Analysis of various Characteristics of Online User Behavior Models

Dhanashree Deshpande
PhD Student,
SGBAU, Amravati,
India

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

ABSTRACT

Accurately identifying online user behavior is challenging task because while identifying malicious users, legitimate user should be separated correctly. Normal and suspicious users should be differentiated. Various classification methods are useful in this behavior detection process. Some of them give good performance and accurate results. Few metrics are used to deviate malicious users from good one. Security is the main concern need to provide to various online applications. Characteristics of user behavior can be studied on variety of OSNs, online news website, shopping web site for prize comparison, browsing behavior, search engine behavior through queries, users' communication behavior through various online messaging platforms etc. This paper gives analysis of various characteristics of online user behavior models. User behavior methods are compared and analyzed.

Keywords

k-means, Markov Logic Network, online user behavior, social network, user behavior model

1. INTRODUCTION

The characteristics of online user behavior analysis created wide scope for research and it is found as new dynamic area of research. User's behavior study and its prediction create target area to focus for many services and business industries. In this paper, various characteristics of user behavior are analyzed. This analysis takes place on various models described. The methods and metrics are used to classify users and to measure deviation from another user. Behavior Predictions can be generated through session traces and these sessions' study gives historical behavior of user. Sales prediction is uncertain task as it depends upon market situation and product demand. An optimal sales prediction model was constructed by using useful data mining methods. These methods are used to forecast sales trends. This sales trend analysis is beneficial for Ecommerce companies. The online user behavior such as sales behavior and browse behavior are considered in some frameworks. Users' behavior is classified and their profiles are created by using their features. Different feature extraction techniques are also used to identify user behavior and anomalies in the data. Some classification algorithms perform best for the classification of malicious and legitimate users. The behavior on various OSNs are studied and compared. The anomaly in social network is detected. One product differs by various prizes on different web sites. User compares prizes to purchase the product in minimal rate. Here, behavior of user is studied through the search query placed by them. User's search intention differs and to identify their characteristics search log is studied from search engine. This paper analyzes various characteristics of online user behavior 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 characteristics of 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 coefficient of reliability was introduced to estimate the value of deviation in user behavior. The IDSs classify acceptable behavior and report other irregular behavior as intrusive. Users' behavior profiles were created by considering temporal features such as events' lengths, and possible relations between them. While building user profiles, different characteristics were considered such as consumed resources, command count, typing rate, command sequences etc. The deviated values are considered as a consequence of abnormal behavior [1].

An efficient layout method was proposed which gives comparison and similarities in behaviors of different users. TargetVue visualizes the behaviors of suspicious users in behavior rich context by using unsupervised learning model. TargetVue consists of few modules like data collection module, preprocessing module, analysis module, visualization module. TargetVue introduces 3 new ego-centric glyphs to visually summarize a user's behaviors. These glyphs effectively represent the user's communication activities, features and social interactions [2].

Graph based anomaly detection was used for detecting anomaly in social networks. It is the process of extracting useful knowledge from graph data by using the techniques such as machine learning, data mining, statistics, pattern recognition, and graph theory [3].

PCA (Principal Component Analysis) method is the feature extraction technique used to analyze user behavior. It identifies anomalies in the data and detects user abnormal behavior [4].

The time characteristics of users' behaviors were investigated for 4 popular web forums of China. They were analyzed for user's post and reply behavior. Some interesting features of human behaviors were revealed and were helpful to the behavior-based detection of spam and spammers. The people's online behaviors have similar time characteristics with that of real life [5].

It is found that REP Tree is the best classification algorithm for the classification of malicious and legitimate users on the 2 datasets from Facebook-links and Live Journals [6].

Data mining framework was presented for the sales prediction by using online searching and browsing user behavior. The optimal data mining model is selected as the predictor by using the cross-validation method. This selected predictor with proper parameter and best feature subset is used to forecast sales trend [7].

The behavior of users who belong to both facebook and twitter were analyzed and compared. The purpose of this analysis is to investigate the behavior of 3 different types of users: declared – declared account on both OSN, hidden – do not declare account on both OSN and other - remaining users. It is found that users who have more accounts are less active than users who have only account in Twitter. The users who have more accounts but not declared facebook account are very inactive. The people who are active on social networks are more likely to feel connected. The social activity may improve social presence and increase social influence [8].

Behavior of users is studied on prize comparison shopping web site. It has included 2 aspects, data analysis and behavior prediction. Users are classified by their search queries. So the data includes the distribution of users coming to the website based on geographical location, time, week, clicked session, repeat users, phone brands visited and compared. There is strong correlation between change of product and popularity (based on number of visits). A model was created using Markov logic which uses the history of the user's activity (training data) in a session to predict whether a user is going to click to convert in that session [9].

Users' behavior was analyzed to study diverse characteristics under different search intentions. Log was analyzed from a real commercial search engine. The users' searching habits were found out such as the length of query, ratio of refining, ratio of clicks. This user behavior analysis gives direction to improvements of search engines. 2 kinds of special search requests were extracted and analysis was performed based on 3 sets i.e. 1) all 2) navigation searching set 3) sex-related set [10].

3. EXISTING METHODOLOGY

3.1 Coefficient of reliability in classification

Coefficient of reliability is calculated at each step of classification. During classification a coefficient of reliability is changed. Based on this, normal or anomalous user behavior is identified. While classifying the user behavior the system monitors deviations between expected user behavior and current one. Coefficient of reliability is used to estimate the value of deviation in user behavior. If this value crosses a certain threshold then it is considered as a case of abnormal behavior. That means the parameters in actions of user are not in admissible intervals. Coefficient of reliability is measuring the same individual twice and it correlates the 2 sets of measures. Every user action class was characterized by statistic parameters of time distribution – mean and standard deviation. Deviations from current values of sequential and temporal parameters are considered as consequence of abnormal behavior. The tools which are used to classify user behavior are N action classes and a relational matrix. These tools describe the model of user behavior. The approach is simple and easy to visualize. It is fast as it does not require many calculations [1].

3.2 Layout Method

TargetVue is a novel visual analysis system which detects anomalous users via an unsupervised learning model. It visualizes the behavior of suspicious users in behavior-rich context through novel visualization designs and multiple coordinated contextual views. 3 ego-centric glyphs visually summarize a user's behaviors. These glyphs effectively present the user's communication activities. The layout method is proposed to capture similarities among users and facilitates comparisons of behaviors of different users. High level features are used for detecting anomalies in online communication systems. Those features are behavior features, content features, interaction features, temporal features, network features, user profile features. In future, more advanced methods can be designed and integrate into the system based on active learning techniques [2].

3.2.1 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 Social Network Anomaly Detection

In the social network anomaly detection, the input networks which are used by various methods are categorized as static/dynamic or attributed/unattributed. Various state of the art methods for mining anomalies from static social networks have been discussed. This node is used for spotting anomalous nodes in static unattributed network [3].

3.3.1 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.

3.3.2 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.

3.3.3 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 Principal Component Analysis (PCA)

This method is used to analyze user's behavior. It is one of the most popular methods of feature extraction. PCA is used widely in pattern recognition, digital image processing, signal processing and other fields. PCA can identify anomalies in data and it detects user abnormal behavior through the change of principal direction. Covariance matrix is obtained and feature values and Eigen vectors of the matrix is calculated. By considering flaws, PCA requires a lot of storage space for computational complexity if original dimension of the space is n. Decomposition of a non-sparse matrix takes place in PCA. Due to this calculation efficiency of PCA decreases apparently when dealing with large amount of data. The detection algorithm can detect normal and abnormal user behavior precisely and effectively [4].

3.5 REP Tree

REP tree is the best classification algorithm for the classification of malicious and legitimate users for the datasets from social media networks like facebook and live journal. REP tree is compared with Naïve Bayes Multimonial Updateable, complement Naïve Bayes, classification via clustering. Classification is done on these algorithms on the basis of 10 fold cross-validation training and testing algorithm. Different parameters like TP, FP rate, precision, recall, F-measure and ROC area are calculated to find which classification algorithm is better. Accuracy of the REPTree is higher than other 3 algorithms. [6].

3.6 Online user behavior based sales prediction method

Sales prediction is complex task under the uncertainty of product demand. Some useful data mining tools such as neural networks and support vector regressions are employed to construct efficient sales prediction models. The set of data mining methods are proposed to forecast sales trend. An optimal data mining model is selected as the predictor by using the cross-validation method. 2 kinds of neural networks are employed – BPNN and RBFNN while 2 SVRs with different kernels are used. Each kind of SVR selects 3 different kernel functions, which are Linear, RBF, Polynomial and Sigmoid function. BPNN perform worse than RBFNN but RBFNN needs more time to get the final prediction. BPNN gives more accurate predictions. In future, an online sales prediction system can be developed to help managers in E-commerce companies for sales trend analysis [7].

3.7 Comparison metrics of twitter and facebook user behavior

3.7.1 CFFr (common twitter friend fraction)

CFFr measures the fraction of friends of users *u* in twitter who are also friends of *u* in facebook. CFFf considers only facebook and CFFt considers twitter. The metrics were summed, averaged and standard deviation was computed. The result showed that there was no significant overlap among the friends of the 2 accounts of a user in twitter and facebook. Overlap measured in twitter was higher than facebook. Average degree of twitter accounts was much higher than facebook. It has also shown that users of twitter have a degree lower than the users of facebook.

3.7.2 Normalized Activity Coefficient (NAC)

The behavior of 3 different types of users were investigated: a) declared, the users who declared account both in fb and twitter. b) hidden, the users who have account on both OSN but have not declared it. c) other, remaining users.

To analyze this, normalized activity coefficient (NAC) of a twitter account was defined as tc/ya, tc is the number of tweet posted and ya is the number of years since the account was created. NAC was calculated by using logarithmic binning function and its value was computed for each typology of users.

The study of OSN analysis has some limitations, a) this study is not generalized as it restricted only for 2 OSNs. b) only public information could be retrieved with limited size of sample. c) casual conclusion could not be drawn.

In future work, other OSNs can be considered with another extraction techniques [8].

3.8 K-Means

Users have been classified based on their search query terms. K-means clustering was applied on the 15 derived features from the dataset. Grouping was done on feature keywords. This clustering algorithm was executed for 100 iterations with n=6. Query keywords have beneficial correlation with the buying behavior of the users. These keywords can be used by the site owners to increase their market gains [9].

The users were also classified according to the number of clicks already done. Dataset was characterized based on various brands and on different price range. User behavior was characterized based on the price of different phones.

3.9 Markov Logic Network

To predict the future behavior of user, the model needs to build which is based on past user data i.e. session history. The past user data contains transitions, time spent, geography, and target value is required to predict i.e. whether the user converted or not. Session history was used to predict whether the user was going to convert or exit the site. Markov logic is good prediction model. This model also helps to devise a mechanism which tries out various features. Markov chain generates session traces between the existing states. Markov chain has time-homogeneity property which means the probability of going from one state to another does not depend on time at which chain is inspected. KL-divergence is used to determine the homogeneity in the Markov chain. It is a distance measure. The performance of MLN was compared with other 2 algorithms like SVM and CART and it has obtained similar results of predicting tasks [9].

4. ANALYTICAL EVALUATIONS

During classification, the system detects deviations between predicted and current user behavior. Coefficient of reliability estimates the value of deviation. 2 parameters describe a distribution. Each time while classifying, mean and standard deviation is calculated. Distribution is described by μ and σ. 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. After applying classifier on 2 datasets, REP tree is the best classification algorithm for the classification of malicious and legitimate users. Social network analysis is a huge source of information about people which may result in strategic and useful knowledge. Multiplicity of social networks is an important aspect which has taken into consideration while studying the complex phenomenon of OSNs like twitter and facebook. While comparing different mobile phones, users' behavior characteristics have been studied. The ranking of mobile phones which dominates the market has provided. Price of phones affects popularity and popularity prediction can be possible. User behavior characteristics at different times of day, days of week and date of month have been obtained. Chinese search engines' performance could be improved and algorithm would be optimized by studying power law distribution i.e. distribution of length of query, submission time of query, modification of the query, time interval between query submission and click etc. Most of the users give attention to the top five results returned by search engine. After understanding users' intention, search engine can be improved for information retrieval and knowledge mining.

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

Method	**Database used**	**Implementation**	**Result**
Layout method [2]	Online social communication data such as tweets (4 million), emails, and instant messages in an offline procedure.	58 communication features were extracted from the data. TargetVue system is used as primary analysis tool.	It captures similarity among users and compares communication behaviors of different users.
PCA [4]	Clinic db, dataset includes the application of article 8900 of the SQL trace. It includes 130 tables and a total of 1201 columns.	The system converts the newly-arrived request into a data point, and then checks the relative changes of these data points to the principle direction using PCA.	The model and detection algorithm can detect normal and abnormal user behavior precisely and effectively.
REP Tree, Naïve Bayes, Classification via Clustering algo[6].	2 real datasets were used. One was facebook links and other was live journal	Classification was done on the basis of 10 fold cross-validation training and testing algo. Parameters were calculated to see which classification algorithm was better.	REPTree is the best classifier among others.
NAC (normalized activity coefficient) of twitter account [8]	757 common users' set of pairs of accounts from Twitter and Facebook	NAC is discretized by applying the logarithmic binning function. Computed the value for each type of users.	More accounts' users are less active than others who have only twitter account. The users having more accounts and not declared FB account are very inactive.
K-means clustering [9]	Data from Smartpix.com. 3,274,505 sessions with 2,675,202 distinct users, 266,323 repeat users, 126,103 sessions.	Clustering was based on search query terms. Grouping was on 15 derived feature keywords. Algorithm was executed for 100 iterations with n=6	Query keywords have significant correlation with the buying behavior of the users which can be used for increasing the market gains.
Distribution of length of queries, position of clicks, click number of same query, etc [10]	Search log during a period of 31 days. 756 million entries of users' search requests which contain 101 million sessions.	Length of query – no of words or terms, position of click – after submitting queries,	Average length of queries – 1.85 terms, top 5 positions received more than 90% of clicks

5. PROPOSED APPROACH

A method which will extract user's behavior characteristics, it will describe user's behavior more completely and will improve the efficiency of the algorithm.

In data acquisition and preprocessing, web log file can be processed or database records can be analyzed. This data is put into data set vector. Required features will be extracted. 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.

The metrics like coefficient of reliability can be used to measure the value of deviation in user behavior and if the value is beyond the threshold then user is considered as malicious or abnormal user.

Methods like Gaussian Mixture Model and fuzzy logic in combining can be used to differentiate between normal and anomalous individuals.

Principal component analysis (PCA) can detect user abnormal behavior through the change of principal direction.

REPTree can classify original user behavior data into malicious and legitimate classes.

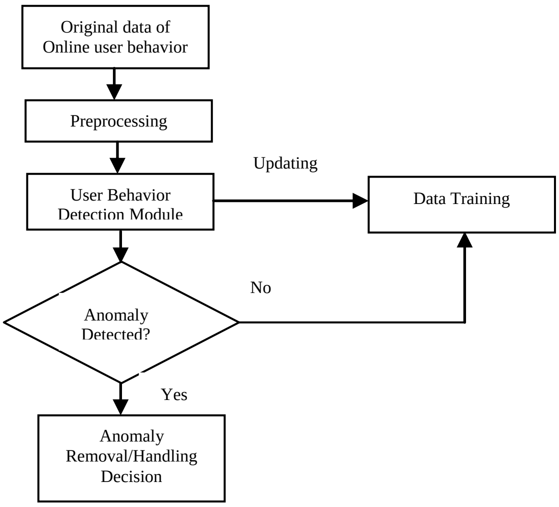
Original data of Online user behavior
Preprocessing
Updating
User Behavior Detection Module
Data Training
No
Anomaly Detected?
Yes
Anomaly Removal/Handling Decision

Fig. 1: Block diagram of proposed Anomaly Detection System

Table 2. Analysis of algorithms by considering parameters

Method	**Parameters Considered**	**Improvement**
Coefficient of Reliability [1]	Mean, standard deviation	If area limited to one σ – 68%, strong match. CR - encouraged. If area between 1 and 2 σ-27%, weak match, CR – should not be changed. If area between 2 σ, the value of CR pengalty should calculated.
PCA [4]	TP, FN, FP, TN, TPR, FPR, ACC	After removing application situation 7.0, overall average performance is very good
REP Tree, Naïve Bayes, Naïve Bayes and Classification via Clustering algorithm. [6]	TP rate, FP rate, precision, recall, F-measure, and ROC area, correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, root mean squared error, relative absolute error, root relative squared error	REPTree has the highest value. Lowest FP rate.
K-means [9]	Query keyword	Site owners use keyword to increase their market gains.
Markov Chain [9]	KL Divergence, time step	KL Divergence = 0 when clicks are from 5 to 30, then users have stable behavior. 13% data falls in this range.

6. SUGGESTIONS FOR FUTHER IMPROVEMENTS

Different classification method can be used and compared with coefficient of reliability to improve the performance of accuracy in the process of identification of normal and malicious user.

REP tree is found as the best classification algorithm for the classification of malicious and legitimate users from social media networks. Here, REP tree can be compared with more algorithms like random forest to improve the classification accuracy and other parameters. Random forest performs best in the prediction. The required input will be number of features and number of trees. As number of trees grow the accuracy increases. Random sets can be applied. Again neural network algorithms can be compared with random forest for further improvement.

CFF and NAC are the comparison metrics used to compare behavior of users on facebook and twitter. This study is limited for only 2 social networks. More users from different online social networks can be taken for further study.

By using k-means algorithms, user behavior was characterized based on prize of different phones. Prize was used as one of the feature. Instead brand of phone, different phone’s features like screen resolution, camera size, screen width, operating system etc.

MLN has obtained similar prediction results as compared to SVM and CART algorithms. Here, Linear and logical regression can be used to obtain improvement in result.

7. CONCLUSION

This analytical study discusses different characteristics of online user behavior models by using various methods and algorithms. Various domains are considered while detecting normal and anomalous online user behavior.

The characteristics of Intrusion detection system (IDS) is studied which was used to classify user behavior. TargetVue is the visualization system studied with different communication features of user behavior. The feature extraction technique like PCA is used to detect anomaly in user behavior. REPTree is found as the best classification algorithm which is studied with different parameters. Different neural network algorithms studied comparatively which are used for accurate sales prediction task. Different metrics are studied which were used to analyze users' behavior on 2 OSN platforms. How search engine's service can possible to improve is studied by analyzing users' queries.

8. FUTURE SCOPE

In future, Users' behavior on more social network platforms other than facebook and twitter will be studied.

9. REFERENCES

[1] Alexandr Seleznyov, Finland, "A methodology to detect temporal regularities in user behavior for anomaly detection", Network Security and Intrusion Detection, part 9, pp. 339-352, 2016.

[2] Nan Cao, Conglei Shi, Sabrina Lin, Jie Lu, Yu-Ru Lin, Ching-Yung Lin, "TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems", IEEE Transactions on Visualization and Computer Graphics, vol. 22, pp. 280-289, 1 January 2016.

[3] P V Bindua, P Santhi Thilagama, India, "Mining Social Networks for Anomalies: Methods and Challenges", Journal of Network and Computer Applications, Elsevier, pp. 1-22, 25 Feb 2016.

[4] Meng Bi, Jian Xu, Mo Wang, Fucai Zhou, "Anomaly detection model of user behavior based on principal component analysis", Springer, 21 January 2016.

[5] Guirong Chen, Ning Wang, Fengqin Zhang, Hua Jiang, China, "Understanding the Time Characteristic of User Behavior on Online Forums", IEEE International Conference on Big Data (Big Data), 978-1-4799-9926-2/15,pp. 2300-2306, 2015.

[6] Pran Dev, Dr. Kulvinder Singh, Dr. Sanjeev Dhawan, India, "Classification of Malicious and Legitimate Nodes for Analyzing the Users' Behavior in Heterogeneous Online Social Networks", 1st International conference on futuristic trend in computational analysis and knowledge management, 978-1-4799-8433-6/15, pp. 359-363, ABLAZE 2015.

[7] Hui Yuan, Wei Xu, Mingming Wang, China, "Can Online User Behavior Improve the Performance of Sales Prediction in E-commerce?", IEEE International Conference on Systems, Man, and Cybernetics, 978-1-4799-3840-7/14, pp. 2347-2352, October 5-8, 2014.

[8] Francesco Buccafurri, Gianluca Lax, Serena Nicolazzo, Antonino Nocera, "Comparing Twitter and Facebook user behavior: Privacy and other aspects", Computers in Human Behavior, Elsevier, 0747-5632, pp. 87-95, 10 June 2015.

[9] Mona Gupta , Happy Mittal , Parag Singla , Amitabha Bagchi, "Characterizing comparison shopping behavior: A case study", ICDE Workshops, IEEE, 978-1-4799-3481-2/14, pp. 115-122, 2014.

[10] Ke XIE, Huijia YU, Rongwei CEN, "Using log mining to analyze user behavior on search engine", Front. Electr. Electron. Eng., Higher Education Press and Springer-Verlag Berlin Heidelberg 2011, pp. 254-260, 2011.