PREDICTING GENDER OF ONLINE CUSTOMER USING ARTIFICIAL NEURAL NETWORKS

GOKHANSILAHTAROGLU

Istanbul Medipol University MIS Department
Email: gsilahtaroglu@medipol.edu.tr

Abstract - Customer age and gender are very important parameters for both retailing and marketing. It is well known that they both play very important roles in purchasing habits. In this study, we propose a model to predict the gender of an online customer by analysing his/her mouse movements. To accomplish this purpose, we have developed a novel data cube model. The model consists of six dimensions which are customer demographic data, customer visits, mouse movements, online shopping cart, external data and time dimension. To detect customer gender we used artificial neural network model. Our results show that using the derivatives of the data cube and the model, gender of an online customer may be predicted with up to 80% of success rate. In the study we have also applied a data mining decision tree analysis in order to find the most significant parameters for detecting an online customer's gender. Our analysis shows that time spent on the site, average time intervals between clicks, items clicked and order of the clicks are important and can be used to predict online user gender and it may be used for promotional and marketing purposes.

Keywords - Gender, Prediction, Artificial Neural Networks, Data Mining, Marketing.

I. INTRODUCTION

Data mining is to extract useful information from large data sets[1]. The data used for data mining may come from a corporate database, search engine queries, network traffic or communication logs, social media comment, customers' online service/product reviews, online newspaper headlines etc. If the data mostly come from web sources like internet sites, social media and so on, the mining process is named as web mining[2]. No matter where the data come from or it is called as data mining or web mining, the data to be used must be restructured for a fast and efficient usage. This special form is data warehouse [3]. Creating data warehouses is one of the early steps of data mining process. When data warehouse is created or raw data set is rendered into a data warehouse format, the data should be cleaned and reformatted [4]. For web mining data from web sources should be collected and rendered into a new format and sometimes they are blended with data from other sources. A data warehouse should be subject oriented[4]. This means web data blended with other data must be reformatted in accordance with a predetermined subject. For example, if you are planning to use product review entries to cluster customers as satisfied or unsatisfied, words and phrases taken from web must be collected and re organised accordingly. Since a data warehouse should carry a timestamp, time values like day, hour, minute or even seconds must be attached to data[5]. After data warehouse is ready web mining analysis may start. For web mining there are various models, tools and algorithms. For a predictive mining, classification model may be used with decision trees, genetic algorithms and artificial neural networks. Algorithms like K-means [6], [1], K-Medoids [7], DBSCAN[8], [1] and K- Nearest neighbour[9] may be used for clustering model. Apriori and AprioriTID[10], [9]algorithms are to be used to find out the relation between data items.

In this study we performed a web mining application using clickstream data of online customers. For mining, we used both artificial neural networks and decision tree models to find out elements to help predict online customer gender.

II. RELATED WORK

With the increase of online shopping, customer behaviour has gained a lot of importance. E-commerce firms started to be interested in examining online customer behaviour and find information in order to use it for various business purposes. There have been some case studies on examining online customer behaviour. To understand what customers demand, like and hate, not only demographic data[12] but also mouse movements of customers are used and analysed. This may be done with visual analysis using plots, diagrams, maps etc. or any other predictive analysis tools or techniques [11]. As it is well known, word-of-mouth affects way of doing business and making shopping decisions. In the course of development of Internet, the form of word-of-mouth may be considered as electronic- word-of-mouth or e-WOM; in this sense the channel of e-WOM and personality characteristics of online shoppers have been collected and analysed in order to find out what kind of customers an e-tailer encounter with [13] .WOM, e-WOM and web site characteristics affect customer loyalty in B2B (Business to Business) as well. There are studies on detecting the influence of web site characteristics over customer loyalty [14], [15]. To measure customer satisfaction and build an automated model to spot satisfied and unsatisfied customers, online firms analyse customers' reviews with a data mining featured system. A robot may extract the most

representative words and phrases from customers' reviews and analyse them to find out which customer is satisfied or not satisfied with the product s/he has bought[16]. Besides decision trees, artificial neural networks, parallel sequence mining algorithm [19] and support vector machines are also very useful for web mining [18].A similar study held for 'search engine query entry words' have yielded satisfying results [17]. Online buyers' and window shoppers' data are analysed and compared to extract the reason behind why and when some customers switch from window shopping to a real purchase transaction and sometimes do not [25].There are many other case studies and model proposals to analyse online customer behaviour, cluster online users, form a customer profile, understand the needs of customers shopping or just online window shopping and so on[20], [21], [22]. All these studies and many others have been performed in order to find useful patterns and information about customers to be used for marketing, management, production and so on [24].However, none of the studies above cover customers' characteristics and features like gender, age or social status.

III. DATA CUBE MODEL

For any kind of web mining (or a clickstream data mining as it is in this study) a special data warehouse is needed. The data warehouse which we introduce here is a novel one for e-commerce web sites or online marketing firms. We suggest a six dimensional data warehouse for a clickstream analysis. Our model is a cube-shapeddata warehouse which may be used for information extraction or data mining.

The dimensions of the data cube are

1. Customer demographic data dimension.
2. Customer visits dimension.
3. Online market basket dimension.
4. Time dimension.
5. Customer mouse movements dimension.
6. External data dimension.

Remember that contents, data fields and other details of each dimension may differ from one application to another. Fig. 1 shows the data cube.

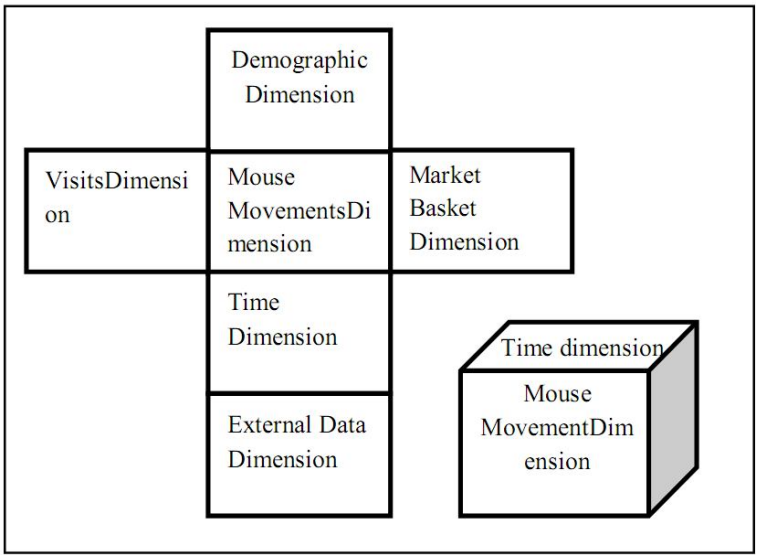

Fig.1. Data cube model for a clickstream data analysis.

Now let's take a look at each data dimension briefly.

Customer demographic data may be extracted from the firms' customer databases. Today, almost all online firms collect this kind of data. In this dimension, several fields may be employed. Some of them are age, neighbourhood, marital status, number of children, job, tenure etc. These items partly depend on the aim of analysis and partly on data available. Customer visits dimension may include last visit time, average visit duration, total payment, number of items bought so far and so on. Online market basket dimension may consist of product sex, colour, type, size of the product, number of items, cost of items, discounted or promoted items and so on. Time dimension may have hour, day, month, season and duration of shopping etc. Mouse movements data dimension must be extracted from client side browsers. These data may include items such as type of item clicked, colour of item clicked, pages visited, order of each click, mouse idle time etc. External data dimension includes outside data that affect online shopping and customer behaviour. Outside temperature, GPS location that customers connect, type of the device they use (i.e tablet, PC, Laptop), operating system may be some of them.

IV. ANALYSIS

4.1. Data

In this study we used a derivation of the data cube introduced above. The data set which has been used comes from an e-commerce company. The company

sells consumer products online and has more than 15 pieces of sales on average every day. With the permission of the company we installed a program on the server and also used 'client side java scripts' to collect clickstream data of online shoppers. For 280 days, data were collected on a server side database. Total records collected and organised are 20.000. Within 280 days there were more than 20.000 visits paid to the site. However, majority of the users were women, so we deliberately and randomly selected 10.000 visit records from men and another 10.000 visits made by women. Thus, our man/woman ratio was one to one in the datawarehouse.

The data dimensions and fields used in the analysis are as follows:

For Time Dimension, fields collected and used are total time spent on the site, first item add time (if there is one), period of the day (as morning, afternoon, evening hours or late at night), day of the week.

Mouse Movements Dimension consists of count of all clicks made, order of clicks (as 1st, 2nd, 3rd etc.), category of clicks (three categories: dress- jacket-sweater, shoes-boots-sandals or underwear) and average time spent between clicks and time spent before each click (as seconds).

Visits Dimension has got two data fields; they are search (indicates if user has conducted a product search) and category of search.

Market Basket Dimension is made up of product sex category in basket (as child-boy, child-girl, unisex, woman, man), number of items in basket, discounted item in basket (whether there is one or not).

For the study External Data Dimension has not been used.

For Demographic Data Dimension we used only gender field which is the class variable at the same time.

4.2. Method and Tools

For the analysis KNIME data mining program [23]has been used. KNIME is an open source, Eclipse based program designed for data mining and machine learning applications. The program has various clustering and classification algorithms. Since other open source programs can be embedded and used inside KNIME, it is quite useful. For the analysis and prediction we used artificial neural networks and decision tree algorithms.

For decision tree analysis we used C4.5 algorithm [24]. The algorithm uses entropy function, as it is in Equation 1, in order to find the best fields to create branches, when constructing the tree.

$$H(p_1, p_2, ..., p_n) = \sum (p_i \log(1/p_i)) \qquad (1)$$

For artificial neural network analysis we used Multi-layer Perceptron predictor. The artificial neural network is made up of ten hidden layers which consist of ten nodes. Throughout the network sigmoid activation function has been used. For both decision tree and artificial neural network analysis we divided the data in two parts with a 70%/30% ratio. Namely, 70% of data, 14.000 records have been used for decision tree and multi-layer perceptron learning; 30% of data, 6.000 records have been used for prediction. The success of the predictions have been measured and assessed in a confusion matrix. The model for analysis is depicted in Fig.2.

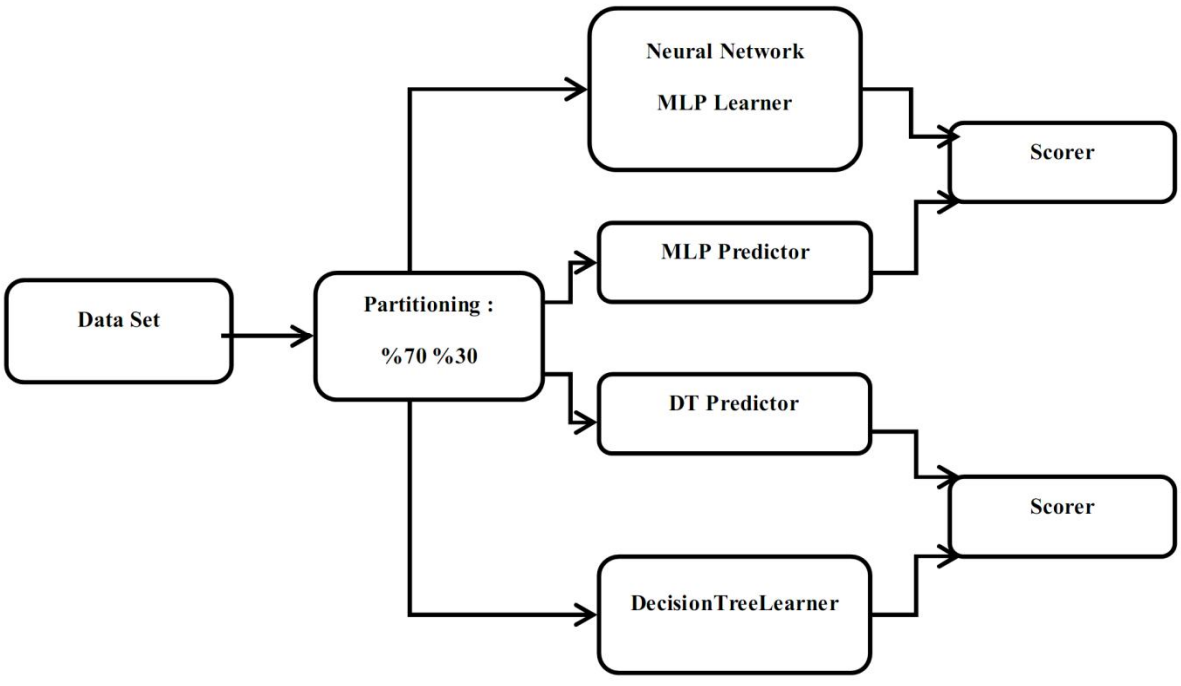

Fig.2 The model used for analysis.

TABLE 1 CONFUSION MATRIXES FOR SCORER NODES

Overall Accuracy:	77.23%
Overall Error:	22.77%

Gender Prediction **(Artificial Neural Networks)**	Predicted Man	Predicted Woman	Accuracy	Sensivity	Specifity	F-Measure
True Man	7890	2110	78.9%	49%	52.8%	0.871
True Woman	1886	8114	81,1%	51%	47.1%	0.710
Correct classified:	16004					
Wrong classified:	3996					
Overall Accuracy:	80.02%					
Overall Error:	19.98%					
Gender Prediction **(Decision Tree)**	Predicted Man	Predicted Woman	Accuracy	Sensivity	Specifity	F-Measure
True Man	7555	2445	75.6%	49%	53.6%	0.852
True Woman	2109	7891	78.9%	51%	46.3%	0.640
Correct classified:	15446					
Wrong classified:	4554					

4.3. Findings

Decision tree data mining performance has given more than twenty-five patterns. Patterns consist of nodes between two and five. That means in some cases two pieces of data or click information may give an idea about the gender of the user. For this, maximum number is five,however, in most of the cases click numbers are as high as twenty-three, namely the system has to wait the 23rdclick to decide about the gender of user.

Our findings show that using the proposed model and a proper data mining model or algorithm, one could predict the gender of online users up to 80% probability.

Table **1** shows confusion matrix of both decision tree and artificial neural network nodes (see Fig.2).

Table **1** shows some statistics related to artificial neural network and decision tree predictions. The table suggeststhat both models may be used for predictions with a minimum 0.640 F value. Indeed, accuracy statistics are all over 75.6% and overall error is 19.98% for artificial neural networks prediction and 22.77% for that of decision tree.

Some of the patterns which are produced by decision tree algorithm are as follows:

Rule 1:

If 3rd click is dress, jacket or sweater,
8th click is shoes, boots or sandals,
23rd click is click is dress, jacket or sweater
then
Gender prediction is man, (confidence 80%, lift 3.4).

Rule 2:

If 2nd click is underwear,
8th click is underwear,
16th click is dress, jacket or sweater,
23rd click is click is dress, jacket or sweater,

Search is True (customer searched a certain category or product),
Then
Gender prediction is man, (confidence 73.2%, lift 2.5).

Rule 3:

If 4th click is shoes, boots or sandals,
5th click is shoes, boots or sandals,
6th click is shoes, boots or sandals,
Time interval before 7th click is >100 seconds
Then
Gender prediction is woman, (confidence 70.4%, lift 4.5).

Rule 4:

If 3rd click is underwear,
4th click is underwear,
10th click is underwear,
Average time interval between clicks <150 seconds
Then
Gender prediction is man, (confidence 75.3%, lift 3.5).

Rule 5:

If day is Friday,
1st click is underwear,
3rd click is underwear,
Average time interval between clicks < 200 seconds
Then
Gender prediction is man, (confidence 82.0%, lift 2.6).

Rule 6:

If 20th click is underwear,
21st click is underwear,
23rd click is underwear,
25th click is underwear,
Average time interval between clicks < 40 seconds
Then
Gender prediction is woman, (confidence 87.0%, lift 4.1).

Rule 7:
If day is Thursday,
Clicks from 1st to 5th are shoes, boots or sandals,
Then
Gender prediction is woman, (confidence 86.0%, lift 4.2).

Rule 8:
If total time spent > 1700 seconds,
Average time spent < 90 seconds,
6th click is skirt, jeans, shorts,
Then
Gender prediction is man, (confidence 75.0%, lift 3.7).

Rule 9:
If average time spent > 85 seconds,
Average time spent < 170 seconds,
First 7 clicks are skirt, jeans, shorts,
Then
Gender prediction is man, (confidence 78.4%, lift 3.5).

Rule 10:
If there is a discounted item in the basket,
First item add time is < 100,
Day is Sunday,
First 5 clicks are shoes, boots or sandals,
Then
Gender prediction is woman, (confidence 88.1%, lift 5.5).

CONCLUSION

In this study we have introduced a novel data cube model. The data cube model which is a data warehouse (or data mart) is to be used for clickstream data mining. In the study we realised the data cube model by collecting data from an e-commerce firm which operates as an on-line retailer. In the study the data cube model has been used to predict the gender of online users while they are shopping, surfing on the site or just window shopping. For predictions, artificial neural networks (ANN) and decision tree (DT) algorithms have been used. After the learning processes (ANN and DT) and predictions, a scoring node has been used to measure the accuracy and reliability of predictions. The tests suggest that predictions are accurate enough to be used for business purposes such as marketing, production or general management.

As it can be seen from the patterns, the most important elements to decide the gender of users are:

- Average time spent before making a categorized click.
- Total time spent on the site.
- Average time spent between clicks.
- Item clicked i.e. shoes, pants, underwear etc.
- Order of the item clicked i.e. 1st, 2nd, 3rd etc.
- Day of the visit
- Existence of a discounted item in the basket.

However it must be stressed that the patterns and results found through this study is valid only for this data and the period the study was held. What we introduce here is to reliability, accuracy and feasibility of predicting online customer gender via proper data mining models and a decent data cube. The study may be repeated with similar data warehouses to predict not only genders but also, ages, social statuses, income levels or any other features of online customers.

REFERENCES

[1] SilahtaroğluGökhan, VeriMadenciliğiKavramveAlgoritmaları, PapatyaYayıncılık, 2013.

[2] Russell, Matthew, A. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More, O'REILLY MEDIA, 2013.

[3] Corr, Lawrence and Stagnitto Jim, Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema, DecisionOne Press, 2011.

[4] W. H. Inmon, Building the Data Warehouse, Wiley, 2005.

[5] Vaisman, Alejandro and Zimányi, Esteban, Data Warehouse Systems: Design and Implementation (Data-Centric Systems and Applications), Springer,2014.

[6] MacQueen, J. B. , Proceedings of Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press. pp. 281–297, 1967.

[7] Kaufman, L. and Rousseeuw, P.J. , Clustering by means of Medoids, in Statistical Data Analysis Based on the –Norm and Related Methods, North-Holland, 405–416, 1987.

[8] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei, A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)., AAAI Press. pp. 226–231, 1996.

[9] Everitt, B. S., Landau, S., Leese, M. and Stahl, D., Miscellaneous Clustering Methods, in Cluster Analysis, 5th Edition, John Wiley & Sons, Ltd, Chichester, UK, 2011.

[10] Rakesh Agrawal and RamakrishnanSrikant, Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[11] Pragarauskaite, Julija; Dzemyda, Gintautas, Visual decisions in the analysis of customers online shopping behavior, NONLINEAR ANALYSIS-MODELLING AND CONTROL Volume: 17 Issue: 3 Pages: 355-368, 2012.

[12] Phang, Chee Wei; Kankanhalli, Atreyi; Ramakrishnan, Karthik; et al., Customers' preference of online store visit strategies: an investigation of demographic variables, EUROPEAN JOURNAL OF INFORMATION SYSTEMS Volume: 19 Issue: 3 Pages: 344-358 Published: JUN 2010.

[13] Li, Chunqing; Ding, Jianlan; Zhu, Zhian , Online Word-of-Mouth: Motives and Channels Based on the Personality Characteristics of Customers, Proceedings of 8th Wuhan International Conference on E-Business, Wuhan,CHINA, 2009.

[14] Hsu, Li-Chun; Wang, Kai-Yu; Chih, Wen-Hai, Effects of web site characteristics on customer loyalty in B2B e-commerce: evidence from Taiwan, SERVICE INDUSTRIES JOURNAL Volume: 33, Issue: 11, Pages: 1026-105, 2013.

[15] Sun, Ying; Su, Xuan; Jiao, Aiying, Research based on the Web Shopping Service Quality and Customer Satisfaction and Loyalty, Proceedings of International Conference On Economic, Business Management And Education Innovation (EBMEI 2013), VOL 17, Volume: 17, Pages: 212-218, 2013.

[16] Wang, Dingding; Zhu, Shenghuo; Li, Tao ,SumView: A Web-based engine for summarizing product reviews and customer opinions, EXPERT SYSTEMS WITH APPLICATIONS Volume: 40, Issue: 1, Pages: 27-33, 2013.

[17] Ortiz-Cordova, Adan and Jansen, Bernard J., Classifying Web Search Queries to Identify High Revenue Generating Customers, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, Volume: 63, Issue: 7, Pages: 1426-1441, 2012.

[18] Sun, Lei and Duan, Zhu , Web Potential Customer Classification Based on SVM, Proceedings of 2012 INTERNATIONAL CONFERENCE ON INDUSTRIAL CONTROL AND ELECTRONICS ENGINEERING (ICICEE), Pages: 568-570, 2012

[19] Demiriz, A, webSPADE: A parallel sequence mining algorithm to analyze web log data, Proceedings of 2002 IEEE INTERNATIONAL CONFERENCE ON DATA MINING, Pages: 755-758, 2002.

[20] Tuzhilin, Alexander, Customer relationship management and Web mining: the next frontier, DATA MINING AND KNOWLEDGE DISCOVERY Volume: 24, Issue: 3, Special Issue: SI , Pages: 584-612, 2012.

[21] Knezevic, Blazenka; Renko, Sanda; Bach, MirjanaPejic , Web as a customer communication channel in the confectionery industry in South Eastern European countries, BRITISH FOOD JOURNAL, Volume: 113, Issue: 1, Pages: 17-36, 2011.

[22] Sanchez-Franco, Manuel J. and JavierRondan-Cataluna, Francisco , Virtual travel communities and customer loyalty: Customer purchase involvement and web site design, ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS, Volume: 9, Issue: 2, Pages: 171-182, 2010.

[23] Berthold MR., **KNIME: the Konstanz information miner** In *Data analysis, Machine Learning And Applications*, Springer-Verlag, Pages: 319-326, 2008.

[24] Quinlan,J.Ross, Simplifying decision trees, International Journal of Man-Machine Studies,issue: 27(3), (pp. 221 – 234), 1987.

[25] Mu Jiankang , An Empirical Study on Customer Conversion Behaviour of Online Window Shopper, Proceedings of International Conference on Engineering and Business Management, CHINA , VOLS 1-8 Pages:2402-240, 2010.

★★★