Indian J.Sci.Res. 17(2): 29 - 32, 2018 ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

CLICKSTREAM DATA PREDICTION USING RANDOM FORESTCLASSIFIER

[1]Dr.B.Sateesh Kumar, [2]MsB.Sindhuja
[1,2]Associate Professor, Computer Science Engineering, JNTUHCEJ, Jagtial

Abstract-Now a day s e- Commerce websites are on the fly. Since the e- commerce websites are plenty there is high competition among the competitors to draw the attention of a site visitor. What makes the site visitor research a product and buy it? What product do visitors tend to buy together, and what they are most likely to buy in the future. Where should a competitor spend resources in such a way that the user experience will get enhance. To mine the knowledge of a site visitor, there is a solution called click stream data. Click stream data is a record of an individual's movement through time at a website and it contain information like time, URL content, user's machine address previous URL visited and Browser type. Random forest algorithm makes use of click stream data to predict the user's buying interests. Random forest is a tree based classifier which consisting voting of several decision trees. Random forest classifier algorithm plays a major role in classification to classify the data with minimal percentage of error rate.

Key words: e-commerce, Click stream data, Classification, Random forest classifier, Decision tree.

I. Introduction

Due to this increase in significance and market share, e-commerce companies have to adopt new strategies that fit the needs of the online customers. This type of customers have different behaviors than the physical ones. Furthermore, web users have a much easier job in comparing different companies through simple online queries, making it hard to maintain customer loyalty and retention the same way it is done in the traditional stores.Predicting user intentionality towards a certain product, or category, based on interactions within a website is crucial for e-commerce sites and ad display networks, especially for retargeting. By keeping track of the search patterns of the consumers, online merchants can have abetter understanding of their behaviors and intentions[1]. e-commerce marketers attracting consumers by giving the offeron choosing product with the help of click stream analysis by using the Random forest classification algorithm. Click stream analysis is the process of collecting, analyzing and reporting aggregate data about which pages are visited by a site visitor and in what order they visits the site. Clickstream is the path the visitor go through it. So the order in which the site visitor visits the webpages, based on that online e-commerce marketer can analyze the site visitor behavior on which product the site visitor showing interest. Therefore, it is important to be able to predict purchasing behavior early during the visiting process, in contrast to previous research that predicted purchase behavior at the end of the visit. Allthe existing researches regarding Clickstream based purchase prediction modeling have investigated predicting purchasing behavior throughout the Clickstream. So given that predicting online purchasing behavior throughout the visitors' Clickstream and usage of Random forest classifier is a relatively unknown area, this research will act as an important first step to gain more insight into this topic.

II.Related work

There is a tremendous change in the consumers vision on way of purchasing, showing much interest in buying the products online rather than the offline because of their busy lives. It is not an easy task for e-commerce website marketer to predict the consumer behaviour, because the needs and goals of online consumer may change dynamically based on the certain factors such as low cost, demand of a product that they choose and the budget. To get attention of site visitors towards their products is a primary key to get success of e-commerce websites. Visitors following a directed buying navigational pattern intend to make a purchase and poses substantial information before making the purchase decision. They tend to follow a focused and goal directed search pattern since their search is nearing and making a purchasing decision is nearby. Search also follows a goal directed search pattern but visitors following this pattern have planned their purchase in the near future. However, they have not yet decided which product to buy in a specific category[2]. Since consumersmay change their buying intention in a session it is very difficult to predict the purchasing behavior of a consumer. Wendy W. Moe Peter S. Fader presented a click stream model in simple way and highlighted relationship between site visitor and purchasing conversion. But the limitation is that the data do not reveal when a customer first visits the site[3]. These models are not sophisticated to analyze the consumers buying behavior. To identify user behaviors today, we need a sophisticated Clickstream analysis system that meets three requirements. First, it must scale and function well on large, noisy Clickstream datasets. Second, the

[1]Corresponding Author

system should be able to capture previously unknown user behavior that is capture behavior without categories or labels defined a priori. This is critical, because users often utilize popular services in unexpected ways, and adapting to these behaviors can determine the long-term viability of a service. Finally, the system should be interactive, and help others understand user behavior by presenting detected behaviors in an intuitive and understandable way. The authors of Unsupervised Clickstream Clustering for User Behavior Analysis Proposed a unsupervised method to model online user behaviors. By building and partitioning a Clickstream similarity graph, Capture the detailed user behavior models as hierarchical clusters in the graph. In addition, the tool provided by them automatically produces intuitive features to interpret the meaning of the behavioral clusters[4].

III. Clickstream Analysis

A Clickstream is the recording of what a site visitor clicks on while browsing the web. Every time he or she clicks on a link, an image, or another object on the page, that information is recorded and stored. E-commerce marketers can find out the habits of one individual, but more useful is when marketers record thousands of Clickstreams and see the habits and tendencies of their site visitors. Clickstreams will tell about consumer behavior. E-commerce marketers could record 1,000 Clickstreams and find out where people are clicking and where they aren't and realize that lots of people are clicking to one specific web page. All of this information combined is valuable data that goes beyond normal analytics. Clickstream tracking is ideal for e-commerce websites and websites that depend on ad clicks .Cickstream data helps online websites to identify each customer as a unique individual, assembling the various data points to form a 'data-driven profile' of each user. For any customer centric businesses, it is crucial to know the user behavior patterns. Clickstream data can be used to quantify search behavior using machine learning techniques [8], mostly focused on purchase records. Modern search engines use machine learning approaches to predict user activity within web content.

Path Analysis:

A web site consists of a hyper linked set of pages. One web site can be composed of one up to several thousands of pages. During visit to a web site, a visitor navigates through the site by either clicking on the hyperlinks, or performing internal searches, or using his or her bookmarks to jump to pages and areas of interest. Example of such navigation is shown in below Fig.3.1.

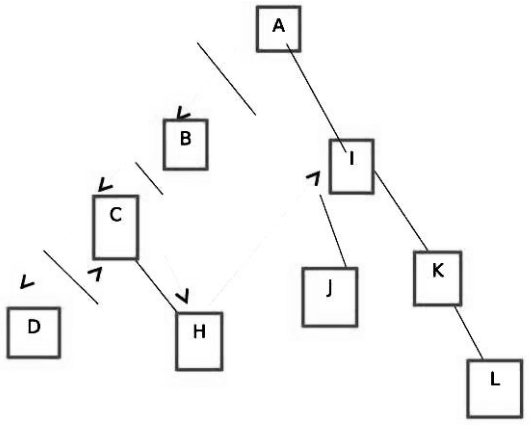

Fig 3.1 A website with sample path

As we can see, the visitor performs different types of moves: forward steps (e.g. from node A to B), backward steps (e.g. from node D to C) as well as forward jump steps indicated by the arrow from H to I. Each of the pages viewed by this visitor is captured in the web log as a separate record[9].

This sequence of moves is known as a "path" or sometimes called "Clickstream" and is recorded in the form of a log file. Paths can be analysed to determine the sequences in which users have navigated the web site. This information is also known as "e - intelligence" and is very advantageous to organizations that are active in e-commerce. The real challenge arises when many visitors navigate through a site that contains thousands and thousands of pages scattered across hundreds of web servers.

The order in which visitors choose to view pages indicates their steps through the browsing (buying) process. The similarities and differences in navigational behaviour of various classes of visitors, such as new visitors vs. repeat visitors, purchasers vs. non purchasers, first time purchasers vs. repeat purchasers can be described by several patterns and could hold clues towards improving the web site design, offer personalization opportunities, and help streamline the e-commerce environment[5].

IV. Random Forest Algorithm

Random forests is a trademark of Leo Breiman and Adele Cutler for an ensemble of decision trees. Random forest algorithm can make use of both classification and the regression kind of problems. Random forest is supervised classification algorithm, it creates more number of trees, higher the number of trees so high accuracy results. When aninputistobeclassified, eachtree classifiestheinputindividually.

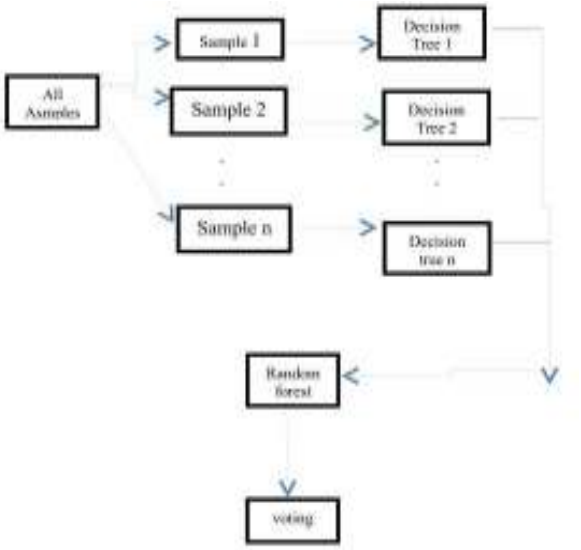

Fig.4.1: Random forest model

The online marketers can apply the classification technique in order to predict the best buying strategies of a consumer based on the data available from Clickstream. Clickstream data consist of information of the pages that recently visited by thecustomer. Construct forest of decision trees with the information available from the cart and Clickstream data information in order to predict the users buying interest. Random forest classification technique can be used with theClickstream data in order to predict the users buying interests more accurately.

Steps for Random forest creation based on Clickstream data

Step 1: randomly select K -attribute from Clickstream data.

Step 2: from the chosen K- attributes select best split.

Step 3: Splits the node into internal node based on best split measurement.

Step 4: repeat the step 1 to 3 until 'L' number of nodes has been reached.

Step 5: build more number of trees by repeating the steps 1 to 4 for n number of times.

Finding the rules, based on the tree generated and compare with the items which are of choice of the site visitor and these items are tested against the training data.

V. Discussions and Analysis

e-commerce marketerscollcect Clickstream data which contains log information of customers.

Example of click stream data:

IP	Country	State	City
172.189.252.8	USA	VA	Dulles
98.29.25.44	USA	OH	Cleveland
68.199.40.156	USA	NY	Freeport
155.100.169.152	USA	UT	Salt LakeCity
38.68.15.223	USA	TX	Dallas
70.209.14.54	USA	FL	Tampa
74.111.6.173	USA	VA	Arlington
128.230.122.180	USA	NY	Syracuse
128.122.140.238	USA	NY	New York
56.216.127.219	USA	NC	Raleigh
54.114.107.209	USA	NJ	Jersey City
74.111.18.59	USA	NY	Syracuse
8.37.70.170	USA	CA	Los Angeles
8.37.70.77	USA	CA	Los Angeles
8.37.70.112	USA	CA	Los Angeles
8.37.70.226	USA	CA	Los Angeles
8.37.70.99	USA	CA	Los Angeles
8.37.71.43	USA	CA	Los Angeles
8.37.71.25	USA	CA	Los Angeles
8.37.71.69	USA	CA	Los Angeles
8.37.71.9	USA	CA	Los Angeles
8.37.71.57	USA	CA	Los Angeles

Table 5.1 iplookup.csv[10].

The above table gives the information about the ip address of a visitor and region concept hierarchy and from the above table we can conclude thatthere are frequent number ofvisitors from los Angeles comparing to all other cities. Similarly we can train a dataset which consist click stream data with attributes

DATE

TIME

USER QUERY

BROWSER

REFERRER

TIME TAKEN

from these attributes the algorithm can learn that the type of user, visiting website and type of items they are going to purchase or visit. Once online business analysts are having

these information they can attract more number of visitors. As per our sample datasets, here are the details about accuracy of Random forest classification algorithm.

Relation1:iplookup.csv

Number of attributes: 4

Number.of instances: 22

	Decision Tree	Random forest
Accuracy	45.4545%	95.45%
Run-time	0 sec	0.01 sec

Relation 2:IIS.csv

Number of attrinutes: 10

Number of instances: 20

	Decision Tree	Random forest
Accuracy	88.88%	94.44%
Run-time	0 Sec	0 sec

Random forests, as compared to Decision trees offers consistent and marked improvements in accuracy particularly true for multiple class classification tasks. This discussion can be extended to more number of instances and anlayze with the existing Decision Tree classification algorithms

VI. Conclusion and Futurework

Clickstream information can be used to find out the customers buying pattrens by making use of classification algorithms. We used the Random forest model to describe the association between general Clickstream information concerning visits and whether a visitor will engage in online purchasing behavior during his visit to the website. Random forest requires far more processing time and omputational burden increases with number of features.

Finally, we wish to stress that, given that the field of Clickstream data research is still in its infancy, Clickstream analaysis will be advantageous for both consumers and online marketers in forthcoming years and much research still needs to be done to get higher benefits in e-commerce website.

Acknowledgment

We thank our colleagues and friends who provided insight and expertise that greatly assisted the research.

References

[1] Predicting online user behaviour using deep learning algorithms Armando Vieira Redzebra Analytics 1 Quality Court WC2A 1HR London, UK armando@redzebra-analytics.com May 27, 2016

[2] https://pdfs.semanticscholar.org/0cc6/25afd38ad2bcdf8552ff784e 5907b0b1b5c8.pdf

[3] capturing evolving visit behavior in Clickstream data by Wendy W. Moe Peter S. Fader 2004.

[4] Clickstream User Behavior Models by GANG WANG ,XINYI ZHANG and SHILIANG TANG, CHRISTO WILSON, HAITAO ZHENG and BEN Y. ZHAO, 2016.

[5] https://dataaspirant.com/2015/01/24/recommendation-engine-part-1/

[6] Fraud Detection Using Random forest Algorithm@ EeshaGoel et al. / International Journal of Computer Science Engineering (IJCSE)

[7] Advances in Computer Science and Ubiquitous Computing: CSA-CUTE2016

[8] Kim, J.B., Albuquerque, P., Bronnenberg, B.J.: Online Demand Under Limited Consumer Search. Marketing Science 29(6), 1001-1023 (2010)

[9] Web Mining & Clickstream Analysis Stefan KolekOezkanKirmaci Fribourg, 2006

[10] https://github.com/mafudge/datasets/tree/master/Clickstream