APPLYING DATA MINING TO IDENTIFY THE FACTORS INFLUENCING THE RETAIL STORE SHOPPERS

DR. RUCHI MITTAL[1]

ABSTRACT

The Indian retail sector is undergoing a phenomenal change today. It is the right time for the Indian retailers to scale up and invest in technology. Retailers like US' Wal-Mart and Germany's Metro have already started experimenting with the latest RFID technology while most Indian retailers are yet to completely adopt bar coding completely but most retailers do not have IT systems in place. This study seeks to bring in a more scientific and empirical approach to retail marketing by applying the techniques of Data Mining to customer data with the objective of understanding customers better.

[1] Associate Professor, MCA, Maharaja Agrasen Institute of Management and Technology, Jagadhri, Haryana, India

INTRODUCTION

Given a greater access to data and analytical tools, information technology has enabled the retailers to answer increasingly complex questions pertaining to retail strategy and operations—for themselves as well as their customers, suppliers, and other partners. Suppliers, partners, and upstream and downstream intermediaries are plugged in through real-time interfaces that rely on EDI and the open standards of the Internet. The tools also supplement data capture and storage technologies by converting information into intelligence. A simple purchase at any retail store can enable the store to gather a vast amount of information about its customers and products. The use of systems to organize, retrieve, search and manage that data is termed as database management which is a major Retail Information System (RIS) tool. A firm may compile and store data on customer attributes and purchase behavior, compute sales figures by vendor, and store records by product category. Each of these would represent a separate database (Berman and Evans 2004). Two important approaches viz. Data Warehousing and Data Mining put database management into effective practice. Data Warehousing is the coordinated and periodic copying of data from various sources, both inside and outside the enterprise, into an environment ready for analytical and informational processing (Inmon 2002). Data Mining is a technique used to identify patterns in data, typically patterns that the analyst is unaware of prior to searching through the data. One application of Data Mining is Micromarketing, whereby the retailer uses differentiated marketing and develops focused strategy mixes for specific customer segments (Berman and Evans 2004).

RETAILING IN INDIA

The Indian retail industry has over 12 million outlets, which is the largest no. of retail outlets in the world. It contributes over 10% to the GDP of the country and is estimated to provide employment to over 18 million people, around 8% of the country's employment, being the largest employment provider after agriculture. Of the 12 million retail outlets present in the country, nearly 5 million sell food and related products. Even with this large number of outlets, organized retail accounts for only 4 % of the total market, opening huge growth potential in this segment.

According to the FICCI 2007 Retail Report, the share of the organized retail sector is likely to increase from the current 4% to over 20% by 2010, as the overall retail sector grows from $328 billion to $430 billion. The organized retail is likely to grow at a CAGR of 50% and set to be worth $90 billion by 2010. Driven by the changing lifestyles, strong income growth and favorable demographic patterns, Indian retail is expanding at rapid pace. Mall space, from a meager one million sq. ft. in 2002, is expected to touch 60 million sq. ft. by end-2008, according to Jones Lang LaSalle's third annual Retailer Sentiment Survey-Asia.

According to FICCI over $30 billion of investment is likely to be made in the next five to seven years, 92% of which is expected to be in urban areas. The investment estimates are as under:-

Hypermarket	38%
Supermarket	21%
Specialty store	22%
Warehouse	16 %
Department Store	2%

Bigger cities are likely to corner 62% of the total investment expected to be made in urban retail. To sustain this growth rate, the report suggests that India needs to generate at least 110 million sq. ft. of additional retail space a year for several years. The whole concept of shopping has altered in terms of format and consumer buying behavior, ushering in a revolution in shopping in India. Modern retail has entered India as seen in sprawling shopping centers, multistory malls and huge complexes that offer shopping, entertainment and food all under one roof. The Indian population is witnessing a significant change in its demographics. A large young working population with median age of 24 years, nuclear families in urban areas, along with increasing working-women population and emerging opportunities in the services sector are going to be the key growth drivers of the organized retail sector in India.

LITERATURE REVIEW

In earlier nineties data mining through its predictive modeling techniques brought revolutionary growth in the banking sector. This section throws light on the work that has already been done towards retail sector. According to **Melab (2001),** Organizations are facing a data explosion. The organizations are drowning in data but starving for knowledge. A solution is to exploit the emerging data mining technology that is a rich, promising and young field with a wide range of applications like Insurance, Banking, Retailing and medical industries. According to **Levinson (2002),** Retailers collect vast amount of data but do not have means to apply it effectively. The technology available today can offer retailers better results for their planning and buying. According to **Collier (1998),** during the 1990s many companies reach the conclusion that their data is a valuable asset and the companies such as Wal-Mart recognized the benefit of applying Data Mining to these rich stores of historical data. They further note that the data mining industry has, and will, provide real advantages to those who employ it. According to **Lee (2009),** perception, experience and quality have positive and significant influence on consumers' purchase intention. Their study investigated four determinant factors such as trust, security, satisfaction and perception that influenced the customers during online purchasing.

In a study of 301 grocery shoppers, who had shopped at 10 different super markets, **Min (2006)** identified 13 store attributes relevant to super market service quality using decision tree due to its visual appeal, simplicity, and efficiency. These attributes in terms of their ranks (starting with 1) were found as quality of products, cleanliness, competitive price, variety, fast check out, proximity to residence, length of store operating hours, quality of prior services, good price labeling, employee courtesy, ease of payment, special products and word of mouth reputation.

OBJECTIVE OF THE STUDY

The objective of this study is

To identify customer perceptions of grocery store image.

Grocery Store attributes are a mix of functional and psychological attributes of a retail outlet. Functional attributes include merchandise selection, price ranges, credit policies, tore layout and other factors that can be measures to some degree and used to compare one outlet objectively with its competitors. Psychological attributes include objective considerations as a sense of belonging, a feeling of warmth, or friendliness, or a feeling of excitement **(Lindquist 1974-75)**.

RESEARCH METHODOLOGY

For the purpose of this research the statistical data-mining technique used is Factor Analysis which is primarily used for data reduction and summarization. There may be a large number of variables, most of which are correlated and must be reduced to a manageable level. Relationships among sets of many interrelated variables are examined and represented in terms of a few underlying factors. For example, store image may be measured by asking respondents to evaluate stores on a series of items. These item evaluations may then be analyzed to determine the factors underlying store image. **(Dunteman 1989)**. Mathematically, factor analysis is somewhat similar to multiple regression analysis, in that each variable is expressed as a linear combination of underlying factors. The amount of variance a variable shares with all other variables included in the analysis is referred to as *communality*. The co variation among the variables is described in terms of a small number of common factors plus a unique factor for each variable. These factors are not overtly observed. If the variables are standardized, the model may be represented as:

$$Xi = AilFl + Ai2F2 + Ai3F3 + \ldots + AimFm + VPi$$

Where

Xi = *ith* standardized variable

Aij = standard multiple regression coefficient of variable *i* on common factor j

F = common factor

Vi = standardized regression coefficient of variable *i* on unique factor *i*

Vi = the unique factor for variable *i*

m = number of common factors

The unique factors are uncorrelated with each other and with the common factors **(Dillon and Goldstein 1984)**. The common factors themselves can be expressed as linear combinations of the observed variables

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \cdots + W_{ik}X_k$$

where *Fi* = estimate of *ith* factor *Wi* = weight or factor score coefficient *k* = number of variables It is possible to select weights or factor score coefficients so that the first factor explains the largest portion of the total variance. Then a second set of weights can be selected, so that the second factor accounts for most of the residual variance, subject to being uncorrelated with the first factor. This same principle could be applied to selecting additional weights for the additional factors. Thus, the factors can be estimated so that their factor scores, unlike the values of the original variables, are not correlated. Furthermore, the first factor accounts for the highest variance in the data, the second factor the second highest, and so on.

SAMPLING DESIGN

The researcher used the mall intercept convenience sampling technique. The researcher distributed 500 questionnaires, out of which 300 questionnaires were usable leading to a success rate of 60%. The respondents were intercepted outside grocery outlets, supermarkets and other stores dealing in food & grocery goods. The respondents were adult shoppers (more than 20 years of age) and were administered the questionnaires over a period of 3 months.

INSTRUMENT DEVELOPMENT

A total of 23 store attributes have been included in the questionnaire. These attributes have been taken from the "CONSUMER RETAIL STORE IMAGE SCALE" developed by Dickson & Albaum (1997). This scale called CIRS (Consumer Image of Retails Stores) encompasses attitudes towards retail prices, products, store layout and facilities, service and personnel, promotion, and "others" (Dickson & Albaum 1997). The original CIRS had a test-retest reliability of 0.91. According to Hair et al (1998) reliability is used to determine the degree of consistency of a scale. Cronbach's alpha is the most widely used measure of reliability. It measures the consistency of an entire scale. Generally, .70 is an acceptable lower limit, with **.60** being acceptable for exploratory research.

Responses to the 23 store attributes comprising this scale were measured on 7-point Likert-type scales. The 23 store attributes included in the study are: Store advertisements, quality of merchandise, Store Layout, Fresh products, National Brands availability, Quality of Store Brands (PLs-Private Labels), Low Prices, Promotional Schemes, Location of store, Number of salesmen, Services provided, Sales return policy, Variety of merchandise, Spacious store, Clean store, Fast check-out, Attractive visual-display, easy to search goods, attractive loyalty-card, ample parking space, acceptance of credit / debit card, one-stop shopping facility and Air Conditioning.

RESULTS & FINDINGS

Previous research suggests that store attributes produce factors. Factor analysis is used to reduce the environmental dimension scales into smaller, more manageable factors. This multivariate technique is also used to identify the underlying patterns or relationships for a large number of variables (Hair et al., 1998). Factor analysis was used to summarize the variables by examining correlations between the variables, and to create an entirely new set of variables to replace original variables. Factors were derived using component or principal components, which summarizes the original information into factors for prediction. Only factors with latent roots or eigen values greater than 1 were included. Factors were rotated using the varimax rotation method. According to Hair et al., factor loadings at □ □ .30 are considered minimal, 0□ .40 more important, □ .50 or greater practically significant. Items with loadings greater than or equal to □ □ .50 were retained. However,those with several high loadings on more than one factor, variables with low loadings, and those that did not load on any factor were evaluated for possible deletion. Exclusion of a variable was dependent upon its overall contribution to the research. In addition to the variable loading, the communality, total amount of variance shared with other variables was evaluated before deleting the variable. Variables with loadings less than or □ □ .50 and variables that did not load with communalities less than .50 were deleted. After the factors were formed, they were named according to those variables with higher factor loadings.

DESCRIPTIVE STATISTICS

	Mean	Std. Deviation	Analysis N
REPUTATION	5.12	1.515	300
QUALITY	5.14	1.505	300
LAYOUT	4.63	1.159	300
FRESH	5.03	1.444	300
BRANDS	4.90	1.349	300
PLs	4.82	1.337	300
LOWPRICES	5.36	1.570	300
SCHEME	5.07	1.625	300

LOCATION	5.05	1.489	300
SALESMAN	3.97	1.103	300
SERVICE	4.90	1.584	300
RETURN	4.95	1.638	300
VARIETY	4.97	1.231	300
SPACIOUS	4.73	1.083	300
CLEAN	4.74	1.170	300
CHECKOUT	4.78	1.144	300
DISPLAY	4.78	1.031	300
EASYSEARCH	4.61	1.240	300
LOYCARD	5.00	1.583	300
PARKING	5.16	1.459	300
DEBITCARD	5.38	1.312	300
OSS	5.17	1.056	300
AC	3.82	1.208	300

ADEQUECY OF DATA FOR FACTOR ANALYSIS

For checking the suitability of data for factor analysis, four recommended techniques have been applied.

(a) Validation through Correlation Coefficient Matrix of Explanatory Variables, (b) Validation through Anti-Image Correlation Matrix,

(c) Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy,

(d) Bartlett's Test of Sphericity.

(a) **Validation through Correlation Coefficient Matrix of Explanatory Variables**: It is a lower triangle matrix showing simple correlations among all possible pairs of variables included in the analysis for the application of factor analysis. It is obligatory that the data matrix should have good correlations of visual inspection reveals no substantial number of correlations greater than 0.30 then factor analysis is probably inappropriate (Heir, 2003). A perusal of the output indicated that there are enough correlations between the variables greater than 0.30 indicating the suitability of data for application of factor analysis.

(b**) Validation through Anti-Image Correlation Matrix**: The Anti-image matrices contain the negative partial covariances and correlations. They can give an indication of correlations which aren't due to the common factors. Small values in the table indicate that the variables are relatively free of unexplained correlations. Most or all values off the diagonal should be small (close to zero). Each value on the diagonal of the anti-image correlation matrix shows the Measure of Sampling Adequacy (MSA) for the respective item. Values less than .5 may indicate variables that do not seem to fit with the structure of the other variables. The values in the diagonal are all more than 0.5 thus indicating sampling adequacy and suitability of factor analysis.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin **(KMO)** Measure of Sampling Adequacy		.895
Bartlett's Test of Sphericity	Approx. Chi-Square	6132.839
	Df	253
	Sig	.000

c) **Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy**: It is an index used to examine the appropriateness of factor analysis. High value (between 0.5 and 1.0) indicate adequacy of data for the use of factor analysis (Malhotra, 2002). Here, the computed value of KMO statistics is .895 indicating the adequacy of data for factor analysis. This index compares the magnitudes of the observed correlation coefficient to the magnitudes of the partial correlation coefficients.

Small values of the KMO statistic indicate that the correlations between pairs of variables can not be explained by other variables and that factor analysis may not be appropriate.

d) **Bartlett's Test of Sphericity**: It can be used to test the null hypothesis that the variables are uncorrelated in the population; in other words, the population correlation matrix is an identity matrix. In an identity matrix, all the diagonal terms are 1, and all off-diagonal are 0. The test statistics for sphericity is based on a chi-square transformation of the determinant of the correlation matrix. A large value of the test statistic will favor the rejection of the null hypothesis that the variables are uncorrelated in the population. The test statistics computed is 6132.839 ($\chi2$) which is very high leading to the rejection of the hypothesis as mentioned earlier. Thus according to the Bartlett's test of Sphericity, factor analysis is appropriate for this data. **(Malhotra 2004; Hair et al 1998).**

The principal component analysis was used for extraction of factors through Varimax Rotation.

The number of factors to be retained was based on

(a) the Latent Route Criterion (Eigen Value Criterion),

(b) Determination Based on Percentage of Variance.

COMMUNALITIES

	Initial	Extraction
REPUTATION	1.000	.855
QUALITY	1.000	.889
LAYOUT	1.000	.796
FRESH	1.000	.606
BRANDS	1.000	.812
PLs	1.000	.879
LOWPRICES	1.000	.852
SCHEME	1.000	.732

LOCATION	1.000	.915
SALESMAN	1.000	.768
SERVICE	1.000	.886
RETURN	1.000	.921
VARIETY	1.000	.745
SPACIOUS	1.000	.635
CLEAN	1.000	.889
CHECKOUT	1.000	.608
DISPLAY	1.000	.649
EASYSEARCH	1.000	.819
LOYCARD	1.000	.954
PARKING	1.000	.913
DEBITCARD	1.000	.587
OSS	1.000	.770
AC	1.000	.457

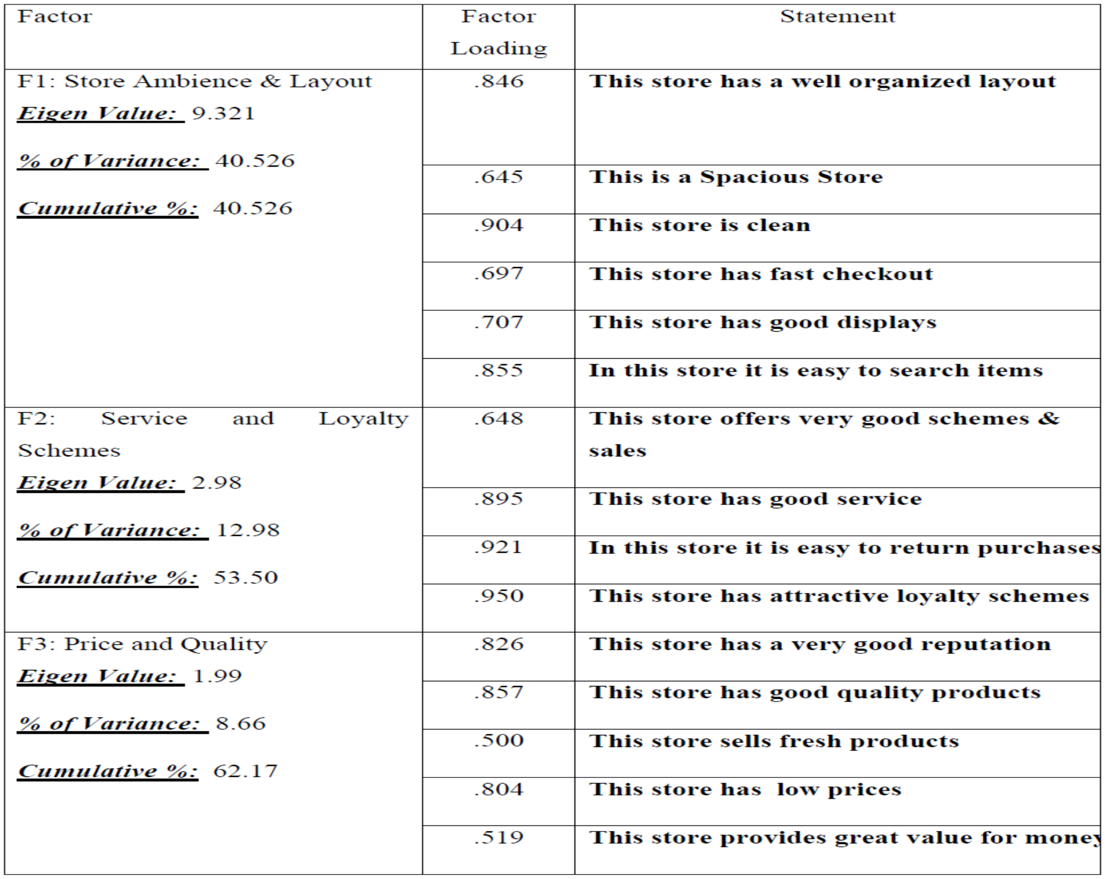
Extraction Method: Principal Component Analysis

Communalities are considered high if they are all 0.8 or greater but this is unlikely to occur in real data. Generally accepted communalities are between 0.40 to 0.70 (Costello and Osborne 2005). In this study all communalities are above 0.457 so they will all be retained for further analysis.

Varimax Rotated Component / Factor Loading Matrix (a)

	Component					
	1	2	3	4	5	6
REPUTATION	.147	.256	**.826**	.193	.211	.059

QUALITY	.154	.317	**.817**	.244	.182	.062
LAYOUT	**.846**	.042	.243	.062	-.014	-.122
FRESH	.052	.183	**.500**	.371	**.413**	-.105
BRANDS	.270	.209	.261	**.762**	.203	.068
PLs	.274	.203	.236	**.827**	.135	.067
LOWPRICES	.188	.234	**.804**	.218	.258	.047
SCHEME	.146	**.648**	**.502**	.165	.003	.106
LOCATION	.130	.034	.195	.114	**.907**	.153
SALESMAN	.010	.051	.127	-.019	.152	**.852**
SERVICE	.097	**.895**	.243	.106	.062	.035
RETURN	.071	**.921**	.173	.175	.078	.027
VARIETY	.260	.114	.274	**.752**	.112	-.100
SPACIOUS	**.645**	.138	.030	.310	.207	.245
CLEAN	**.905**	.041	.162	.157	.064	-.119
CHECKOUT	**.697**	.097	.079	.181	.140	.233
DISPLAY	**.707**	.178	-.038	.306	.086	.126
EASYSEARCH	**.855**	-.012	.193	.183	.066	-.113
LOYCARD	.060	**.950**	.158	.144	.047	.012
PARKING	.134	.042	.197	.112	**.911**	.105
DEBITCARD	.183	.346	**.414**	.184	**.419**	-.231
OSS	.234	.106	.143	**.827**	.012	.026
AC	.142	.091	**.519**	.182	-.005	.354



Factor	Factor Loading	Statement
F1: Store Ambience & Layout ***Eigen Value:*** 9.321 ***% of Variance:*** 40.526 ***Cumulative %:*** 40.526	.846	**This store has a well organized layout**
	.645	**This is a Spacious Store**
	.904	**This store is clean**
	.697	**This store has fast checkout**
	.707	**This store has good displays**
	.855	**In this store it is easy to search items**
F2: Service and Loyalty Schemes ***Eigen Value:*** 2.98 ***% of Variance:*** 12.98 ***Cumulative %:*** 53.50	.648	**This store offers very good schemes & sales**
	.895	**This store has good service**
	.921	**In this store it is easy to return purchases**
	.950	**This store has attractive loyalty schemes**
F3: Price and Quality ***Eigen Value:*** 1.99 ***% of Variance:*** 8.66 ***Cumulative %:*** 62.17	.826	**This store has a very good reputation**
	.857	**This store has good quality products**
	.500	**This store sells fresh products**
	.804	**This store has low prices**
	.519	**This store provides great value for money**

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization (a: Rotation converged in 7 iterations.).

FACTOR LOADINGS

F4: One Stop Shopping *Eigen Value:* 1.45 *% of Variance:* 6.31 *Cumulative %:* 68.48	.762	**This store stocks well known brands**
	.827	**This store's own products are of good quality**
	.752	**This store has a vast variety of products**
	.827	**This store offers everything under one roof**
F5: Convenience *Eigen Value:* 1.14 *% of Variance:* 4.98 *Cumulative %:* 73.46	.907	**This store has a very convenient location**
	.911	**This store has good parking facilities**
	.419	**This store offers option to pay by credit/debit card**
F6: Salesman *Eigen Value:* 1.04 *% of Variance:* 4.526 *Cumulative %:* 77.986	.852	**This store has helpful salesmen**

INTERPRETATION OF FACTORS

The factors or dimensions that have emerged have been named based on the types of attributes that have converged to form a particular factor / dimension. The following six factors are described as follows:

Factor 1- Store Ambience and Layout: This dimension is catching up especially in metros where consumers are increasingly seen thronging modern retail stores and malls. Modern retailers such as Reliance Fresh, More, Easy Day, Big Bazaar etc. seem to invest a good deal in Store ambience.

Factor 2- Service and Loyalty Schemes: Prompt service, problem solving, return of goods and loyalty schemes have always been important to consumers. Offering loyalty schemes enables retailers to pre-empt attractions from competitors. Services combined with attractive offers can ensure long term store loyalty.

Factor 3- Price and Quality: This dimension includes variables such as store reputation, quality products, fresh products, low prices and value for money. Pricing and quality related store attributes ensure long-term sustainable customer loyalty. It is also very interesting and logical that consumers have chosen to consider all of these attributes and a single dimension.

Factor 4- One Stop Shopping: Availability of different popular brands, good quality store rands, and availability of several products under one roof ensures that the modern consumer can shop at the same store and not move around several stores to buy merchandise. This is the modern consumer who likes visiting large super markets, hypermarkets and shopping malls. This consumer seems hard pressed for time.

Factor 5- Convenience: For today's grocery shopper convenience includes easy location, easy payment options and comfortable parking. These attributes converging into one single dimension prove that convenience is a very important dimension and grocery retailers must incorporate this in their overall retail strategy.

Factor 6- Salesmen: This is an important attribute and salesmen play a pivotal role in makingthe shopping environment friendly and personalized. Even in modern stores, where self-service is the norm, salesmen can be helpful without being obtrusive and at the same time can ensure enduring relationships between customers and the store.

CONCLUSION AND FUTURE RESEARCH

This study on store attributes takes a very customer-centric approach in understanding have consumer perceive the store image, which is represented by a combination of the store attribute dimensions. In this study consumers perceive that all store attributes can be summarized into 6 major dimensions: Store Ambience and Layout; Service and Loyalty Schemes; Price and Quality; One Stop Shopping; Convenience and Salesmen. The data mining approach used here is the Principal Component Analysis (PCA). This is a factor analysis approach where varimax rotation was applied. The PCA approach allows us to calculate exact factor scores, the benefit of which is that we can use these factor scores to develop a predictive model based on the multiple regression method. This will be the future research to be conducted based on the findings of this paper.

REFERENCES

1. Melab N.(2001), Data Mining: A key Contribution to E-business. Information and Communication Law. Volume 10. Issue 3.
2. Collier K. et. al (1998), "A Perspective on Data Mining," July, The Centre for Data Insight, t Northern Arizona State University, pp 1-38.
3. Min, H. (2006), "Developing the Profiles of Supermarket Customers Through Data Mining", The Service Industry Journal, 26(7), October, 2006, pp 747-763.
4. Lindquist, Jay D. (1974-75), "Meaning of Image" Journal of Retailing, Vol. 50, p 31.
5. Berman, B. and Evans, Joel R. (2004), Retail Management: A Strategic Approach (9/E), Pearson Education, New Delhi.
6. Inmon, William H. (2002), Building The Data Warehouse, John Wiley and Sons, New York
7. Malhotra, N. (2004), Marketing Research, Pearson Education, New Delhi.
8. Hair, J.F. et. al. (1998), Multivariate Data Analysis, (5th Edition). Upper Saddle River, NJ: Prentice Hall
9. Dillon W.R., Goldstein M. (1984), Multivariate Analysis, Methods and Applications. New York, Singapore: Willey. ISBN 0471 08317 8.
10. Levinson, M. (2003), The RFID Imperative. CIO Magazine. Retrieved from http://www.cio.com/archive/120103/retail.html
11. Lee Y.J. (2009) Exploring the Influence of Online Customers' Perception on Purchase Intention as Exemplified with an Online Bookstore, Vol. 5 Issue 2.
12. Dunteman, G., 1989. Principal Component Analysis. Sage University Paper 07-69, Thousand Oaks, CA.
13. Costello A.B., Osborne J.W. (2005), "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the most from Your Analysis", Practical Assessment, Research and Evaluation, Vol. 10, No 7, ISSN: 1531-7714.