IJRSET NOVEMBER 2018 Volume 5, Issue 11 Pages: 23-29

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

LEARNING PROBABILISTIC USER BEHAVIOR MODELS FROM DATABASE ACCESS LOG

[1] Jeevanandhini .P, [2] A. Nirmala,
[1] Research Scholar, [2] Assistant Professor,
[1, 2] Dr. NGP arts and science college, Tamilnadu. India.

ABSTRACT- Client behavior modeling is a standout amongst the most imperative and fascinating issues should have been understood when creating and abusing present day programming frameworks. By client behavior modeling we mean finding examples of client action and developing prescient models dependent on point of reference behavior data. These models allow gauging next client activity based on the present movement. Essentially such technique was arranged to the commercial applications in proposal frameworks. At present time the region of its application is altogether more extensive. These methods assume an incredible job in PC security frameworks, where they are utilized for detecting malicious or unqualified client activities. In this paper proposed to learning probabilistic user behavior model approach used into database log process and their experimental results are checked into better performance of existing methods.

Keywords - [Probabilistic, Behavior Model, Learning, Discovery, Knowledge.]

1. INTRODUCTION

Data mining has pulled in a lot of consideration in the information business and in the public arena all in all lately, because of the wide accessibility of immense measure of data and the fast approaching prerequisite for changing over such data into valuable information. The information and the data picked up can be utilized for applications running from market study, extortion acknowledgment and client protection to create, sort out and science discovery. Basically to affirm, data mining alludes to extricating or "mining" vital information from immense measure of information. The articulation is really a misnomer. Hence, a misnomer that conveys both "data" and "mining" turned into a very much loved determination. A few different terms hold a comparable or to some degree diverse significance to data mining, for example, knowledge mining from information, data/design analysis, data prehistoric studies, data digging and knowledge extraction. Many people treat data mining as a synonym for one more widely used term, Knowledge Discovery from Data (KDD). On the other hand, data mining can be viewed simply as an essential step in the process of knowledge discovery.

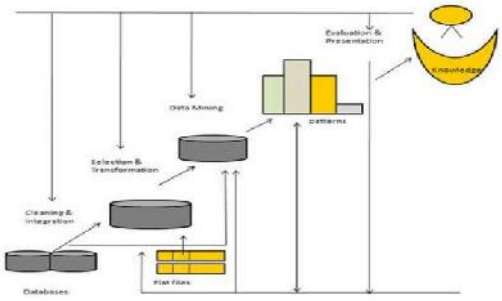

Figure 1: Process of Knowledge Discovery

Figure 1 demonstrates the Knowledge Discovery comprises of an iterative arrangement of the accompanying advances: 1. Data cleaning: Data cleaning is a technique that is connected to expel the uproarious data and right the irregularity in information. Data cleaning includes transformation to precise the mistaken data. Data cleaning is executed as data preprocessing venture before setting up the data for a data distribution center. 2. Data combination: Data Integration is a data preprocessing technique that blends the data from various heterogeneous data sources into a lucid data store. Data reconciliation may ingest capricious data in this way needs data cleaning. 3. Data determination: Data Selection is where data important to the analysis errand is recovered from the database. Sometimes data transformation and union are performed preceding data determination technique. 4. Data transformation: In this progression data is changed or merged into structures proper for mining by performing rundown or accumulation activities. 5. Data mining: In this progression keen techniques are connected with the end goal to separate data designs. Stages 1 to 4 are distinctive types of data preprocessing, wherever the data is prepared for mining. The data mining step may interrelate with the client or the information base. The fascinating examples are offered to the client and might be put away as unique data in the information base. As per this analysis, data mining is reminded just as one stage in the entire procedure, despite the fact that a critical one since it covers known examples for estimation. We concur that data mining is a stage in the data discovery method. In any case, in industry, in media and in the database investigating environment, the term data mining is ending up more mainstream than the more extended expression of data discovery from data. Along these lines, here we utilize the expression data mining. We execute a wide perspective of data mining usefulness. Data mining is the improvement of finding intriguing knowledge from immense measure of information put away in databases, data distribution centers or extra information stores. In view of this analysis, the engineering of a particular data mining structure may have ensuing significant components.

2. LITERATURE SURVEY

1. Liang Zhao, Zhikui Chen, Yueming Hu, Geyong Min and Zhaohua Jiang (2014) proposed framework for productive analysis of high-dimensional financial huge data dependent on imaginative disseminated include determination. In particular, the framework joins the techniques for monetary element determination and econometric model development to uncover the concealed examples for financial improvement. The usefulness lays on three columns: (I) novel data pre-handling techniques to get ready high-quality monetary data, (ii) a creative dispersed element distinguishing proof answer for find vital and delegate financial markers from multidimensional data sets, and (iii) new econometric models to catch the concealed examples for monetary improvement. To start with, fundamental components are connected to depict the instrument of financial development. The monetary development can be advanced by expanding utilization and venture, and in addition influencing related unequivocal variables. When moving toward monetary analysis, the contributing variables are chosen to recognize the relations among them and financial improvement. Second, from the point of view of cost sparing, urbanization can bring more workforces into city, which lessens the financial expenses and lifts offices sharing to chop down exchange costs. In the interim, through the agglomeration and dissemination impacts, the financial development can be quickened. Third, components and inside affiliations are included to exhaustively clarify the connections among's economy and its definitive elements. **2. Mohammad Shorfuzzaman (**2017) proposed the effect of big data examination on knowledge the board

and proposes a cloud-based theoretical framework that can break down big data progressively to encourage upgraded basic leadership planned for upper hand big data because of its different properties like high volume, assortment, and speed can never again be successfully put away and investigated with conventional data the executives techniques . New technologies and designs are required to store and break down this data and thus produce indispensable ongoing information for basic leadership in associations. This has opened the entryway for the scientists to center around big data examination which are probably going to assume an extreme job in the accomplishment of associations. The test is to gather, store, and break down the undertaking big data at the correct speed from sources, for example, deals, inventory network, research, and client relations to manufacture the knowledge base for compelling basic leadership of the associations. Ongoing investigations additionally uncovered the way that the usage of big data has enlarged prominently in basic leadership and both open and private associations are receiving rewards from this rising innovation. There is no endless supply of big data accessible in the writing. Be that as it may, big data is for the most part portrayed by three substances, for example, volume, speed and assortment. Volume speaks to the substantial measure of data that are gathered which for the most part go from terabytes to Exabyte. This isn't sufficient to express the genuine importance of the idea. **3. Dr. Venkatesh Naganathan** (2018) proposed Big data its difficulties and its future degree where it is driving as well and Big Data Analytics strategies utilized by various associations that encourages their business to settle on a solid speculation choices. Data and examination are at the core of the advanced upheaval. They are a basic over all enterprises. To endure and flourish in the advanced period, right now is an ideal opportunity to drive data and examination into the center of your business and scale outward to each worker, client, provider, and accomplice. Scaling the estimation of data and investigation requires a culture of data enablement that reaches out all through each aspect of your association. 4. **RakeshRanjan Kumar& BinitaKumari** (2015) proposed Big Data mining, issues identified with mining and the new chances. DATA MINING PARAMETERS The most usually utilized techniques in the data mining are: Association - Looking for examples where one occasion is associated with another occasion. Counterfeit neural networks - Non-straight prescient models that learn through preparing and take after natural neural networks in structure Classification - is a systematic procedure for acquiring imperative and important information about data, and metadata – data about data. Clustering - the way toward recognizing data sets that are like each other to comprehend the distinctions and in addition the similitudes inside the data. **5. Vivekananth.P , Leo John Baptist.A** (2015)proposed distinctive data investigation techniques, for example, text examination, audio analytics, video analytics, online networking investigation and prescient examination The resultant impact of having such an enormous measure of data will be data investigation. Data examination is the way toward organizing big data. Inside big data, there are distinctive examples and relationships that make it workable for data examination to improve ascertained portrayal of the data. This makes data examination a standout amongst the most critical parts of information innovation. There are diverse techniques that are at present being used. As a rule, the techniques can be summed up to: 1) Association Rule Learning 2) Classification tree analysis 3) Genetic calculations 4) Machine learning 5) Regression analysis 6) Sentimental Analysis 7) Social network analysis.

3. PROPOSED WORK

3.1 Learning Probabilistic User Behavior Model Approach

Before we turn to the problem of constructing function (1) we need to define the structure of database access log in the form of sequences of actions $A_i \epsilon$. Most of database access logs consist of records of similar structure:

[user id , event , sql, time , other features]

where user id is user login; event is a type of event(e.g., start or finish of a query execution); sql is SQL text of a query; time is a timestamp; other features can be divided into execution group that includes numerical characteristics of query execution (e.g. number of read/write operations,
duration, etc); and identification group with discrete characteristics of query such as identifiers of client process, server's process, user aliases, etc. Thus the problem is to map such structure into a finite alphabet. We suggest the following procedure.

DB Access Log Pre-processing Procedure:

Step 1. "Uninteresting" attributes reduction.

Step 2. Numeric attributes descritization.

Step 3. Extracting templates(skeletons) from SQL statement.

Step 4. Mapping discrete attributes combination to finite alphabet Ω .

On the first step we exclude attributes that are not interesting for analyzing. For example, db server process id, as a rule, is not interesting for the model. On the second step the rest numerical attributes are discretizied by some unsupervised discretization algorithm. In particular, we use equal frequency interval method with small (3-10) number of intervals. On the next step SQL statement text is processed. We extract its so called skeleton or, in other words, template, that presents the query syntax with removed user parameters. We use the approach similar to. SQL statement is converted into the sequence of tokens, where each token has either keyword type (for SQL language keywords) or name type (for db related names, i.e. table names, fields, stored procedure, etc.). Let us clarify this idea on the example. Assume we are given the following query:

SELECT FROM USERS WHERE NAME='Bob' ANDCITY='London'

We convert it to the sequence of tokens:

(SELECT, keyword) (FROM, keyword) (USERS, name) (WHERE, keyword) (NAME, keyword) (AND, keyword) (CITY, name)

Thus, before the fourth step the initial log document record has the type of vector of discrete attributes, where each attribute is either discrete attribute of the initial record or SQL layout identifier, or interval id of discretizied initial numeric attribute. Records with the equivalent SQL format, the equivalent discrete attributes and close numeric attributes have a similar portrayal, i.e. a similar blend of coming about discrete attributes. Such portrayal of the comparative records recognizes the conceivable activity, to which a one of a kind image from the alphabet is appointed. In this way the alphabet Ω determines the set of all conceivable client activities. At the principal look the recommended method can be censured for the likelihood of unbounded growth of the span of the alphabet Ω . By and by, for creation frameworks being in stable abuse, it is discovered that the growth arrives at stop quickly enough, only for couple of hundreds. Additionally, number of various conceivable activities can be lessened by gathering them, utilizing bunching or visit scenes or master's space knowledge.

Subsequent to applying this technique for mapping db get to logs structures into the alphabet Ω we can utilize traditional data mining methods dependent on affiliation rules, sequential models or autoregressive classification. However, as we sketched out previously, these methods don't take time highlight into record, just the request. To evade this issue we propose novel approach. Its primary thought is building empirical element mapϕ that expressly maps a

discretionary succession of images from Ω with timestamps into a limited measurement metric space H. Above all else, we have to expand the portrayal of client activities by adding time names to them. Then each activity from S(U) or H(U) is depicted by the match (A,tm)∈Ω×Time .

4. EXPERIMENTAL RESULTS

Detection

Existing 1	Existing 2	Proposed
0.09	0.04	0.13
0.14	0.08	0.2
0.19	0.13	0.28
0.25	0.19	0.39
0.3	0.22	0.45

Table 1:Comparison table of Detection

The comparison table of detection explains different values of existing method and proposed method. While comparing the existing and proposed method the proposed method shows the highest value. Existing 1 value starts from 0.09 to 0.3 existing 2 values starts from 0.04 to 0.22 and proposed method values start from 0.13 to 0.45. The proposed method shows the better results.

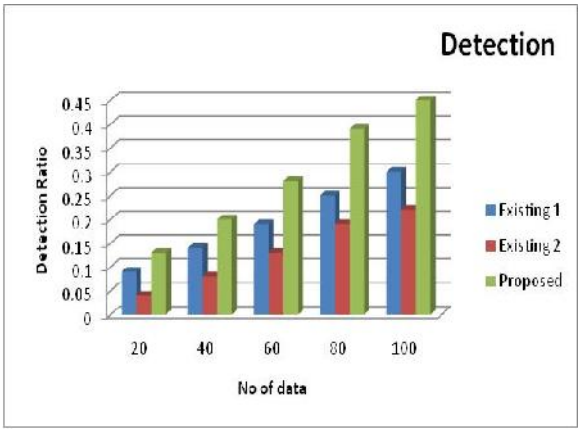

Figure 2: Comparison chart of Detection

The comparison chart of detection explains the values of existing and proposed method. Detection ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value 0.09 to 0.3 existing 2 values are 0.04 to 0.22 and proposed method values are 0.13 to 0.45.

Effectiveness

Existing 1	Existing 2	Proposed
57	69.5	83
59	69.9	84.8
62	69.5	87.9
66	70.9	90.2
69	72	93.6

Table 2:Comparison table of Effectiveness

The comparison table of effectiveness explains different values of existing method and proposed method. While comparing the existing and proposed method the proposed method shows the highest value. Existing 1 value starts from 57 to 69 existing 2 values starts from 69.5 to 72 and proposed method values start from 83 to 93.6. The proposed method shows the better results.

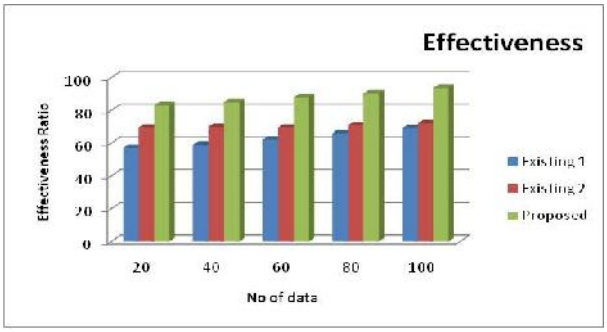

Figure 3: Comparison chart of Effectiveness

The comparison chart of effectiveness explains the values of existing and proposed method. Effectiveness ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value 57 to 69 existing 2 values are 69.5 to 72 and proposed method values are 83 to 93.6.

Deployment

Existing 1	Existing 2	Proposed
31.9	39	66
37.7	45	72
42.6	49	76.5

50.4	55	79.8
55.23	58	85

Table 3: Comparison table of Deployment

The comparison table of deployment explains different values of existing method and proposed method. While comparing the existing and proposed method the proposed method shows the highest value. Existing 1 value starts from 31.9 to 55.23 existing 2 values starts from 39 to 58 and proposed method values start from 66 to 85. The proposed method shows the better results.

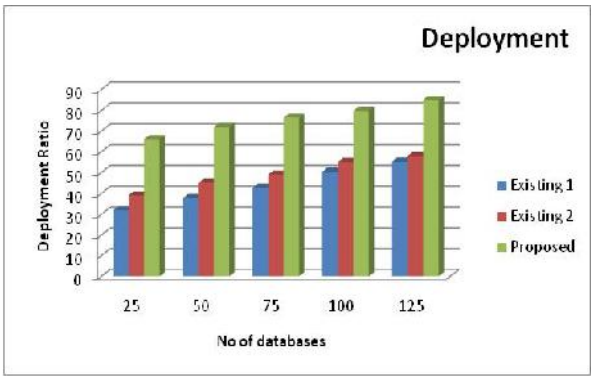

Figure 4: Comparison chart of Deployment

The comparison chart of deployment explains the values of existing and proposed method. Deployment ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value 31.9 to 55.23 existing 2 values are 39 to 58 and proposed method values are 66 to 85.

Discovery

Existing 1	Existing 2	Proposed
55	67	75
58.6	70.1	78.9
62.3	74.8	83.86
68.9	78.89	88.21
72	81	92.06

Table 4: Comparison table of Discovery

The comparison table of discovery explains different values of existing method and proposed method. While comparing the existing and proposed method the proposed method shows the highest value. Existing 1 value starts from 55 to 72 existing 2 values starts from 67 to 81 and proposed method values start from 75 to 92.06. The proposed method shows the better results.

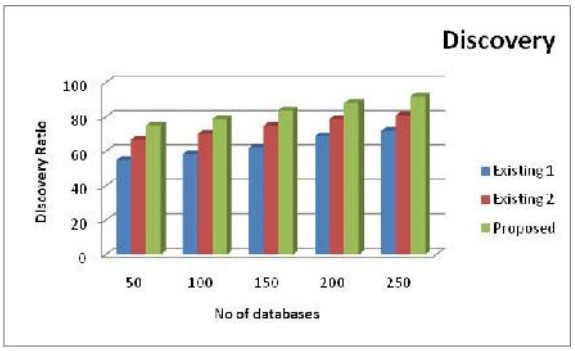

Figure 5: Comparison chart of Discovery

The comparison chart of discovery explains the values of existing and proposed method. Discovery ratio in x axis and no of databases in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value 55 to 72 existing 2 values are 67 to 81 and proposed method values are 75 to 92.06.

CONCLUSION

Novel method for mining probabilistic client behavior models has been defined. Unlike other existing data mining methods it fuses time highlight in the client model. The empirical element outline, by potential functions theory, has been proposed for that. Joining this element outline with choice tree algorithm we get new method with following focal points: it is exact enough; it takes into record time intervals between client activities; it gives reasonable for a human master interpretation of produced behavior models as "Assuming… THEN" rules. Experimental execution evaluation on realworld data has been led. It has illustrated that database get to logs can be effectively utilized for client behavior modeling and dependable models can be built. In these tests, our proposed method has shown extraordinary outcomes in the "following activity expectation" situation

and competitive outcomes in "anomaly detection" situation.

REFERENCES

[1]. **Liang Zhao, Zhikui Chen, Yueming Hu, Geyong Min and Zhaohua Jiang**, "Distributed Feature Selection for Efficient Economic Big Data Analysis", JOURNAL OF LATEX CLASS FILES, VOL. 13, NO. 9, SEPTEMBER 2014.

[2]. **Mohammad Shorfuzzaman**, "LEVERAGING CLOUD BASED BIG DATA ANALYTICS IN KNOWLEDGE MANAGEMENT FOR ENHANCED DECISION MAKING IN ORGANIZATIONS", International Journal of Distributed and Parallel Systems (IJDPS) Vol.8, No.1, January 2017.

[3]. **Dr. Venkatesh Naganathan**, "Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends", 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis.

[4]. **Rakesh Ranjan Kumar & Binita Kumari**, "Visualizing Big Data Mining: Challenges, Problems and Opportunities",) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3933-3937.

[5]**. Vivekananth.P , Leo John Baptist.A**, "An Analysis of Big Data Analytics Techniques", Volume-5, Issue-5, October-2015 International Journal of Engineering and Management Research Page Number: 17-19.

[6]. **Dr.M.Padmavalli**, "Big Data: Emerging Challenges of Big Data and Techniques for Handling", IOSR Journal of Computer Engineering (IOSR-JCE).

[7]. **Althaf Rahaman.Sk,Sai Rajesh.K,.Girija RaniK**, "Challenging tools on Research Issues in Big Data Analytics", International Journal of Engineering Development and Research.

[8]. **Sofiya Mujawar , Aishwarya Joshi**, "Data Analytics Types, Tools and their Comparison", International Journal of Advanced Research in Computer and Communication Engineering.

[9]. **Yolanda Gil**, "Teaching Big Data Analytics Skills with Intelligent Workflow Systems", National Conference of the Association for the Advancement of Artificial Intelligence (AAAI), Phoenix AZ, 2016.

[10]. **Revanth Sonnati**, "Improving Healthcare Using Big Data Analytics", NTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 6, ISSUE 03, MARCH 2017.

[11]. **Mengru Li, Hong Fu , Ruodan Sun and Che Che**, "The Application of Big Data Analysis Techniques and Tools in Intelligence Research", International Conference on Communications, Information Management and Network Security (CIMNS 2016).

[12]. **M. Dhavapriya, N. Yasodha**, "Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016.

[13]. **Manisha R. Thakare, S. W. Mohod and A. N. Thakare**, "Various Data-Mining Techniques for Big Data", International Conference on Quality Up-gradation in Engineering, Science and Technology (ICQUEST2015).

[14]. **M.Chalapathi Rao, A.Kiran Kumar**, "Challenges arise of Privacy Preserving Big Data Mining Techniques", International Research Journal of Engineering and Technology (IRJET).

[15]. **B R Prakash , Dr. M. Hanumanthappa**, "Issues and Challenges in the Era of Big Data Mining", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).