classification. Computing Journal, 11, 185–195.

[14] Wallace, C.S., and Freeman, P.R. 1987. Estimation and inference by compact coding (with discussions). JRSSB, 49, 240–251.

[15] Sarantopoulos, G. (2003), 'Data mining in retail credit', Operational Research, Springer Berlin / Heidelberg, Vol. 3, No. 2, pp. 99-122.

[16] Rogers, M. (2005), 'Customer strategy: observations from the trenches', Journal of Marketing, Vol. 69 No.4, pp.262.