International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-01, April 2019

# User Behavior Analysis using Machine Learning

[1]Sanat Deshpande, [2]Sandesh Todkari, [3]Pratik Gagare, [4]Nilakshi Mule

[1,2,3]Student, [4]Professor, Government College of Engineering Karad, Maharashtra, India,

[1]sanatdeshpande1@gmail.com, [2]sandesh.todkari21@gmail.com, [3]gagare.pratik.d@gmail.com,

[4]nilmule@gmail.com

**Abstract: In recent years, overwhelming amount of data consisting of user details makes it difficult to identify and authenticate users based on their online activities only. For creating better user profiles, user behavior patterns must be identified from their online activities. User behavior is the new way of authenticating users which increases chances of getting better results for predicting future activities of users, for better recommendations and policy making decisions. The proposed project will focus on mining of hidden behavioral patterns from user activities. The project makes use of sentiment analysis to analyze sentiments of users regarding various posts using VADER sentiment analysis. The project further clusters users based on their interests in a particular subject for better recommendations using DBSCAN algorithm. Further the project represents the mined data in the form of histograms and pie charts for better visualization thereby making it easier for decision making. The project also provides recommendations to users by analyzing past data of user to provide personalization to user.**

***Keywords —Behavior analysis, clustering, data visualization, DBSCAN, sentiment analysis, VADER, recommendations***

## I. INTRODUCTION

With the rise of internet, the world has witnessed a huge rise in the amount of data. The major concern that the world is facing right now is how such huge amount of data can be managed. Furthermore, the businesses nowadays make use of past data for predicting future sales and for taking certain strategic decisions [3]. For this purpose, the data must be stored and must be available whenever needed. In other words, the data must be available in such a manner that it should be possible to extract meaningful information from the data. For this purpose, data mining is used. Data mining is the process of examining data from huge data stores such as databases or data warehouses with the intention of finding meaningful information from the data. This information can be used as a basis for making critical strategic business decisions [2].

The rise of internet has witnessed a huge increase in number of active internet users. Websites have become an integral part of their life and the users spend a large amount of time on the websites. In order to provide better suggestions and assisting users in finding the required information, these websites try to provide a more user specific experience to the users by analyzing the user behavior patterns on the internet. This is termed as user behavior analysis. User behavior analysis not only helps in providing a better user experience but also helps in understanding user's interests.

The proposed project makes use of sentiment analysis for analyzing user behavior patterns using different parameters. Sentiment analysis analyzes user's sentiments for a particular post or subject and classifies into either positive, negative or neutral [5]. Further, the project makes use of DBSCAN clustering algorithm to form clusters of users with similar interests [7]. These clusters help in analyzing the users' interests and thereby suggesting policies to a particular group of users. For providing easy analysis of data, the proposed project provides visualization of data in the form of histograms and pie charts.

## II. RELATED WORK

The traditional web mining technologies deal with extraction of three types of data: 1) Web log mining; 2) Web content mining; 3) Web structure mining. The traditional web mining approaches deal with analyzing data from large data stores to find patterns in the data. Furthermore, the web digging activity deals with identifying hidden data from web sites and web documents [2].

Prediction of user's navigation patterns can be predicted using Naïve Bayesian method for better recommendations. User's web usage can be mined for analyzing and predicting user's navigation patterns. Furthermore, it is possible to analyze changing user's interests by analyzing user's navigation patterns [3].

It is proved that a user's interest depends on his browsing behavior on the web. Web log mining is the process of extracting user's web log data and analyzing the data to find meaningful information about user's behavior patterns. Based on these patterns, web sites can be made more user-centric thus providing a better user experience to the user. This also helps in providing better web services that cater to

 DOI : 10.18231/2454-9150.2019.0348 © 2019, IJREAM All Rights Reserved.

the needs of potential customers and to improve quality of the services [4].

Analyzing user's responses on social media websites such as tweets on Twitter helps to understand user's response to certain events and situations. A machine learning approach is used to classify Twitter users by considering features such as tweeting behavior, linguistic content of tweets and social network information. The project makes use of Gradient Boosted Decision Trees (GBDT) as a learning algorithm for user classification [5].

Behavior of users change overtime which is reflected through their online behavior, especially social media. In such cases, continuous analysis of user sentiments is necessary to keep updated with recent trends. Past work deals with analyzing data and storing it in parts where each part is considered as a past dataset. A new dataset is analyzed and compared against the past dataset. If it is found similar, then it is ignored, else the past dataset is updated with the new dataset. This new dataset is further used for future sentiment analysis. Thus, the work deals with continuous supervised learning of user sentiments considering a large scale dataset [6].

Density based clustering methods prove useful in case of unlabelled dataset and Density Based Spatial Clustering of Applications with Noise or DBSCAN algorithm is one of the important density based clustering algorithms. DBSCAN forms clusters based on two parameters viz. minimum number of points and distance between two points. Based on these two parameters, DBSCAN forms dense or sparse clusters. Although DBSCAN is proved to be useful in many cases, the formation of meaningful clusters is not guaranteed [7].

Some past work has considered ant colony model for identifying user's browsing patterns by taking into account the frequency of visits of user and the total time spent by the user on a website as measures for identifying the interest level of user regarding a particular topic [1].

An attempt to increase the efficiency of DBSCAN algorithm was made in [8] where the factor that causes the computation to be slow is found to be the distance calculation between the clusters. This was eliminated by calculating mean of two clusters under consideration and the mean was compared against threshold times three value. If the mean was less than the thrice of threshold value then the two clusters are combined. This has resulted in increasing the efficiency of the density based clustering algorithm.

User interest is analyzed by capturing click streams of the user on certain social networking sites and also by considering the social network topology of the social networking site. Based on this data, there has been work done as in [9] that showcases how frequently users navigate through their friends' accounts based on which their interest levels are calculated.

A study was performed in [10] to analyze whether Vader sentiment analysis behaves similar to how a human would analyze sentiments in a sentence. The work was done using Twitter and the tweets were analyzed using Vader sentiment analysis and also by a group of 10 people. The results showed that there was not a significant difference between a human and Vader thus making Vader a good sentiment analysis tool.

A new approach to web log mining and identifying patterns in the web log data was proposed in [11] where the web log data is initially preprocessed by cleaning and eliminating all the robot or fake requests and then using a density based clustering algorithm to find navigation pattern of user. Further, the work was able to identify user sessions using a time-out based heuristic with a good false negatives to the total sessions ratio.

## III. PROPOSED WORK

The proposed project focuses mainly on two things: 1) Analyzing users' internet activity to predict his/her behavior; 2) Generating user profiles based on his/her interests.

**I. Analyzing users' internet activity:**

In order to analyze users' internet activity, we developed a news feed website for capturing user data such as number of visits, time spent on a particular article and whether the user likes the post or not [1]. The news feed portal also allows users to express their views on that particular topic. The comments are further used for analyzing the sentiments of users for that particular post. From the number of visits, time spent and the likes of users, the interest score of the user is calculated [9], [11]. We have developed a formula for calculating the interest score of user as follows:

**Interest_score = (NOV * 0.1) + (TS * 0.02) + LIKE * 3**

Where, NOV is the number of visits on a particular article of a user, TS is the time spent by the user on an article and LIKE is whether the user has liked the post or not. It is important to note that TS is the average time spent by the particular user on that particular post.

**TABLE1 USER BEHAVIOR EVENT DATA**

| Post ID | NOV | Time Spent | Like | Score |
|---|---|---|---|---|
| 1 | 4 | 181 | 0 | 4.02 |
| 2 | 3 | 93 | 1 | 5.16 |
| 3 | 2 | 22 | 0 | 0.64 |
| 4 | 4 | 284 | 0 | 0.4 |

The above table shows the event sets used for analyzing the user behavior. Like column in the table contains values 0 or 1 where 1 denotes liked and 0 denotes not liked.

 DOI : 10.18231/2454-9150.2019.0348 © 2019, IJREAM All Rights Reserved.

The user behavior event data helps in calculating the interest score of user which analyzes whether the user is interested in a particular post. However, this interest can be positive or negative. User gives his opinions through comments on the articles. These comments are used for analyzing the sentiments of users for a post in the news feed [6]. We have tested the comments on two different algorithms for sentiments analysis. The first model makes use of logistic regression for analyzing the sentiments. This model was trained using IMDB movie reviews dataset [15]. However, it was not efficient in analyzing complex sentences and negated sentences since the algorithm calculates sentiments for each word separately by tokenizing the entire sentence. The second algorithm that we used in this project is Valence Aware Dictionary and Sentiment Reasoner or VADER which is lexicon and rule-based tool used for sentiment analysis [14]. The VADER makes use of a dictionary of known words along with their polarity scores. The sentiment analysis is done by calculating sentiments of each word in the text and checking it against the polarity of words in the dictionary. The VADER returns a compound score which takes an aggregate of the overall positive, negative and neutral scores in a sentence. This compound score is in the range of -1 to +1 [10].

**II. Generating user profiles based on user interest:**

For generating user profiles, we have used Density Based Spatial Clustering of Applications with Noise or DBSCAN clustering algorithm [7], [8]. We make use of the compound score obtained in I for clustering together similar or closely spaced users based on his/her interests. For forming clusters, we make use of Euclidian distance to calculate distance between two points that are a combination of user interaction score and user sentiment score. Users with similar interest groups are grouped into one cluster. Users are grouped based on keywords for each post that the user is interested in.

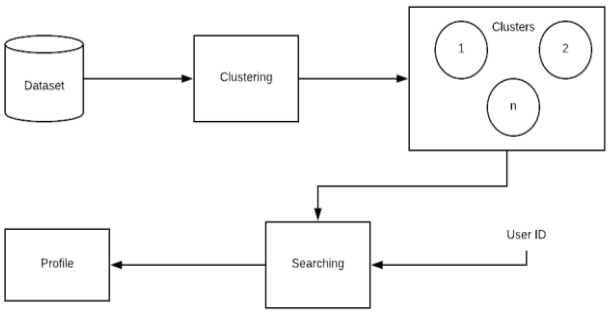


**Fig 1: Architecture of user profile creation**

As described in the above architecture, it is possible to search in the clusters using user ID. Performing this search provides the user with a visualization of data in the form of graphs and pie chart that can be used to better understand the information present inside the data.

## IV. IMPLEMENTATION

Implementation of the project takes the following steps:

**I. User:**

1. Users can log in to the news feed portal to view the news. If they do not have an account, they can signup using the link provided.
2. Once logged in, the users can interact with the news articles by liking the post and commenting on the same.
3. The user can end his session by clicking on the logout option.

Whenever a new user signs up on the portal, an online profile of the users is created automatically at the backend. This profile consists of the posts the user liked, his interest in posts, etc. Since each post has a specific keyword assigned to it, it helps to uniquely identify what the post is about using that keyword. If a user is inclined towards posts of a particular keyword, e.g. football, then the system updates the profile of the user and adds football as an interest and calculated the user's interest levels using the method explained in the Implementation part of this paper. In such manner, the profile of the user is automatically and continuously updated as the interaction of the user with the portal increases. As the user interacts with the portal, the system automatically recommends posts that the user may be interested in. This is done by ranking the posts that are most similar to the posts which the user has liked earlier or has interest in.

**II. Admin:**

1. Admin is responsible for uploading news articles for the user to interact.
2. Once the users have interacted with the news feed, the behavior event set is used to generate profiles of users.
3. Sentiments for all the comments are calculated and stored for later use.
4. Admin is able to visualize data such as clustering of similar interest groups of users and sentiments of users on posts.
5. Admin is able to view the histogram of user interest scores plotted against the keywords of posts he is interested in.

## V. RESULTS

The results of our system showcase the interest levels of users using histograms denoting the number of users interested in a particular subject. The following figure shows a histogram of users where X-axis consists of keywords or the topics in which the user is interested and Y-axis consists of Number of users interested in a particular topic [13].

 DOI : 10.18231/2454-9150.2019.0348 © 2019, IJREAM All Rights Reserved.

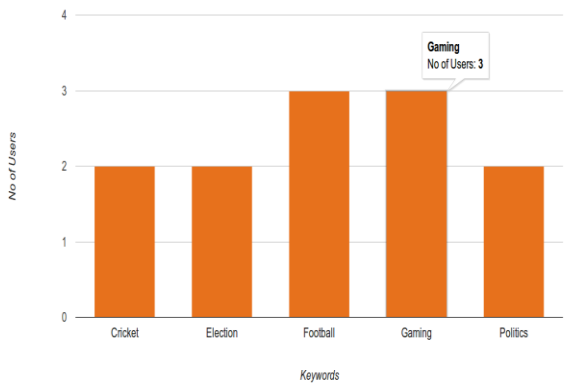


**Fig 2: Histogram of users interested in a particular topic**

The sentiment analysis compound score returned by VADER sentiment analysis algorithm is represented using pie chart where the sentiment is positive if score is greater than 0.05, negative if the score is less than -0.05 and neutral if the score lies between -0.05 to 0.05. Furthermore, the sentiments are classified into weak positive, moderately positive and strong positive if the scores are greater than 0.05, 0.33 and 0.66 respectively. The same is done in case of negative sentiments except that the score value range is negative. The following figure visualizes the sentiments of user for a particular topic based on his comments [12]. This is represented using a pie chart as shown below:

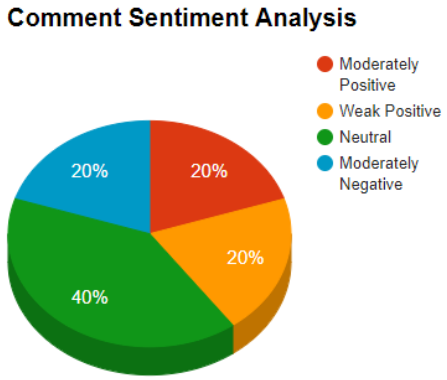


**Fig 3: Pie chart of sentiments of users for a particular topic**

We trained a clustering model using DBSCAN algorithm for generating clusters based on interest levels and comment sentiment polarity (CSP) score using a dataset that we obtained by allowing people to interact with our portal. A new user is thus classified into particular cluster based on his interest and comment sentiment polarity score. The minimum distance Ɛ is set to 0.3 which is an empirical value.


**Fig 4: Generated clusters based on interest level and CSP score**

The green cluster represents interested users, red cluster represents not interested users and yellow cluster represents fake users or bots.

The proposed system further provides recommendations of the posts based on user's interaction with the portal, capturing his/her likes and interests, based on which, posts that are most related to the user's interest are displayed first. Thus, the system acts as a recommendation engine.

## VI. CONCLUSION

The proposed project is able to successfully identify and analyze user's interest levels regarding a particular subject. The project also analyzes users' responses to various posts and predicts the sentiments of users with a very good accuracy. Further, the project helps the user by providing better visualization of information thereby enhancing the user experience. The system also recommends articles by analyzing the user's interests in a particular subject and displays the articles that the user is most likely to get interested in.

Given any amount of data under any scenario, our system is able to generate automatic profiles of users based on their interests and likes and help the organization to make better policies or plans for group of people belonging to a particular interest group. Further, the system is also able to successfully analyze the sentiments of the people belonging to a particular interest group, thereby making it easier for the authorities to make changes accordingly.

## REFERENCES

[1] Xipei Luo, Jing Wang, Qiwei Shen, Jingyu Wang, Qi Qi, "User Behavior Analysis Based on User Interest by Web Log Mining," 2017 27th International Telecommunication Networks and Applications Conference (ITNAC).

[2] Yin B, Zhang Z, Wang X, et al. Research and Application of Data Mining Technology Used in the Analysis of Smart Home User Behavior[C]// Sixth

 DOI : 10.18231/2454-9150.2019.0348 © 2019, IJREAM All Rights Reserved.

International Conference on Measuring Technology and Mechatronics Automation. IEEE, 2014:476-479.

[3] Mahdi Khosravi, Mohammad J. Tarokh, "Dynamic Mining of Users Interest Navigation Patterns using Naïve Bayesian Method," Proceedings of the 2010 IEEE 6th Internation Conference on Intelligent Computer Communication and Processing.

[4] Haifeng Ling, Yezheng Liu, Shanlin Yang, "An Ant Colony Model for Dynamic Mining of Users Interest Navigation Patterns", 2007 IEEE International Conference on Control and Automation, Guangzhou, CHINA, 2007

[5] Marco Pennacchiotti, Ana-Maria Popescu, "A Machine Learning Approach to Twitter User Classification", Proceedings of the Fifth International AAAI Conference on Weblogs and Social-Media.

[6] Rui Xia, Jie Jiang, Huihui He, "Distantly Supervised Lifelong Learning for Large Scale Social Media Sentiment Analysis", IEEE Transactions on Affective Computing, Vol.8, No.4, December (2017).

[7] Amey K. Redkar, Prof. S. R. Todmal, "A Survey on DBSCAN Algorithm to Detect Cluster with Varied Density", International Journal of Advanced Research in Computer Engineering and Technology, Vol. 5, Issue 7, July 2016

[8] Satyasai Jagannath Nanda, Ganpati Panda, "Design of computationally efficient density-based clustering algorithms", Elsevier Data and Knowledge Engineering Journal 95 (2015) 23-38

[9] Fabricio Benevenuto, Tiago Rodrgiues, Meeyoung Cha, Virigilio Almeida, "Characterizing user navigation and interactions in online social networks", Elsevier Information Sciences Journal 195 (2012) 1-24

[10] Ajla Kirlic, Zeynep Orhan, "Measuring human Vader performance on sentiment analysis", Invention Journal of Research Technology in Engineering and Management, Vol. 1, Issue 12-Version-4, December 2017

[11] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley, Reda Alhajj, "Effective web log mining and online navigational pattern prediction", Elsevier Knowledge-Based Systems Journal 49 (2013) 50-62

[12] Google API for generating pie chart, https://google-developers.appspot.com/chart/interactive/docs/gallery/piechart

[13] Google API for generating histogram, https://google-developers.appspot.com/chart/interactive/docs/gallery/columnchart

[14] Vader sentiment analysis use documentation and usage, https://www.nltk.org/_modules/nltk/sentiment/vader.html

[15] IMDB movie reviews dataset for sentiment analysis, http://ai.stanford.edu/~amaas/data/sentiment/

 DOI : 10.18231/2454-9150.2019.0348 © 2019, IJREAM All Rights Reserved.