BIG DATA ANALYSIS FOR RETAIL INDUSTRY AND DATA MINING TECHNIQUES

P.Manikandan*

S.Yuvarani*

Dr.C.Jothi Venkateswaran**

ABSTRACT

This research paper is analyzing big data from retail industry. The retail industry provides an exciting way of human life for their livelihood in this sector of the Indian economy. Retailers provide the goods and services from food, auto parts, apparel, home furnishings, appliances, and electronics to advice, home improvement, and skilled labor. The Big data collected from Supermarkets, Super centers, Hard-lines stores, Discount stores etc… The data mining techniques are used wherever large volumes of big data need to be processed for decision support Retail industry.

Keywords: Big data; retail industry; data mining techniques;

* Asst.Professor, Department of Computer Application, Thanthai Hans Roever College, Perambalur -621212

** Asst. Professor & Head, Department of Computer Science, Presidency College, Chennai-600 0005

I. Introduction

Data Mining (DM) is a well honored field of Computer Science. Data Mining aims to discover Valid, complex and not obvious hidden information from large amounts of data. Data mining techniques are the processes designed to identify and interpret data for the purpose of understanding and deducing actionable trends and designing strategies based on those trends [6].

A retailer or retail store is any business enterprise whose sale volume comes primarily from retailing. These are the final business entities in a distribution channel that links manufacturers to customers. Manufacturers typically make products and sell them to retailers or wholesalers. Wholesalers resell these products to the retailers and finally, retailers resell these products to the ultimate consumers.

Any organization selling to final consumers whether it is a manufacturer, wholesaler or retailer-is doing retailing. It does not matter how the goods or services are sold (by person, mail, telephone, vending machine, or internet or where they are sold-in a store, on the street, or in the consumer's home). A Retailer thus, provides value creating functions like assortment of products and services to the consumers, breaking bulk, holding inventory and provides services to consumers, manufacturers and wholesalers.

Retailing broadly involves:

1. Understanding the consumers' needs.
2. Developing good merchandise assortment.
3. Display the merchandise in an effective manner so that shoppers find it easy and attractive to buy.

II. Review of literature

This Study classifies existing retail industry based on data mining technique. Each data mining technique can perform one or more of the following types of data modeling:

➢ **Association** intends to determine relationships between attributes in databases (Mitra et al., 2002; Ahmed, 2004; Jiao et al., 2006). The focus is on deriving multi-attribute correlations, satisfying support and confidence thresholds [1]. Examples of association model outputs are association rules. For example, these rules can be used to describe which items are commonly purchased with other items in grocery stores.

➢ **Classification** aims to map a data item into one of several predefined categorical classes (Berson et al., 1999; Mitra et al., 2002; Chen et al., 2003; Ahmed, 2004). For example, a classification model can be used to identify loan applicants as low, medium, or high credit risks [2].

➢ **Clustering,** similarly to classification models, aims to map a data item into one of several categorical classes (or clusters). Unlike classification in which the classes are predefined, in clustering the classes are determined from the data. Clusters are defined by finding natural groups of data items, based on similitude marks or probability bulk models (Berry and Linoff, 2004; Mitra et al., 2002; Giraud-Carrier and Povel, 2003; Ahmed, 2004). For example, a clustering model can be used to group customers who usually buy the same group of products [3].

➢ **Forecasting** estimates the future value of a certain attribute, based on records' patterns. It deals with outcomes measured as continuous variables (Ahmed, 2004; Berry and Linoff, 2004). The central elements of forecasting analytics are the predictors, i.e. the attributes measured for each item in order to predict future behavior. Demand forecast is a typical example of a forecasting model whose predictors could be for example price and advertisement.

➢ **Regression** maps a data item to a real-value prediction variable (Mitra et al., 2002; Giraud-Carrier and Povel, 2003). Curve fitting, modeling of causal relationships, prediction (including forecasting) and testing scientific hypotheses about relationships between variables are frequent applications of regression.

➢ **Visualization** is used to present the data such that users can notice complex patterns (Shaw, 2001). Usually it is used jointly with other data mining models to provide a clearer understanding of the discovered patterns or relationships (Turban et al., 2010). Examples of visualization applications include the mind maps [4].

III. Big Data Analysis from Retail Industry

The Big Data and analytics hold the potential to dramatically increase three imperatives of retail industry

A) Ownership

B) Strategy-mix

C) Service vs. Goods retail mix.

A) OWNERSHIP BASED RETAILERS:

Depending on the ownership pattern, stores can be divided into six categories as:

1. Independent Stores

i. Owned by a single retailer-

ii. Low entry barriers

iii. Low initial investments

iv. Simple licensing procedures

v. Owner holds the right to decisions

vi. Can act as specialized stores

2. Chain stores

i. Have two or more retail outlets

ii. Common ownership & control

iii. Centralized purchase & merchandising.

iv. Sell similar lines of merchandise

v. Bulk purchaser, high bargaining power

vi. E.g.: Bata, Liberty, Kodak, Archies, Titan, Raymonds, LG, McDonald's, Barista etc

3. Franchise stores

i. Store based on contractual agreement between a Franchiser (manufacturer) & a Franchisee, which allows the franchisee to conduct a given form of business under an established name & according to a given pattern of business.

ii. Franchisee gets well known brands.

iii. Exclusive rights to sell.

iv. Benefit of the nationwide promotional activities.

v. Exposure to standard operating procedures.

vi. E.g.- Aptech, McDonald's, Monte Carlo, Koutons, Pizza-Hut,etc

4. Leased Departmental store

i. A department in a retail store that is rented to an outside party.

ii. The lessee is accountable for all activities of the leased department.

iii. Adds variety to the merchandise offered by the store.

iv. No cannibalizing of sales of existing product lines of the stores.

v. Reduced cost of establishment.

vi. Increased customer traffic.

5. Vertical Marketing system

i. A distribution system in which the producers, wholesalers, & retailers act in a unified manner to facilitate the smooth flow of goods & services to the end-user.

ii. One channel member owns the other or has contracts with them.

6. Consumer Cooperatives

i. Retail operations owned & managed by its customer members.

ii. A group of customers invest in the retail operations in return of stock certificates, which entitle them to a share in the profits of the retail store.

B) STRATEGY-MIX BASED RETAILERS

Depending on the strategic mix retailers adopt, they can be classified into two groups:

I. Food oriented Retailers-

1. Convenience stores-

i. Small stores located near residential areas.

ii. Open long hours, seven days a week, & carry a variety of products with limited assortment of merchandise.

iii. Operate in 3000-8000 sq.ft. area.

2. Conventional supermarkets-

i. Similar to departmental stores.

ii. Focus on food & household maintenance products.

iii. Self-service operations

iv. Variety of merchandise with deep assortments

.

3. Food-based supermarket-

i. Larger & more diversified than a conventional supermarket.

ii. Operates in 25,000-50,000 sq.ft. area.

iii. Range includes- grocery items, garden supplies, flowers, & small household appliances.

4. Combination Stores-

i. A Blend of a supermarket & general merchandise (>40%) store.

ii. Maintains identity of both food store & drug store

iii. One stop shopping experience.

iv. Operate in 30,000-100,000 sq.ft. area.

5. Box (Limited-line) store-

i. Food based discount store that concentrates on a small selection of goods.

ii. Limited- shopping hours, service & stocks.

iii. Refrigerated perishable goods are not available.

iv. Prices are displayed on the shelf/overhead signs. Priced 20-30% below market price.

v. Self-service.

6. Warehouse stores-

i. Discount food retailers, offer low price deals.

ii. Average size of 100,000 sq. ft.

iii. Merchandise is displayed in cut boxes or shipping pallets & services are limited.

iv. Lack consistency of products available as they warehouse retailers buy goods only when a manufacturer or a wholesaler offers deep price or quantity discount.

II. General Merchandise retailers-

1. Variety store-

i. Offer deep assortment of inexpensive & popular goods like stationary, gift items, woman's accessories, house wares etc.

ii. Also called as 5 and 10-cent stores.

2. Department store-

i. Large retail units offering wide variety and a deep assortment of goods & services.

ii. Separate depts. for separate types of merchandise

iii. One-stop shopping experience.

iv. Offers clothing, shoes, cosmetics, gifts, luggage, jewelry & other household items.

v. E.g.- Shopper's stop, Westside, Lifestyle & Pantaloons etc.

3. Off-price Retailer-

i. Offer an inconsistent assortment of branded fashion-oriented soft goods at low prices.

ii. Purchase goods from manufacturers who have excess inventory.

iii. Purchase in bulk & sell at off-prices.

4. Membership club

i. Customer has to pay annual fee to become the members of the club.

ii. Membership allows them to purchase goods at low price.

iii. Purchases directly from manufacturers.

5. Flea Market (Outdoor Bazaar)

i. An outdoor or indoor facility that rent out space to vendors who offer merchandise, services & other goods.

ii. Many retail vendors offering a variety of products at discount prices at places where there is high concentration of people.

C). SERVICES Vs GOODS RETAIL MIX BASED :

❖ **Service Retailing-**

The retail entities primarily selling services rather than products are in retailing of services. Services also play a significance role in the retail merchandise mix of the retail organization selling merchandise as a core product. The main differences between retailing of products and retailing of services are on account of the intangibility, simultaneous production and consumption, perishability and inconsistency.

i. Sale or rental of an intangible activity, which usually can not be stored or transported, but satisfies the needs of the user.

ii. Service can be along with goods or pure service.

iii. E.g. - Hospitals, banks, beauty saloons, entertainment firms etc.

❖ **NON-STORE RETAILERS**

I. Traditional-

1. Direct Marketing

i. Customer is informed about the product through non personal medias like TV, radio, magazine, newspaper, internet etc.

ii. The customer places an order through mail or phone.

iii. Less investment as compared to store based retailing.

iv. Wide geographic area is covered.

2. Direct selling

i. Door-to-door selling

ii. Person-to-person selling (Eureka Forbes)

iii. Multilevel (network) marketing (Amway)

3. Vending machines

i. Involves coin or card operated dispensing of goods & services.

ii. Round the clock sales

iii. Machines are placed at the most convenient places for the customers

iv. Soft drinks vending machines, ATMs,coffee vending machines etc.

4. Catalog marketing

i. Sales made through catalogs mailed to a select list of customers or made available in a store.

ii. Delivery or order can be through mail, express service, and parcel post.

5. Tele-Marketing

i. Telephone as a media for sales.

ii. Informing customers about new merchandise & upcoming sales events.

6. TV home shopping

i. Shop – by- TV, demonstration of the product, its features, benefits etc (Asian sky shop, tele-shopping etc.)

II. Nontraditional:

1. e-tailing or World Wide Web

i. Internet as a medium for promoting their products.

ii. People access information about products using the web address of the retailer's homepage.

iii. Retailers' website allows customers to order with a click of mouse.

2. Video kiosk

Freestanding interactive computer terminal that displays product & related information on a video screen, are often touch screen.

3. Video catalog

A retail catalog on a CD-ROM disk, to be viewed on a computer monitor

IV. Retail industry in data mining Techniques

Data mining techniques can be applied to various areas like marketing, finance and sales. The tools are used wherever large volumes of data need to be processed for decision support Retail industry collects large amount of data on sales and customer shopping history. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce.

Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business.

Data mining techniques for the various data mining tasks described above. Some of the most commonly used techniques are described as follows.

- **Rule Induction and Statistical Analysis**

Rule induction is the process of looking at a data set and generating patterns. By automatically exploring the data set the induction system forms hypotheses that lead to patterns. The process is in essence similar to what a human analyst would do in exploratory analysis. For example, given a database of demographic information, the induction system may first look at how ages are distributed, and it may notice an interesting variation for those people whose profession is listed as professional athlete.

- **Decision Trees and Neural Nets**

A decision tree is a technique for partitioning data into a set of rules that represent decisions. These decisions generate rules for the classification of a data set. A decision tree consists of nodes and branches. The beginning node is called a root. Depending upon the results of a test the data is partitioned into various subsets. The end result is a set of rules with all possibilities. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) (Kamble, 1999).

Neural Nets are non-linear predictive models and are inspired by the human brain. They learn through training and resemble biological neural networks in structure. They are better suited for financial applications and medical diagnosis (Kamble, 1999). Both decision trees and neural nets are effectively equivalent, except that neural nets typically use some form of parallel processing, while decision trees are linear.

- **Fuzzy Logic and Genetic Algorithms**

Unlike the yes-no system of conventional logic, fuzzy logic assumes a continuum of truth values between 'completely true' and 'completely false'. It can be used as a means of generating probabilistic analyses of data. Genetic algorithms take models derived from the natural world - for example, evolution, inheritance and epidemiological activity - and apply them to data. A model can be overlaid on data to see if the data fits (Herman, 1997).

Genetic algorithms are optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design, based on the concepts of evolution (Kamble, 1999). Genetic algorithms are not just for rule generation and may be applied to a variety of other tasks to which rules do not immediately apply, such as the discovery of patterns in text, planning and control, and system optimization, etc. (Holland, 1995).

- **Clustering and Correlation**

Clustering (or segmentation, which is effectively the same thing) is one of the first steps in analyzing an undifferentiated body of data. It is used to produce first-approximation classifications. Clustering takes data items which may display a large number of attributes or characteristics (usually expressed as database records) and divides them into a smaller number of categories, grouping individual items while at the same time dividing up the whole data set. Data mining tools do this automatically by looking for 'best-fit' simplifications of the data set. For example, a clustering analysis might reveal a statistically significant correlation between location and buying habits in a customer database. In this case, two or more attributes or characteristics of the data appear to be connected (correlated) in such a way that the chances of the connection being random are below an agreed threshold (statistically significant). The correlated attributes

in database records can be combined, so that other tools can work more efficiently on a simplified data set.

- **Association Rules**

Association rules are used to analyze transactions which can be broken down into distinct components. The model for this approach is market basket analysis, derived from the realization within retailing that buying patterns could be detected in the contents of individual shopping baskets. The ability to break a single transaction down into component parts and determine associations between them is a powerful tool for any industry involved in targeted marketing. Associations are typically expressed as confidence-ratings. For example, 50% of transactions in which beer was purchased also included snack foods. In practice, confidence thresholds are set to eliminate spurious trends, but it generally requires human experience to distinguish genuine trends from temporary phenomena (Herman, 1997).

- **Sequence-based analysis**

Sequence-based analysis can be viewed as a form of market basket analysis in which an attribute can be used to generate a time series. For example, loyalty card account numbers or frequent flyer memberships can be used to track customer activity across time. In these cases, one may be interested in analyzing the components of transactions to determine temporal sequence and frequency (how often do consumers buy snack foods and do they buy them before or after they buy beer). This sort of analysis should allow one to determine "precursor" purchases and to forecast certain behavior or activities. It is particularly useful for the early detection of anomalous situations (Herman, 1997).

V. Big Data Analysis using Data Mining Techniques

Some of the following different analysis is available and analysis tools that make data analysis incredibly easy: -

I. Basic Frequency Analysis

Frequency analysis gives you an "Overall" insight into the responses for your survey. This is the **First** level for your analysis. This gives you an "overall" impression of what your respondents are thinking. In the subsequent sections we'll do some data-mining and in-depth analysis.

II. Cross Tabulation Analysis (Banner Tables)

Cross Tabulation Analysis (Crosstabs) give you more insights into your data. It basically involves the interaction between **two questions** and a distribution of how users responded to both of them -- taken together.

III. Grouping and Segmentation Analysis

Grouping analysis is probably the most interactive tools to "delve" into your data. You can create custom groups or filters. After creating the "Filters" or "Groups" you can click on all the Data Analysis links and view the frequency data filtered by the groups that you just created!

IV. T.U.R.F Analysis - Total Unduplicated Reach and Frequency Analysis

It was originally devised for analysis of media campaigns, and has been expanded to apply to product, line and distribution analysis. With Question any MULTIPLE CHOICE (Multiple Answer) question can be analyzed using TURF. The TURF Simulator can give you One-Click access to optimal configurations for maximizing reach. Reach or Coverage is defined as the proportion of the audience (target group) that chooses a particular option

V. Trend Analysis - Analyzing aggregate response data over time

The past, we can see trajectories into the future - both catastrophic and creative projections." John Ralston Saul The Trend Analysis module allows you to plot aggregated response data over time. This is especially valuable, if you are conducting a long running survey and would like to measure differences in perception and responses over time.

The following data points can be measured (Y-Axis)

1. Mean and Mean Percentile
2. Standard Deviation and Variance

The "Time Factor" (X-Axis) can have the following granularity

1. Daily
2. Weekly
3. Monthly
4. Quarterly (Jan-Mar, Apr-Jun, Jul-Sept, Oct-Dec)
5. Yearly

Trend Analysis can be extremely valuable as an early warning indicator of potential problems and issues with product line and service level changes that impact customers. If you see a dip in the "mean" for a Continuous Variable satisfaction question after a particular

"marketing event" you can immediately start investigating the dip and explore causes of the decrease in satisfaction levels. It can also be used to gauge response rates over time.

Trend Analysis can only be performed on "Quantitative" question types like Multiple Choice, Rank order and Constant Sum. Questions that have textual input (Qualitative) cannot be used for trend analysis.

VI Conclusion

In this study has been analysis of big data from retail industry. The retail entities primarily selling services rather than products are in retailing of services. Data mining tools do this automatically by looking for 'best-fit' simplifications of the data set. Some of the techniques are used in retail industry for Rule Induction and Statistical Analysis, Decision Trees and Neural Nets, Fuzzy Logic and Genetic Algorithms, Clustering and Correlation, Association Rules, Sequence-based analysis. Most of the Big Data analysis for Frequency analysis, Cross Tabulation Analysis, Grouping analysis, Total Unduplicated Reach and Frequency Analysis and Trend Analysis

VII. Reference

1. Mitra, S., Pal, S., and Mitra, P. (2002). Data mining in soft computing framework: a survey. IEEE Transactions on Neural Networks, 13(1):3– 14.

2. Berson, A., Smith, S., and Thearling, K. (1999). Building Data Mining Applications for CRM. McGraw-Hill, New York.

3. Ahmed, S. R. (2004). Applications of data mining in retail business. In In- formation Technology: Coding and Computing, International conference on, volume 2, page 455, Los Alamitos, CA, USA. IEEE Computer Society.

4. Shaw, M. (2001). Knowledge management and data mining for marketing. Decision Support Systems, 31(1):127–137.

5. Giraud-Carrier, C. and Povel, O. (2003). Characterising data mining software. Intell. Data Anal., 7(3):181192.

6. Abdullah Al-Mudimigh, Farrukh Saleem, Zahid Ullah(2009) Department of Information System: Efficient implementation of data mining: improve customer's behavior, 2009 IEEE, pp.7-10.

7. Kamble, A. "Data mining and knowledge discovery - an emerging technology" Electronics Information & Planning, Jul-Aug 1999, p.477-479.

8. Herman, G. "Data Warehousing: Transforming customer information into business intelligence" An Financial Times Management Report published and distributed by FT Retail & Consumer Publishing, 1997, p.21-137

9. Holland, J. "Hidden Order" New York: Addison Wesley, 1995.