Analysis of Online User Behavior Detection Methodologies and its Evaluation

Dhanashree Deshpande
PhD Student, SGBAU,
Amravati, India

Shrinivas Deshpande, PhD
Head, P.G. Dept. of Comp. Science & Tech.
DCPE, Amravati,
India

ABSTRACT

With the increasing use of internet, users are accessing information and services easily through various media like social communication, multimedia content, online shopping and banking services etc. It becomes challenging task to accurately identify and differentiate normal and suspicious user behavior. Various businesses need information of next user behavior prediction to enhance their service quality. This paper gives the analysis of online user behavior detection and prediction. Various user behaviors identification methods are compared and analyzed. Their parameters are considered and improvements are suggested. The proposed methodology describes anomalous user behavior detection system. The principal component analysis is the feature extraction method used to detect and differentiate normal and anomalous user behavior.

Keywords

online user behavior; fuzzy theory; GSP algorithm; SGD algorithm; PCA

1. INTRODUCTION

Due to regular improvement in technology, online user creates various directions for behavior. Every time user behavior gives new view for study and this study facilitates good service to online users. The user behavior is constructive as well as destructive. It is constructive when used as web site design improvement, web navigation improvement for better resource utilization, Product and service enhancement in Electronic Commerce, improvement of students' facility in e-learning environment. Personality of a person can be also predicted accurately through sentiment analysis. Users' liking behavior to various posts, comments of social media predicts personality of online user. In destructive type of behavior, a large number of cyber-attacks have been focused on online banking system, online social networks introduce new challenges related to security and privacy, malicious behaviors such as spam and Sybil attacks take place in OSNs and bring severe security threats. Fake accounts are created solely for the purpose of spamming. Behavior of the user may get vary according to the situations. Computers learn or strengthen the memory of the user's interests by analyzing his behaviors. The analysis of users' behavior gives direction to improvements of search engines. In the process of information retrieval and knowledge mining, it is crucial to understand users' intentions. Various methods have been used for online user behavior identification on different domains. Web log file stores web site navigation information of user and this log file is studied to analyze the user behavior using data mining techniques. This paper analyzes online user behavior detection methodologies and compares the result.

This paper is arranged as follows. Section II reviews related work on online behavior detection process while Section III describes various methodologies used in user behavior detection. Section IV gives comparative analysis of various online behavior identification methods. Section V describes proposed methodology and parametric explanation. Section VI gives suggestion for improvement in existing user behavior detection process and parameters.

2. RELATED WORK

The questionnaire was published on the military museum's web site. Purpose of visit and preferred data elements get cleared in these questions. The characteristics of online museum visitor's web search behavior have been explored in the context of digital museum resources. The characteristics of visitors' searching behavior like visual experience, exploratory behavior, broad known item search and meaning making (visitor's searching behavior) were identified. Visitors are interested in the information of specific museum object and its photograph. The method of web questionnaire survey is used to describe why online museum collections are used and by whom [1].

Students' interaction is classified in the education domain. Large heterogeneity observed in the students' behavior. Web site's educational resources are improved to create reasons for students to stay longer and increase their knowledge. The proposed method predicts session end and next user action. Personalized polynomial classifier was proposed to classify dynamically changing data having various obstructions. This classifier was proposed by using attributes' weights which was calculated for each user individually. These weights were dynamically calculated using stochastic gradient descent approach [2].

Improved GSP (generalized sequential pattern) algorithm was used to analyze the sequence of user behavior by comparing accuracy rate of user behavior prediction with the classical GSP algorithm. User's next behavior is also get predicted by matching all the sequences patterns in the file. 6 kinds of behaviors were extracted which includes user login, successful response, fast response, termination response, interruption response and exchange prize. AprioriAll algorithm, Spade algorithm, and GSP algorithm are commonly used to analyze user behavior sequence [3].

The popular social media platform Twitter and online commenting platform Disqus were used to analyze people's usage pattern and personality traits. Person's sentiments were identified clearly on Disqus online platform. But tweets on twitter provide ambiguous or neutral sentiments [4].

There is a connection between user's personality and facebook user's like behavior. The people are considered as extrovert when liking of post activity is high level [5].

Fuzzy logic was used to detect user's abnormal behavior in online banking. For example, If the account is inactive since many days and suddenly used for payment transaction then it is considered as suspicious user. If user is doing one mistake while login and has done 4 very small payment transactions after that. Then the user is considered as normal user [6].

Search log analysis method is used to capture interaction of a large number of users with a search engine. Query log metrics and novel metrics were used to understand the search problems of children. Click positions and click duration were analyzed to explore young users' behavior. The method like analyzing session characteristics was used to find out the search behavior and difficulties of particular demographic group like young users. The characteristics of the query i.e. how to queries can influence the topics that correlate best with young users. The web access characteristics differences were identified among child, young and old users [7].

Consumer behavior is measured in online channel with the help of consideration set, price competition, online search and online purchasing. The price comparison engine is used for the explicit purpose of consumer research. Consumer behavior vary between the markets of small value items where risk level is low or zero and retail banking where risk level is high [8].

3. METHODOLOGY

3.1 Web questionnaire survey

This is the data collection method to find out the answer of few combinations of questions. Online museum visitors' search and interaction behavior with the museum collection database is studied. 24 participants who are men and aged between 32 and 72 were asked about topical information need, information about data element search, historical information about items, their hobby etc. This study takes the data from 132 respondents. Online questionnaires provide information about visitors' area of interests, purpose of visit, and preferred data elements about online museum. Questionnaires are useful to describe why online museum collections are used and by whom. But they have low response rates and low explanatory power as to how users interact with rich museum content such as collection-related information. In future, collection of large cultural heritage will provide access to millions of heritage objects [1].

3.2 Simulated search task

3 simulated search tasks represented well-defined information need. Free text searching was the main search strategy. Participant's search process has been recorded by asking them to retrieve many useful documents to satisfy their information need. The recordings include mouse movements by using Morae software tool, retrospective talk-aloud sessions where participants were asked to comment and explain their search sessions in order to obtain detail search information. The recorded video clips were examined for extracting information on few search attributes. A one-way ANOVA test is carried out in order to analyze differences among above 4 search tasks. The LSD (multiple comparisons) test was used to examine differences in patterns [1].

3.2.1 Data Analysis

Quantitative data like questionnaire, participants' recruitment is explored and support the qualitative data analysis. ATLAS.ti is the qualitative data analysis software used to analyze the comments of participants' during retrospective talk-aloud session together with observation notes from search sessions.

3.2.2 Inductive Content Analysis

4 main characteristics of online museum visitor's searching behavior were identified. Those are highly visual experience, exploratory searching behavior, element searches and meaning making. This is the way to analyze the data collected from qualitative research methods. The text material of questionnaires was coded based on and inductive content analysis. ICA involves allowing themes to emerge from data. For e.g. If many participants were interviewed, then researchers will find raw themes from all transcripts before comparing them. Based on these themes and quotes in the interviews which support them, the researcher then writes an interpretation of the data. Reflexivity may get applied at the end and at last interpretation is given to the interviewee in order to establish credibility.

3.3 Binary classifier

Students' behavior in web based education system is irregular and unexpected. To deal with this limitation, polynomial binary classifier is proposed using the stochastic gradient descent algorithm. This algorithm is able to process the real time data of students' actions and dynamically make predictions for actual sessions. With stochastic approach, all observations are considered only once so the big amount of data can be processed. Few of the attributes are considered for the prediction if the student will leave the website or not. Those attributes are the students, learning objects, students' visits of learning objects, individual sessions and timestamp. Few attributes are calculated like student behavior in the session, her typical behavior, LO characteristics. In binary classification, the accessed observations can be classified into one of two considered classes. The problem of imbalance in multiple classes may occur. To reduce this disproportion, the modified SGD approach is used which results in equal classification precision for both considered classes. The SGD classifier supports classification of multi-class by combining multiple binary classifiers in a "one versus all" scheme. The proposed approach can be used to improve students' experience to interact with system, students' performance is increased from the knowledge point of view. In future, it may help to discover typical behavior patterns and to explore latent dependencies in students' behavior by considering historical actions of students' sessions [2].

3.3.1 TLOF (Time-Adaptive Local Outliner Factor)

This is the unsupervised learning model and anomaly detection technique in online communication system. This model is constructed by taking into considerations few factors like online communications in dynamic environment requires anomalies should be detected quickly with as little training data as possible. Here, changes in user behaviors are sudden and TLOF is adapted to identify anomalies in these sudden changes. This model has few advantages which is suitable for application of anomaly detection. This model requires no training data and in application also no anomalous users are known in advance, it takes the time sequence of user behavior into account instead of just one snapshot of the behavior, it assigns anomaly scores instead of just assigning users as normal or anomaly, it detects outliers based on Euclidean distance. It makes the result easily interpretable by visualizations. TLOF gives an anomaly measurement for every time series (every user).

3.3.2 Detecting anomalies in static unattributed networks

The various approaches are categorized into 3 groups: clustering/community-based, network structure-based, and signal processing-based approaches. An algorithm has been proposed to find the community or neighborhood of each node in the bipartite graph using random walks with restart and graph partitioning. This algorithm is used to detect anomalous nodes in the network. SCAN and GskeletonClu are density based network clustering algorithms used to identify clusters, hubs, and outliers in large networks. OddBall is a network structure-based technique used to discover anomalies such as near-clique, near star, heavy vicinity, and dominant edge patterns from large, weighted networks. A combination of Gaussian Mixture Model and fuzzy logic as a novel method is used to differentiate between normal and anomalous individuals.

3.3.3 Detecting anomalies in static attributed networks

The community-based anomaly detection methods proposed to integrate attribute graph clustering and outlier detection in a single algorithm. GBAD algorithm is introduced for discovering anomalies in network.

3.3.4 Detecting anomalies in dynamic unattributed networks

The various approaches can be grouped into 3 categories: matrix/tensor decomposition-based, community-based, and probability-based approaches. To detect anomalies CMD is used which is the low-rank approximations of input networks are used to summarize the dynamic networks. The signal processing-based approach uses matrix decomposition to find anomalous nodes in dynamic unattributed networks. In order to detect anomalous time windows, a linear ramp filter is applied on the residual matrices and then partial eigen vectors is analyzed. Tensor analysis is very good and powerful tool for detecting anomalies from dynamic and multi-aspect network. NetProbe approach is used to find anomalous nodes. This approach is used to detect fraudsters in online auction networks. Link prediction technique is applied to discover anomalous edges in a dynamic network. Future interactions are also predicted through link prediction.

3.4 Generlized Sequential pattern (GSP)

GSP algorithm is proposed to mine frequent sequential patterns in user behavior. Traditional and improved GSP algorithms are compared for user behavior prediction. The set of parameters used are size of test set, max sequence pattern length, correct prediction number, unable to handle number, prediction accuracy rate. Prediction accuracy rate for improved GSP algorithm is significantly higher than classical GSP algorithm. It is effective to analyze the sequence of user behavior through improved GSP algorithm [3].

3.5 Sentiment Analysis

It is the process of extracting information on a person's behavior, interest and opinion towards different topics. It determines whether a piece of writing is positive, negative, or neutral. A common use of this technology is to discover how people feel about a particular topic. From social media platforms the linguistic features and personality traits have been extracted. Linguistic based analysis is performed by using LIWClite7 tool and social media analysis is performed on the dataset by using Texalytic. Openness, neuroticism, and conscientiousness personality traits of a person are strong in his Disqus comments and extraversion is strong in tweets. Disqus offers comprehensive result than twitter Online profiles of an individual can be merged and this reveals to build a comprehensive virtual profile. This profile can help in research like response prediction, news feed generation, group targeted advertisements. Dataset consisted of random people of different age and geography to it was difficult to reach them. In future, it will be possible to predict how a person would react to a topic in general. Personal behavior of an individual can be identified from different social platforms [4].

3.6 Fuzzy Expert System

Fuzzy expert system was proposed to identify the user behavior in internet banking. This system is optimistic to be used for improving e-banking security and quality. The fuzzy rule base was developed by using input variables and expert views with 120 "if-then" rules, the fuzzy rule base was developed. The fuzzy expert system was implemented by using the information gained from the real environment of the system. By examining receiver operating characteristic curve results, it is understood that a system that combines several important factors as input and examines these factors simultaneously can determine if certain banking transactions are dangerous. In future, performance of the proposed algorithm can be enhanced by applying fuzzy methods, neural network approach or genetic algorithm [6].

3.7 Consideration Set

Consideration set is used for analyzing and understanding consumer behavior. It is one of the measures of consumer behavior in the online channel. The consideration set is the group of suppliers that a buyer actively considers in their decision making before purchasing the product. Price comparison engines is used as an important marketing topic and used for the purpose of consumer research. Retail panel data is used to examine consumer purchasing behavior of a household product. With the help of online panel data, it is possible to analyze and examine the actual behavior of online consumer instead of reports of historical actions. The online consideration set is calculated from online users who visit 2 or more websites because these are the users who are actively researching 2 or more suppliers. The variables like online consideration set and online price competition intensity, online research, online purchasing are the new theoretical constructs used to measure online behavior, In future, online measurement constructs can be develop and more sophisticated models can be created. [8].

4. ANALYSIS & DISCUSSIONS

This questionnaire information in the context of digital museum resources reveals that users search for previously decided object. Free text strategy is applied here. Questionnaires were asked which were inspired by real life information needs. Participants' searching behavior is combined with their explanations and comments in the process of inductive content analysis. Students' behavior is irregular and uncertain at various situations. SGD approach handles bulky and discrete data of online education system in the process of retaining the students. In the layout method, users were efficiently compared based on the glyphs and also similar users were identified. PCA algorithm describes user behavior more completely in the proposed model. Sequence of user behavior is analyzed by improved GSP algorithm and behavior prediction accuracy was improved. After applying classifier on 2 datasets, REP tree is the best classification algorithm for the classification of malicious and legitimate users. Abnormal behavior in online banking was detected by using fuzzy theory.

Table 1. Comparative analysis of various online user behavior identification methods

Method	Database used	Implementation	Result
questionnaire, Inductive Content Analysis [1]	1,705 military museum artefacts	It provides purpose of visit, area of interests, , preferred data elements and demographic data.	4 main characteristics of visitor's searching behavior were identified More time is spent on open topical search tasks.
SGD [2]	Log information of ALEF education system.	SGD classifier brings faster data processing and effective reaction to dynamically changing data.	Stochastic classifier trained individually per user needs high amount of observations to be trained optimally.
GSP [3]	Business db of survey company contains total of 38,658 complete behavior sequences.	User's frequent sequence pattern is obtained. The prediction to the next behavior for this active user is made by matching all the sequences patterns in the file.	Sequence of user behavior is analyzed through improved GSP algorithm, by comparing the accuracy rate of user behavior prediction with the existing classical GSP algorithm.
Sentiment Analysis [4]	173 users were identified who had linked their public twitter profile to Disqus. 3,200 recent tweets per user were collected. In Discus on average a user discussed on 424 distinct articles through 1170 comments.	LIWClite7 tool is used to perform a linguistic based analysis on Disqus and Twitter contents on around 23720 words. Social media analysis is performed on the dataset using Texalytic.	Disqus offers comprehensive result than twitter. Disqus comments provide better result on a person's sentiment. A person's twitter posts on the same topic show neutral sentiment.
Fuzzy expert system [6]	30 information items related to users exist in db of online banking.	Method was implemented using information gained from the real environment of the system.	The performance of method was evaluated using an ROC curve. ROC provides accuracy of 94%.

5. PROPOSED METHODOLOGY

Analysis of user behavior by considering different methods, algorithms are described. Suspicious user behavior and normal user behavior is identified, user's next behavior is predicted. User's personality traits are identified. In this behavior analysis, different parameters are improved by considering various domains.

The purpose is to propose a method which extracts user's behavior. PCA (principal component analysis) is the method can be used in anomaly detection model. It describes user's behavior more completely and improves the efficiency of the algorithm. It reveals the internal structure of the data and explains in detail the variance in the data. In PCA algorithm, covariance matrix of dataset is obtained first. Then feature values and eigenvectors of the covariance matrix are calculated. Eigenvectors with the most obvious features are selected.

In data acquisition and preprocessing, web log file is processed or database records are analyzed. This data is put into data set vector. If the user behavior feature values are in the normal range then this behavior data will be added to the training data. Otherwise the user behavior will be considered as abnormal.

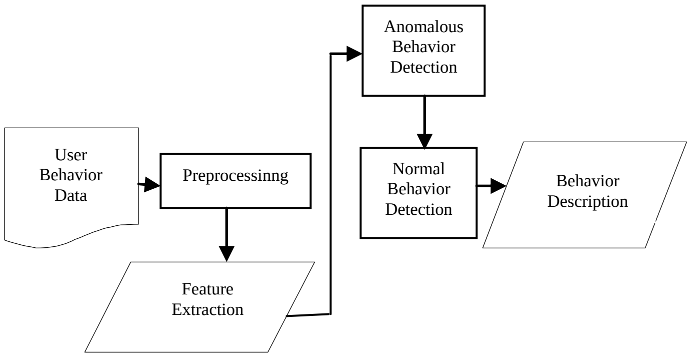

Fig. 1: Block diagram of Anomalous Behavior Detection System

Table 2. Analysis of algorithms by considering parameters

Method	Parametric Considerations	Improvement
Questionnaire, Inductive Content Analysis [1]	Average (search time, items viewed, records viewed, digital photographs viewed, use of zoom function, digital images, search iterations)	Low precision was accepted in order to achieve high recall. This finding supports explorative search behavior.
SGD [2]	Accuracy, precision	Improvement in

		prediction accuracy
GSP [3]	Prediction accuracy rate, maximum sequential pattern length, number of unable sequences	Prediction accuracy rate in improved GSP algorithm is higher.
Sentiment Analysis [4]	Neuroticism, Extraversion, Openness, Agreeableness, conscientiousness	Disqus reveals more neuroticism, openness. Extraversion is more in twitter, Agreeableness is consistent, tweets reveal less information than Disqus comments on conscientiousness
Fuzzy expert system [6]	Error & transfer count, transferred amount, dormant amount, user records, transfer interval, input time, output	The time interval between transactions and prices determines dangerous behaviors.

6. SUGGESTIONS FOR IMPROVEMENTS

Visitors on digital museum library search mostly for specific object. While searching it explores more objects of similar features and visitors can enhance their knowledge. Search time can get increase when visitors will get more knowledge about object through similar type of user by finding their comments for the same object. This can improve the service of digital library. They can discuss and explore their knowledge. Similar type of user searching behavior can recommend required objects to visitors.

Students' behavior prediction includes attrition rate prediction, short term behavior prediction and session end prediction. To retain the student on learning website, repetitive link visit can be found out from log. There are two meanings. Student did not understand the topic or student is more interested for research. Here multimedia content can be recommended. Still student ignores then questionnaire may get provide. This analysis may helpful for retaining students on learning website.

Fuzzy theory is applied for the detection of suspicious user on internet banking. Certain suspicious users' behaviors are taken into account. To improve the accuracy in the suspicious behavior prediction process, ip address of user and browser which user is normally using can be used in the combination of other behavior of same user.

7. CONCLUSION

The analytical study discusses various methods and algorithms used in the process of user behavior detection. Various domains are considered while detecting normal and abnormal user behavior. Items of online digital museum library are considered, threats in online banking are analyzed. Various social media platform uses are now on big demand. Analyzing users' comments, likes, tweets, sharing indicates their personality traits. Prediction of user behavior pattern is studied. User behavior prediction is the upcoming internet star for various business services, marketing campaigns, and to analyze people opinion for political leader. Principal component analysis method of feature extraction is described and proposed for normal and anomaly detection.

8. FUTURE SCOPE

In future, PCA will be applied to extract features and to detect anomalies on domain like shopping cart user behavior.

9. REFERENCES

[1] Mette Skov, Peter Ingwersen, Denmark, "Museum web search behavior of special interest visitors", Library & Information Science Research, Elsevier, vol. 36, pp. 91-98, 20 May 2014.

[2] Ondrej Kassak, Michal Kompan, Maria Bielikova, "Student behavior in a web-based educational system: Exit intent prediction", Engineering Applications of Artificial Intelligence, Elsevier, vol. 51, pp. 136-149, 2 Feb 2016.

[3] Xiaowei Zhu, Shaochun Wu, Guobing Zou, Shanghai, "User Behavior Detection for Online Survey via Sequential Pattern Mining", 5th International Conference on Instrumentation and Measurement, Computer, Communication and Control, IEEE, 978-1-4673-7723-2/15, pp no. 493-497, 2015.

[4] Hasan Al Maruf, Nagib Meshkat, Mohammed Eunus Ali, Jalal Mahmud , "Human behavior in different social medias : A case study of Twitter and Disqus", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ISBN 978-1-4503-3854-7/15/08, pp. 270-273, 2015.

[5] Seyed Morteza Ghavami, Masoud Asadpour, Javad Hatami, Mohammad Mahdavi, Iran, "Facebook User's Like Behavior Can Reveal Personality", 7th International Conference on Information and Knowledge Technology, IEEE, 978-1-4673-7485-9/15, IKT2015.

[6] Saeideh Alimolaei, Iran, "An Intelligent system for User Behavior detection in Internet Banking", 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), IEEE, 978-1-4673-8545-9/15, 2015.

[7] Sergio Duarte Torres, Ingmar Weber, Djoerd Hiemstra, University of Twente, "Analysis of Search and Browsing Behavior of Young Users on the Web", ACM Transactions on the Web, 1559-1131/2014, vol. 8, no. 2, article 7, pp. 7:1 to 7:54, March 2014.

[8] Christopher P. Holland, Gordon D. Mandry, "Online Search and Buying Behaviour in Consumer Markets", 46th Hawaii International Conference on System Sciences, IEEE, 1530-1605/12, pp. 2918-2927, 2013.