User Profile Based Behavior Identification using Data Mining Technique

Mr. Yogesh Kulkarni[1] Mr. Rushikesh Mule[2] Mr. Akash Lakade[3] Mr. Subodh Kulkarni[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Sinhgad Academy of Engineering, Pune, India

Abstract— In regular retail shop the shopkeeper may predict the behaviour of customers using their facial expressions which can results in increase in their sell. However, while considering online shopping it is not possible to see and analyse customer behaviour such as facial expressions, products they check or touch etc. In this case, click streams or the mouse movements of E-Customers may provide some hints about their buying behaviour. In this, we have presented a model to collect, analyse click streams of E-Customers and extract information and make predictions about their shopping behaviour on an online shopping market place. The model we present predicts category of most likely bought products on a digital market place by the customer and according to that it gives recommendations of products to the E-customers. We are also going to provide offers on items added in basket of most likely bought category by customer through basket analysis. For analysis and prediction we are going to use Naive Bayes algorithm. Result of this analysis can be used in Customer Relationship Management and Business Intelligence.

Key words: Data Mining; Naïve Bayes; Clickstream; User Profiling; Online Shopping Market

I. INTRODUCTION

One of the basic advantages of using the digital market is that it offers more number of choices, at lower prices, easy search and also provide access to online customers. Hence, the digital market is expanding day by day. As a result of this customers behaviour analysis and prediction are gaining more importance in buying or not buying products online. It is important to study the E-customers behaviour, so we can predicts about their behaviour in digital market. The behaviour of customer is the study of when, why, how and where customers buy a products or not. Understanding about what customer actually need is important while building an online e-commerce application.

Data mining is the process of discovering meaningful pattern through large amount of datasets which are stored in data warehouses. The key idea behind use of data mining technique to classify the customers data according to the posterior probability. So in this model the data mining concept is used to perform the classification on training data set and also use for prediction.

Benefits of Data mining in E-commerce:

A. Customer Profiling

Customer profiling is also known as customer-oriented strategy in E-commerce site. This strategy allows companies to use business intelligence through the mining of customers' data to plan their business activities and operations. It also helps to develop new research on products or services for e-commerce prosperous. Analysing and classifying the customers' data can help companies to lessen the sales cost. Companies can also use users browsing data to identify whether they actually shopping or just browsing through online sites or buying something. As a result of this companies can plan about their strategy and improve their infrastructure.

B. Personalization of Services

Personalization is the act of providing services and contents to individuals on the basis of history data available in the database.

C. Basket Analysis

Market Basket Analysis is a common retail, analytic and business intelligence tool which helps retailers to know their customers behavior better. This helps them in their future strategies. There are lots of ways to deal with the market basket analysis which contains: Identification of product affinities; tracking not so apparent product affinities and leveraging on them is the real challenge in retail.

D. Sales Forecasting

Sales forecasting is the process that involves the aspect of an individual customer spend time on online site to buy an item or product and in this process it is trying to predict if the customer will buy an item again or not.

E. Market Segmentation

Market segmentation is the best use of data mining technique. The data which gotten from various sources, it can be broken down in various and meaningful segments such age, gender, name, phone no., occupation of customers etc. As result of the market segmentation can also help a organization or company to identify its own competitors. Segmentation of the database of a retail company will also improve the conversion rates. With the help of that company can focus their promotion on a close fitted and highly wanted market.

Finally, in our model we will be dynamically generating the web data of customers and analysis will be performed based on some attributes that are mentioned in the section V. According to that analysis, prediction will be performed. This study introduces a new model and an application on prediction of customer behavior using click stream data which is helpful for further business implementations.

II. MOTIVATION

Click streams are the mouse clicks a user makes when they are surfing on internet can tell us a lot about their behavior if it is analyzed in a right way. We can consider this analysis as web mining approach to discover patterns in the surfing of web contents. By analyzing users' navigation patterns and their relation with web content one can redesign a website or e-business along with the behavior of the online users of it. There are a number of studies going on collecting and analyzing web content, text mining, data mining and also click stream data analysis. We know that a proper web mining may shows useful results about the quality of a web

page. Some issues like web site performance, its quality and online marketing intelligence can be carried out with data mining techniques such as classification, clustering, analysis and prediction etc. Companies can yield a lot by analyzing the relation between customers and products on the online shopping market if they have made a proper platform, especially a web based platform which is suitable for web mining. Classification models can be used for this purpose in a better way.

So, in the sense of customer behavior, a detailed customer profile can be created through such an analysis. So, before starting a web mining or click stream analysis, analysts need to build a model with a proper database or data warehouse. The data warehouse will play a major role in web mining model. This study will cover some important works on data mining technology applied to e-commerce.

III. OBJECTIVES

The objectives of customer behavior are in the fact, that we can improve our sales if we study the customers. We can alter the way we sell our products depending on the ways that customers buy them. Continuous observation of consumer behavior can enable you to find out their interest which can in turn help you to recommend products of their interested category to the ultimate satisfaction of e-customers. As the market trend shifts, a consumer analysis will be the first indicator of the same. Whether it is demand forecasting or sales forecasting, both of them are possible and therein lay the importance of consumer buying behavior. The primary objective of this study is to increase sells of e-commerce through customer behavior analysis by finding loyalty or interest of customers in specific category of products and providing recommendations of products of interested category. We are also going to provide offer on product added in cart by e-customer of interested category. This way we are going to provoke the e-customer to buy the products which will result in increase in sells of e-commerce.

IV. LITERATURE SURVEY

1) ” Analysis and prediction of e-customers’ behaviour by mining clickstream data”, Gokhan Silahtaroglu, Hale Donertasli, 2015 ieee international conference on big data (big data)- in this paper, author have presented a model to analyze clickstreams of e-customers and extract information and make predictions about their shopping behaviour on a digital market place. the model predicts whether customers will or will not buy their items added to shopping baskets on a digital market place. for the analysis, decision tree and multi-layer neural network prediction data mining models have been used.
2) ” Analysis of the internet using behaviour of adolescents by using data mining technique”, Chonnikarn Rodmorn, Mathuros Panmuang, Khuanwara Potiwara, 2015 7th international conference on information technology and electrical engineering (icitee)-the author has investigated the association rule of upbringing of parent to affect behaviourand experience of internet using of teenagers by the apriori algorithm.
3) ” Efficient association rule mining algorithm based on user behaviour analysis for cloud security auditing”, Chunye Zhao, Shanshan Tu, Haoyuchen, Yongfeng Huang, 2016 ieee-apriori and fp-growth algorithm are used for finding associations between product and customer transaction.
4) ” users profiling using clickstream data analysis and classification”, Wedyan Alswiti, Ia'far Alqatawna,Bashar A l-Shboul, Hossam Faris,Heba Hakh, 2016 cybersecurity and cyberforensics conference-author proposed a framework to extract features based on the sequences of api calls and frequency of appearance and identifies malware by using k-nearest neighbor algorithm.
5) “The analysis and prediction of customer review rating using opinion mining”, Wararat Songpan, 2017 ieee sera 2017, june 7-9,2017, london, uk-this paper proposes the analysis and prediction rating from customer reviews who commented as open opinion using probability's classifier model. this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques.

V. DATASETS USED IN STUDY

Dataset has a server side program which is used to collect data from the organization’s web server or company’s web server; also at the same time, another program called java script has been used to collect data from client side. Data attributes used in the study are given as follows:

A. Day and Date

This attribute represents the day of week days and date on which user visit the site.

B. Amount of the Time Spent on the Site

This attribute includes total amount of time spent on the site and is calculated in seconds.

C. Search

This variable represents what the customer search on site and the customer searches a certain product on the site by just typing the keywords, this variable takes 1 as true value and 0 as false value.

D. Category id

Products on the site have been categorized into electronics, cloths, shoes, etc. And one unique id is provided for each category such as electronics as 1, cloths as 2, shoes as 3, etc.

E. Number of Products in the Shopping Cart

This attribute includes the number of different products in the cart. The products in the cart may be of same category but they may differ in their size and colour.

F. Product Category of the Item in the Basket

There are five categories in this field: Female, male, unisex, child (girl), child (boy).

G. Buy Count

This show how many products or item of specific category bought by customers.

H. Click Count

When user click on particular item, its entry will be calculated as click count. We are going to count only the click count made on the item or products.

VI. METHODOLOGY

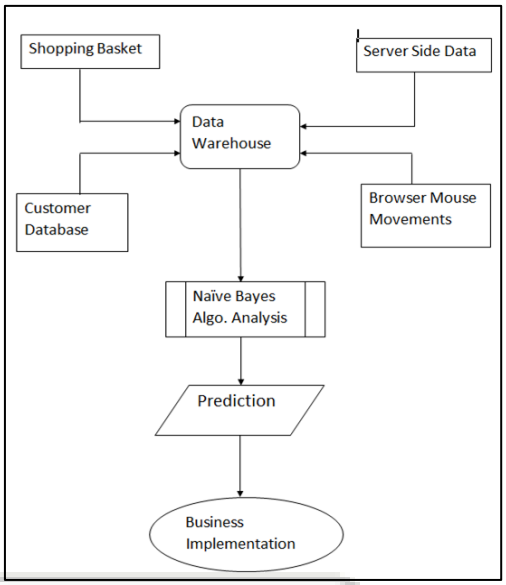

Fig. 1: Overall Model for Applied Analysis

VII. PROPOSED SYSTEM

Figure shows that there are three models such as the admin module, the client module and the server module. The customers can open site and perform various functions such as making registration, login, and search products, like/dislike, click and view products. The server module maintains the activity log of the user. The admin module is used to give offers based on the analysis and prediction performs.

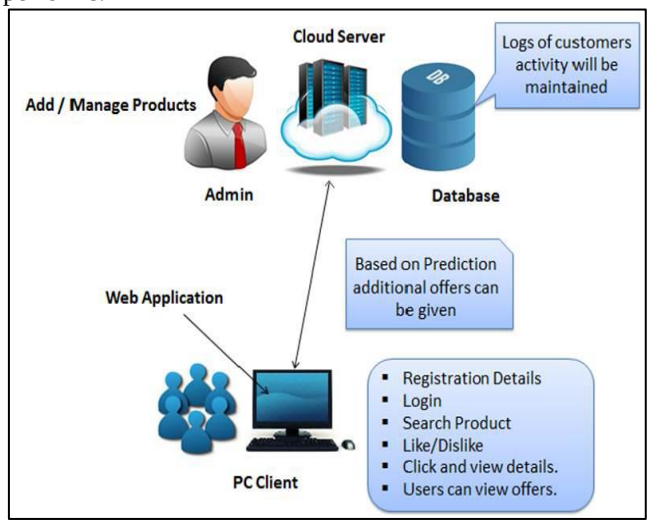

Fig. 2: Architectural Overview

A. Admin Module

Admin module consists of two phases. The first phase consist of add and manage products where admin is able to add, delete and manage products. And second part is analysis part where the actual algorithm is being implemented. There is connectivity between the admin module and the database which is used in the system. Based on the prediction made offers will be given to the individual interested customers. Admin at the admin site will be able to execute queries on the database for managing the products such as add, delete, update.

B. Client Module

Client module is used for the customers activities such as registration, searching, viewing, etc. Firstly, Clients will get registered to the application. After that when they gets logged in into the application, they will start searching for products. The customers will search the products based on the name, category, like/dislike, rating of a particular product.

C. Server Module

The server which we have use is cloud server. The cloud server will be having GlassFish server which will host the web service. Server is responsible for user's authentication and the server also provides the services requested by the users. It also maintains the users logs based on the clicks and the activity. JDBC connectivity is being use for connecting the database where the database is MySQL. Server will apply prediction on logs to analyse the user behaviour and will predict if customer will going to buy product or not in cart.

VIII. ANALYSIS AND PREDICTIONS

Data classification process includes two steps. First step consists of learning process where the training data is analysed by a classification algorithm. The second process is classification; here test data is tested against classification algorithm to evaluate the accuracy of the classification algorithm. When learning is completed this model is then use to classify data into different classes. As in our case the e-customers are classified into classes of interested categories such as electronics, clothing and accessories, sports collectibles, beauty products, books etc. For achieving this first we have to make analysis of e-customer's behaviour from their history data. Here for analysis and prediction of class labels for e-customers we are going to use Naïve Bayes classifier as specified follow:

A. Bayes Rule

A conditional probability is the possibility of some conclusion, C, given some observation, E, where a dependency exists between C and E.
This probability is denoted as P(C |E) where,

$$P(C|E) = \frac{P(E|C)\,P(C)}{P(E|C)}$$

B. Naive Bayesian Classification Algorithm

Bayesian classifiers perform statistical classification. They are used to predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is basically based on Baye's theorem or rule.
The Naive Bayesian classifier works as follows:

1) Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-

dimensional attribute vector, $X = (x_1,x_2,\ldots,x_n)$ shows n measurements on the tuple from n attributes, respectively, A1, A2,.., An.

2) Let's assume that we have m classes C_1, C_2,…, C_m. Given a tuple, X, the classifier will predict class of X which is having the highest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \le j \le m, j \ne i$$

The C_i class for which $P(C_i|X)$ is maximum, the tuple X belongs to that class. By Bayes' theorem,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

3) As P (X) is constant for all classes, only $P(X|C_i)\,P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the probability of classes are equally likely, that is, $P(C_1)=P(C_2)=\ldots=P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_i, D|/|D|$, where $|C_i, D|$ is the number of training tuples of class C_i in D.

4) If the given data sets have many attributes, it may take expensive computation to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made. This assumes at first that the values of the attributes are conditionally independent of one another, given the class label of the tuple. Thus,

$$P(X|C_i)=\prod_{k=1}^{n}(X_k|C_i)$$
$$=P(X_1|C_i)*P(X_2|C_i)*\ldots*P(X_m|C_i)$$

We can easily estimate the probabilities $P(X_1|C_i)$, $P(X_2|C_i)$,… ,$P(X_m|C_i)$ from the training tuples. Here X_k refers to the value of attribute A_k for tuple X. For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute $P(X|C_i)$, we consider the following:

- If A_k is categorical, then $P(X_k|C_i)$ is the number of tuples of class C_i in D having the value X_k for A_k, divided by $|C_i, D|$, the number of tuples of class C_i in D
- If A_k is continuous valued, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined b.

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

So that,

$$P(X_k|C_i)=g(X_k,\mu_{C_i},\sigma_{C_i})$$

We should compute $\mu_{C_i}$and σ_{C_i}, which are the mean and standard deviation, of the values of attribute A_k for training tuples of class C_i. After that we are going to put these in above equation

5) In order to predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \le j \le m, j \ne I$$

In other words, the predicted class label is the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

IX. Conclusions

In this study, application analyse the massive volume of customer data and classify them based on the customer behaviours and Naive Bayes algorithm helps for considering different attributes which required for analysis. Also it gives accurate results for large amount of datasets. This kind of customer behavior analysis will directly produce increase in sells of e-commerce application. This application helps shopping effective and easy.

X. Future work

The application which we have built can also be implemented on Android platform as well because most of the customers use smart phones for daily purposes. Also the algorithm for carrying out the same task can be done using a hybrid approach. We can also provide location base offers to customers. If the number of users for such application goes on increasing this also can be implemented on Hadoop platform.

References

[1] GokhanSilahtaroglu, Hale Donertasli, "Analysis and Prediction of E-Customers behavior by Mining Clickstream Data.", in 2015 IEEE International Conference on Big Data (Big Data)

[2] Chonnikarn Rodmorn, Mathuros Panmuang, behavior of Khuanwara Potiwara, "Analysis of the Internet Using behavior of Adolescents by Using Data Mining Technique." in 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, Thai

[3] Chunye Zhao, Shanshan Tu, Haoyuchen, Yongfeng Huang, "Efficient Association rule mining algorithm based on user behavior analysis for cloud security auditing", in 2016 IEEE.

[4] Wedyan Alswiti, Ja'far Alqatawna,Bashar Al-Shboul, Hossam Faris,Heba Hakh, "Users profiling using clickstream data analysis and classification" in 2016 Cybersecurity and Cyberforensics Conference

[5] Wararat Songpan, "The Analysis and Prediction of Customer Review Rating Using Opinion Mining." In 2017 IEEE.

[6] C. M. Fong, Baoyao Zhou, S. C. Hui, Guan Y. Hong, and The Anh Do, "Web Content Recommender System based on Consumer BEHAVIOUR Modeling." In IEEE Transactions on Consumer Electronics, Vol. 57, No. 2, May 2011.

[7] Fauzan Burdi,Anif Hanifa Setianingrum,Nashrul Hakiem,"Application of the Naive Bayes Method to a Decision Support System to provide Discounts" in 2016 6th International Conference on Information and Communication Technology.

[8] Huma Parveen,Shikha Pandey,"Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm" in 2016 IEEE.

