A Novel Approach for Classification of E-Commerce Data

Chaitanya K
Infosys Limited
Hyderabad
India
Chaitanya_K02@infosys.com

D. V. L. N. Somayajulu
Dept. of Computer Science & Engineering
National Institute of Technology, Warangal
India
soma@nitw.ac.in

P. Radha Krishna
Infosys Limited
Hyderabad
India
radhakrishna_p@infosys.com

ABSTRACT
Classification is one of the most interesting problems in the fast evolving fields such as e-commerce and web-based businesses where the data is growing exponentially. Existing classification techniques over e-commerce data are mainly based on the users' purchasing patterns. However, gender preferences significantly improve in recommending various products, targeting customers for branding products, providing customized suggestions to the users etc. In this paper, we propose a two-phase approach for gender based classification to classify e-commerce data by exploiting hierarchical relationships among products. The first phase reduces the dimensionality of the data by identifying the features that well describes the browsing pattern of the users. The second phase classifies the data based on these features. Experiments are carried out on clickstream data (provided by FPT group) consisting of browsing logs (with list of products formed as a hierarchy), session start time and session end time. We compared our results with standard Bayesian classification model, which shows the applicability of our classification approach for e-commerce data.

Categories and Subject Descriptors
H.2.8. [**Information Systems**]: Database Applications – *Data Mining.*

General Terms
Algorithms, Performance, Experimentation.

Keywords
Gender classification, e-commerce, feature extraction

1. INTRODUCTION
The use of internet is increasing day-by-day by the users across the globe. In the recent times, Internet is further powered by elements such as social media and e-commerce [3]. The initial surfers of the internet were more likely to be young, male, better educated, more affluent and urban [8]. With the capability of Internet in providing information, Internet has become a new flexible channel for the retailers to promote their goods. This also establishes a two way communication channel between the customer and the retailer [1]. Internet is also a perfect place for e-commerce business because the information on the product is immediately available. Buyers can also compare parameters of a product such as its price and reviews given by the users [9]. E-commerce purchases have also become a routine to many people due to the accessibility of the websites from multiple devices such as tablets and smart phones.

Analyzing and extracting the information from the customer data is an important task that can assist the process of improvement of the company's sales through e-commerce [7]. From the users browsing data/logs of an ecommerce website within a given period, it is possible to extract important rules which are useful for the decision-making. One commonly distinguished measure in Internet access and usage is gender of the user who is browsing [8].

Classification is one of the most interesting problems in the fast evolving fields such as e-commerce and web-based businesses where the data is growing exponentially. In general, e-commerce data is classified into *usage data*, *content data* and *user data*. Usage data consists of user sessions. The content data in a site are collection of objects and the relationships that are conveyed to the users. User data may include information of registered users and reviews given the user [2]. The dataset used in this work is the combination of usage data and content data. In this paper, we present a classification approach based on purchasing patterns made by the customers. This model identifies a set of features which are useful for the classification of gender from the browsing data.

The rest of the paper is organized as follows. In section 2, we discuss related work, and in section 3, we describe our approach and provide implementation details. In section 4, we discuss the results of the experiment and demonstrate our approach on an independent dataset. We discuss our findings and conclude the paper in section 5.

2. RELATED WORK
There have been several attempts made by the researchers in the past to find the gender of the Internet users. The use of Internet has started increasing in the last decade of the 20 century [8]. Studies show that during that period women used the internet less than man but this difference was disappeared by early years of 21 century. This study also indicated that women remained less frequent users of the Internet. A similar study was also conducted by Zhang et al [14] to find the role of gender in bloggers' switching behavior. They concluded that bloggers' intention to switch their blog services were related with three factors: satisfaction, sunk costs and attractive alternatives. They also concluded that female had more sensitivity to satisfaction and fewer tendencies to attractive alternatives than male. On the other hand, some studies have been carried out to find the relationship between gender and web design perspective. For instance, Tuch et

Compute'15, Oct 29--31, 2015, Ghaziabad, India.

DOI: http://dx.doi.org/10.1145/2835043.2835049

al [11] showed that only men are sensitive to symmetry of a website into consideration and gave good results when compared to women. Gender difference is the most interesting aspect for the advertising and marketing companies since decades. Hoffman et al [5] showed that men were more likely to browse and purchase a product than women. According to Tracy et al [10], male users have been shown to be 2.4 times more likely than women to shop online

Several machine learning techniques such as Bayesian Logistic Regression, Support Vector Machine, and AdaBoost decision tree can be used to find the gender of the users based on the features identified [3]. According to Cheng et al [3], identifying correct set of features for the gender identification is still an open problem. Attempts were made by Bhaskar et al [2] on the weblog files to examine the purchasing patterns of the users to find whether customers are purposefully shopping or buying products that they are familiar about. Ketul et al [6] discussed about web mining in ecommerce and pattern discovery techniques to find out interesting patterns from the web data. They also discussed on the pattern discovery techniques such as association rules, clustering, sequential patterns and classification to improve the success of ecommerce sites and improve service to the customers. Zhang et al [13] studied the behavior patterns of visitors using web mining by taking gender as a variable for customer clustering. They extracted knowledge from the gender of the classified data. They also performed association rule mining on the data pertaining to web pages, the buying patterns of customers, and predicting model generation for the potential customers.

3. PROPOSED APPROACH

We present a two-phase approach for the classification. In the first phase, we read the training data sequentially, preprocess the data to remove the unnecessary data followed by the identification of features namely *product category* feature, *category* feature, *product prefix* feature. The second phase classifies the data based on the features identified in the first phase. In this process, *Gender Values* for both male and female users are calculated for an element to determine the gender. In this work, we considered log data related to purchasing patterns of users per session. Each record X in the training data is a 5-tupled as shown below:

$$X = \{ S_{id}, G_{M|F}, S_{Start}, S_{End}, D \} \rightarrow (1)$$

where S_{id}: Session ID,
$G_{M|F}$: Gender of the person which can be either 'Male' or 'Female',
S_{Start}: Session Start time,
S_{End}: Session End time,
D: Detail of the products.

Details are formed as a hierarchy starting from category at the highest level followed by subcategory (which may by up to *m* levels) and finally the Product ID. D for a session may have data related one or more products and it can be represented as

$$\{C_1 /SC_1 /SC_2/\ldots.SC_m / Pid_1; C_2 /SC_1 /SC_2/\ldots.SC_m / Pid_2; C_n /SC_1 /SC_2/\ldots.SC_m / Pid_n \}$$

where Cn is the category of $\mathbf{n}^{th}$ product in a session,
$\mathbf{SC_1}$ is the subcategory of level 1,
$\mathbf{SC_2}$ is the subcategory of level 2, ……
$\mathbf{SC_n}$ is the subcategory of level *m*,
$\mathbf{Pid_n}$ is the Product ID of n^{th} product in a session.

Using the above convention, the male data can be represented as

$$X_{Male} = \{ S_{id}, G_{Male}, S_{Start}, S_{End}, D \} \rightarrow (2)$$

where the value of G_{Male} is the string "Male"; and the female data can be represented as,

$$X_{Female} = \{S_{id}, G_{Female}, S_{Start}, S_{End}, D\} \rightarrow (3)$$

where the value of G_{Female} is the string "Female".

3.1 Preprocessing of Data

Data preprocessing improves quality of the data which in turn increases the accuracy and efficiency. As part of preprocessing step, the session out time for every session should be looked upon. The session whose logged out time is NULL is omitted. In this case, the assumption is that the user has forgotten to logout. Also we observed that the time difference between session starting and session ending is indefinite (for example, a session with time gap more than 10 hours for browsing 2 products). These kinds of sessions are also removed from the data and hence they do not participate in any of the functionalities of our approach.

3.2 Feature Extraction

For classification of e-commerce data, we extracted 3 features namely *product category*, *category* and *product ID*. The data (both test and training) is organized in the form of hierarchy starting from category, subcategory level *1*, subcategory level *2*, and so up to subcategory level *m* and finally product ID. The association between the products and categories can be better depicted in the form of a tree in which category ID is root, sub category level 1 is the first level node, sub category level 2 is the second level node and the product ID is at leaf node. This arrangement is analogous to the concept of lattice (like concept hierarchy [4]) where each product can have more than one sub category as a parent. Figure 1 shows an example tree for one session data.

3.2.1 Feature based on product category

The input data of a session (which is formed as a hierarchy) can be portioned into separated sets, which serves as a feature, as shown below:

$$\{ \{ Ci \}, \{ Ci, S1\}, \{ Ci, S1, S2\}, \{ Ci, S1, S2, S3\},\ldots\ldots \{Ci, S1, S2, S3 \ldots.Sn, PID\} \} ;$$
$$\{ \{ Cj \}, \{ Cj, S1\}, \{Cj, S1, S2\}, \{ Cj, S1, S2, S3\},\ldots\ldots \{ Cj, S1, S2,S3,\ldots\ldots.Sn, PID\} \} \rightarrow (4)$$

where *Ci*, *Cj* are the Categories, S1 is level *1* subcategory, S2 is level *2* subcategory and Sn is level *n* subcategory and PID is the product ID related to the product. By following the similar strategy, the product data for all the sessions are partitioned and formed as a larger set. *Time stamps* are also considered, which will help in classifying the records that show ambiguity while classifying records.

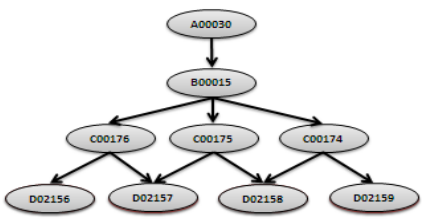

Figure 1. Concept hierarchy of a product

3.2.2 *Features based on Category level*

It is quite common in e-commerce data that products related to male are contained in male categories and products related to females are contained in female categories. Hence, features can be extracted from the training data by taking the session detail information **D** and grouping the data that are same across the products. Consider category level features related to **two products.** The generic form where the category and subcategories for the two products are same except the product ID can be shown as:

$$\{ \{ S_{Start}, S_{End}, \; Ci, S1: Sn \}, \{ P1, P2 \} \} \quad \rightarrow (5)$$

where S1:Sn denote subcategories from S1 to Sn. The generic form where only category is same but different sub categories and product ID's:

$$\{ \{ S_{Start}, S_{End}, Ci \}, \{ S1:Sn \}, \{ P1, P2 \} \} \quad \rightarrow (6)$$

3.2.3 *Feature based on product ID*

It is logical that data in e-commerce websites are arranged in sequential fashion categorized by gender. If Pi is the product related to male, then it is reasonable that product Pi+1 is also a product related to male. Hence, all products viewed by males may have sequential product IDs. Similarly, all products viewed by females have their product IDs in sequence. Hence, **common prefix** feature is found by scanning all product IDs sequentially across all the sessions and finding its length which can be considered as a feature. That is, the average length till which the product IDs are same in majority of the product IDs is taken as the ***prefix length*** (pl_n)**.**

The whole training data is scanned to identify features which are in the form shown in equations (4), (5) and (6).

3.3 Classification

By conceptualizing the data into a concept hierarchy and using the above features, the classification of data into **male** and **female** is done in the following manner.

3.3.1 *Level 1*

The test data related to each session is read in sequential manner. As a first step, *category* level feature (as shown in equation (5)) is used for the classification system. This feature values are used to identify matched records in both male and female lists. If records are found in any of the list, we consider that gender is identified. The matched records are removed from the test data to reduce the size. This is considered as pruning process in our approach. Next, the features that are identified using equation (6) are used for the gender prediction. We followed the same procedure that is used for the other category level feature (for equation (5)) to identify the gender. There are chances that the product may be present in both male and female training data. For such records, we used the product prefix feature (see section 3.2.3). The prefix of size ***prefix length*** (pl_n) is taken from product IDs of each session and followed the same procedure as implemented for features in equation (5). Once the gender is identified, the product information is removed from test data.

Constantly removing the identified products information has two advantages: (a) reduce the search space and (b) decrease the search time.

3.3.2 *Level 2*

After the above steps, there will be cases where (a) Gender is not identified, and (b) Data is present in both male and female training data. Hence, to further classify data, *product category* feature is used.

Operations using the Product Category feature:

Frequency of each element in the feature set is computed from the training data and tabulated as shown in Table 1.

Table 1. Generic table format for product category feature

Element in feature set	Frequency of Male	Frequency of Female	Total
C_i	X	Y	(X+Y)
$\{C_i, S_1\}$	X1	Y1	(X1+Y1)
$\{C_i, S_1, S_2\}$	X2	Y2	(X2+Y2)
. . $\{C_i, S_1, S_2 \ldots C_i, S_1, S_2, S_3, \ldots\ldots, S_n, PID\}$	. . (X_n)	. . (Y_n)	. . (X_n+Y_n)

Element values for male, $EV_{en}(Male)$, is calculated as given below:

$$EVen(Male) = \frac{Male(en)}{Male(en)+Female(en)}$$

Element value for female, $EV_{en}(Female)$, is calculated as,

$$EVen(Female) = \frac{Female(en)}{Male(en)+Female(en)}$$

Here, for instance, *male(e1) is the frequency of males containing the element e1, and female (e1) is the frequency of females containing the element e1.*

The element values for each element are summed up to get the Gender Value of Male ($GV_{(M)}$) as given below:

$$GV_{(M)} = \frac{\sum_{k=1}^{n} EVek(Male)}{n}$$

Similarly, the Gender Value of Female is,

$$GV_{(F)} = \frac{\sum_{k=1}^{n} EVek(Female)}{n}$$

Here, n is the number of elements that are extracted from a record and EV is the element value of each element.

Classification is done based on the gender values. If the male gender score is greater than female gender score, then the product is classified as male class. The records are classified as female in case the female gender score is greater than male score. In the case of unclassified cases (where the gender score for both male and female are either same or zero), Product ID feature is applied. Further, **Timestamps** are always useful when there is a biased situation. It can be treated as a discriminative element for classification and thus useful to precisely classify the data.

4. EXPERIMENTAL RESULTS

We used the dataset provided by FPT group which was published in the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) competition [15]. The training data comprises 15,000 records which are the logs related to the products. Each row in the log file consists of four columns: session ID, session start time, session end time and list of product IDs (in sequential order of viewing the products in that session).

Since distribution of unique product IDs in the data is very sparse, the IDs contain additional information regarding product category hierarchy. Each product ID can be decomposed into four different IDs. The IDs starting with letter 'A' are the **most general categories,** starting with 'D' correspond to **individual products** and 'B' and 'C' are associated with **subcategories and sub-subcategories**, respectively. By following equations (1), (2) and (3), training data can be represented as,

$$X_{Training} = \{ X_{Male} \cup X_{Female} \}$$

The training data is sequentially and is stored in separate *Lists*. The List data structure serves as a wrapper for an array, providing read/write access to the array and automatically resizing the array as needed. While formulating the approach for gender classification, we assume that (i) every record pertaining to a session in the log file contains all the information as specified in equation (1), (ii) the distribution of labels across the training data is not balanced and (iii) labels for the gender are already provided for training data.

Table 2. Total feature count

Feature	Sub-category	Total Count	Male Features	Female Features
Features based on product and category		22,466	4,559	17,907
Features based on product category	till 2nd subcategory	742	323	419
	till 1st subcategory	163	79	84

The features are extracted for the test data and the Table 2 shows the total count of the features that are identified category wise. We observed that majority of the product IDs are same till 4th position. Hence, prefix length is taken as 4. Our approach is explained on a sample record given below (taken from the test data):

11/14/2014 17:15	11/14/2014 18:09	A00002/B00003/C00046/D01169/; A00002/B00003/C00014/D01478/ A00002/B00003/C00046/D01153/;

We first extract the features for the sample record.

Category level features:

a) { {11/14/2014 17:15, 11/14/2014 8:09} {A00002/B00003/C00046}, {D01169, D01153} }

b) {11/14/2014 17:15, , 11/14/2014 8:09} {A00002/B00003}, {C00046, C00014}, {D01169, D01478, D01153} }

Here, Point *a* is the feature that is extracted till second level subcategory and point *b* is the feature that is extracted till first level subcategory.

Product Category level features:

{ { A00002, A00002/B00003, A00002/B00003/C00046, A00002/B00003/C00046/D01169}, { A00002, A00002/B00003/ A00002/B00003/C00014, A00002/B00003/C00014/D01478 } }

Element weights of the features are shown in table 3.

Product ID feature is taken into consideration for the elements whose element weight is 0. Since the product's prefix length is taken as 4, the first 4 digits of the product ID is taken into consideration for concluding the gender. Referring to our example, since the element value of

A00002/B00003/C00046/D01169,
A00002/B00003/C00046/D01153

is 0; the prefix of length 4 is looked up on and the frequency of such product IDs are identified in the training set. The frequency of D0116 in the male set is 3 and frequency in the female set is 2. The frequency of D0115 in male dataset is 1 and the frequency in female dataset is 11. Element Values of the above features are shown in Table 4.

Table 3. Element weights of the features

Element	Male Count	Female Count	Total
A00002	1707	12360	14067
A00002/B00003	337	2537	2874
A00002/B00003/C00046	32	305	337
A00002/B00003/C00046/D01169	0	0	0
A00002/B00003/C00014	51	594	645
A00002/B00003/C00014/D01478	0	1	1
A00002/B00003/C00046/D01153	0	0	0

Table 4. Element values for the sample record

Element	EV_{Male}	EV_{Female}
A00002	0.12	0.87
A00002/B00003	0.11	0.88
A00002/B00003/C00046	0.09	0.90
A00002/B00003/C00046/D01169	0.6	0.4
A00002/B00003/C00014	0.07	0.92
A00002/B00003/C00014/D01478	0	1
A00002/B00003/C00046/D01153	0.08	0.91

The Gender values of male and female are:

$GV_{(M)} = 1.07/7 = 0.15$ and $GV_{(F)} = 5.88/7 = 0.84$

Since the gender value of female is greater than the gender value of male, it can be considered that the product is viewed by female.

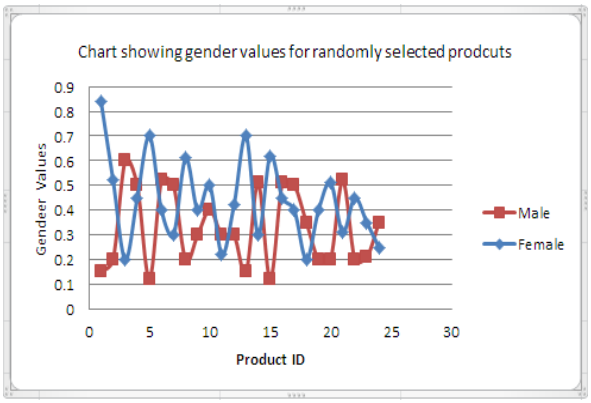

Figure 2. Gender values for randomly selected products

We compared our results with Bayesian approach (presented in [2][12]) and observed that our approach is performed better. Our approach has generated 70% accuracy according to the submission system published in [15], whereas the Bayesian approach ([2] [12]) yielded 30% accuracy which was published at [16]. Figure 2 shows scatter chart for gender values of 25 randomly selected products

5. DISSCUSSION AND CONCLUSION

Recently, the diffusion of the Internet increased as e-commerce has undergone a tremendous change. The ability to perform Internet operations has led to changes in consumers' behaviors. Searching for the products in e-commerce websites is one of the most important and frequent activity these days. Previous studies have shown that both men and women participate equally in Internet activities. When users browse websites, particularly the ecommerce, the data is generated in the form of logs which contains variables such as age, gender and products purchased. Companies are most benefited in finding these variables. The benefits include understanding the customer behavior, improving the customer services and providing the reports to the users which can include the average time spent by user, most frequently purchased products etc.

In this paper, we presented an approach to classify the gender of the user from the log files. We applied a two-phase approach for the classification. In the first phase, we identified several features such as product category feature, category feature and product prefix feature. The second phase classifies the data based on the features identified in the first phase. The data for this experiment is taken from [15]. The training and test datasets contained 15000 records each. We identified 22,466 product and category level features, 905 features of product category (out of which 742 features which are grouped till 2nd product sub category and 163 features which are grouped till first level sub category). Table 5

Table 5. Total female and male counts for various number of products in a session

Number of products	Total Female Count	Total Male Count
> 5	756	191
> 6	508	150
> 7	353	131
> 8	240	107

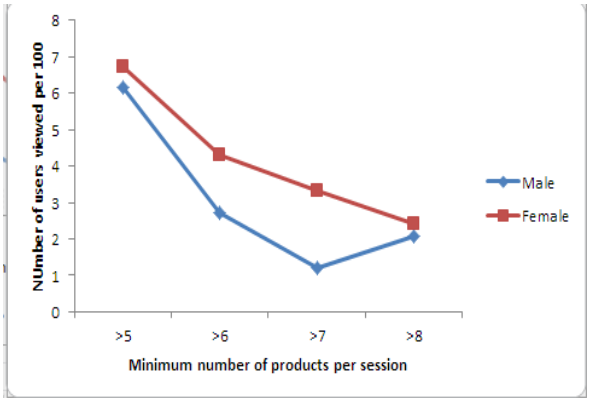

Figure 3. Number of users per 100 who browsed more than 5 products for test data

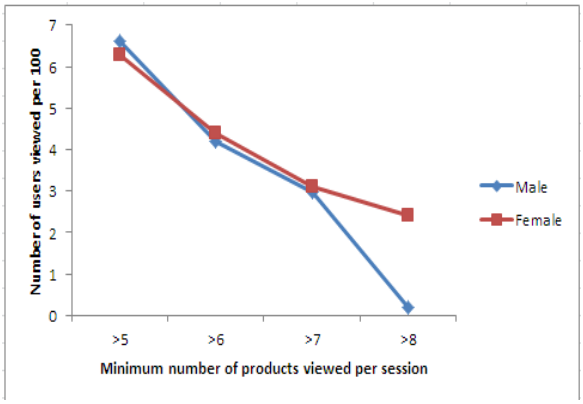

Figure 4. Number of users per 100 who browsed more than 5 products for training data

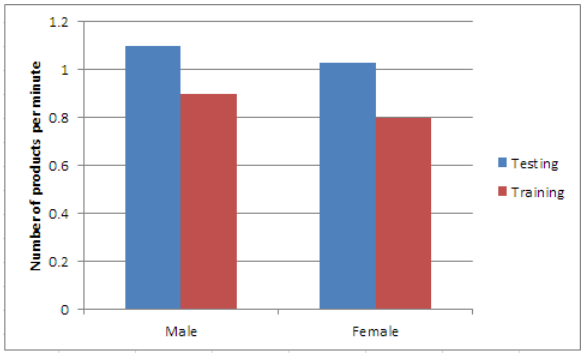

Figure 5. Number of products per minute (for both male and female) for both test data and training data

shows the total number of users (both male and female) and the number of products browsed during a session.

Figures 3 and 4 show the relationship between the product viewing percentage (i.e., scaling to the number of products viewed by 100 users) and the number of products for test data and training data respectively. From figure 5, we can observe that males browse more number of products than females before

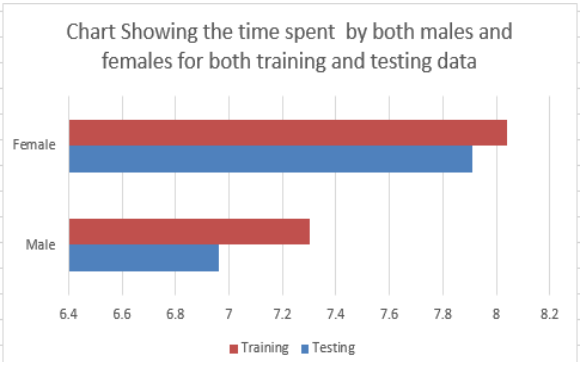

Figure 6. Average time spent by the users

coming to any conclusion. For instance, in the training set, men have browsed as high as 33 products in a single session before coming to a conclusion. We also found from the dataset that (a) male users browse 22% faster than female users, (b) female users browse more number of products and switch between the categories and products when compared to male users, and (c) female spend more time (average 10.5% more) when compared to men on the internet for browsing the data (see figure 6).

6. REFERENCES

[1] Basu, A. and Muylle, S. 2011. Assessing and enhancing e-business processes. Electronic Commerce Research and Applications, Elsevier. 10, 4 (July–August 2011) 437–499.

[2] Bhaskar V., Satyanarayana, .P. and Sreeram, M. 2010. Analyzing target customers behavior by mining ecommerce data, *International journal of Information sciences and computing*, 4, 1, (Jan./Jul.2010) 27-30.

[3] Cheng, N., Chandramouli, R. and Subbalakshmi, K. P. 2011. Author Gender Identification from Text, *Digital Investigation*, 8, 1, 78-88.

[4] De, S. K. and Krishna, P. R. 2002. Mining web data using clustering technique for web personalization, International Journal of Computational Intelligence and Applications (IJCIA), World Scientific, 2, 3, pp. 255-265.

[5] Hoffman, D. L., W. D. Kalsbeek, and T. P.Novak. 1996. Internet and Web Use in the U.S., *Communications of the ACM* 39, 12, 36–46.

[6] Ketul, B. P. Jignesh, A. C. and Jigar, D. P. 2011. Web Mining in E-Commerce: Pattern Discovery, Issues and Applications. *International Journal of P2P Network Trends and Technology,* 1, 3, 40-45.

[7] Mehenni, T. and Moussaoui, A. 2012. Data mining from multiple heterogeneous relational databases using Decision tree classification, *Pattern Recognition Letters*, 33, 13 (Oct. 2012) 1768–1775.

[8] Ono, H. and Zavodny, M. 2003. Gender and the Internet. *Social Science Quarterly*, 84, 1, 111-121.

[9] Srinivasana, S. S., Andersona, R. and Kishore, P. 2002. Customer loyalty in e-commerce: An exploration of its antecedents and consequences. *Journal of Retailing*, Elsevier, 78, 1 (Spring 2002) 41-50.

[10] Tracy, B. "Seasoned Users Lead in E-commerce." *Advertising Age* 69, 26 (1998), 39.

[11] Tuch, A. N., Bargas-Avila, J.A. and Opwis, K. 2010. Symmetry and aesthetics in website design: It s a man's business. *Computer on Human Behaviour*, Elsevier, 26, 6 (August.2010), 1831-1837.

[12] Vitor V. de S. Campos, Carlos E. Bueno, Jacques D. Brancher, Fabio T. Matsunaga. Rafael and R. Negrao. 2015. An Approach Based on Data Mining to Support Management in E-Commerce, In *Proceedings of the Third International Conference on E-Technologies and Business on the Web*, Paris, France 2015.

[13] Zhang, X., Gong, W. and Kawamura, Y. 2004. Customer Behavior Pattern Discovering with Web Mining, In *Proceedings of APWeb 2004*, LNCS 3007, 844–853.

[14] Zhang, K. Z. K., Matthew K.O. Leeb, M. K. O., Christy M.K. Cheungc, C. M. K. and Chend, H.2009. Understanding the role of gender in bloggers' switching behavior. *Decision Support Systems*, Elsevier, 47, 4 (Nov. 2009), 540-546.

[15] https://knowledgepit.fedcsis.org/contest/view.php ?id=107.

[16] https://msdn.microsoft.com/en-us/magazine/jj891056.aspx