iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Monitoring Unusual/Abnormal Behavior of Social Media Users

E.Surya[1], P.Sivaranjani[2],
[1, 2] *Assistant Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai,*

Abstract*: Terrorists are using the online environment to prey on the young and vulnerable. Nowadays a lot of terrorist activities are initiated through social media. A study says that these groups not only use social media for communication but are recruiting vulnerable youths into their groups. How can such activities be prevented? How such activities can be identified and eliminated at root level? In order to overcome these kinds of threats, this project deals about development of a monitoring system, this monitors the unusual behavior of users by comparing the dynamic data with the behavior datasets and process the information using natural language processing.
Keywords*: NLP (Natural Language Processing), Sequence Pattern analysis.

I. INTRODUCTION

Among Mining, Sequential Patten mining is the technique of finding fascinating sequential patterns among the large databases. From a sequence of database it also finds out frequent sub sequences as patterns. Many industries are interested in mining so they are gathering enormous amounts of data are continuously being collected and stored as they are showing interests in mining sequential patterns. Sequential pattern mining has wide range of applications in record analysis, web-log analysis, and client purchase behavior analysis [11]. It is the task of finding patterns which are present in a certain number of instances of data. The identified patterns are articulated in terms of sub sequences of the data sequences and expressed in an order that is the order of the elements of the pattern where it appears. If it appears in a number of instances above a given threshold value, usually defined by the user, then it is measured to be frequent. Sequential pattern mining is used to identify whether any relationship occurs in between the events. The sequential patterns that occur in particular individual items can be found and also the sequential patterns between different items can be found.

The aim of this paper is to characterize and detect unusual or abnormal behaviors of social media users. We propose a monitoring system that observes users' sequential topic patterns using the document streams on the Internet. Textual documents created and distributed on the Internet are ever changing in various forms. In this paper, in order to characterize and monitoring personalized and abnormal behaviors of Internet users, we propose Sequential Topic Patterns Mining. They are unusual on the whole but relatively frequent for specific users, so can be applied in many real- life scenarios.

We solve this problem through three phases: preprocessing to extract topics and identify sessions for different users, generating all the candidates with (expected) support values for each user by pattern-growth, and selecting rare patterns by making user-aware rarity analysis on derived topics. Experiments on both real datasets show that our approach can indeed discover special users effectively and efficiently, which significantly reflect users' characteristics. In order to characterize user behaviors in document streams, we study on the correlations among topics extracted from these documents, especially the sequential relations, and specify them as Sequential Topic Patterns. Each of them records the complete and repeated behavior of a user when she/he is publishing a series of documents, and is suitable for inferring users' intrinsic characteristics and psychological statuses. In this way we can effectively monitor the users' activity.

II. LITERATURE SURVEY

[1] This paper calls for effective ways to precisely monitor analyze and recapitulate the important information present in an online and used conventional term-based and word-based approaches used for information filtering. Topic model has used for discovering unseen topics in a set of credential. The creature of habit mining technique used in field of topic modeling generates model for finding out more meaningful and discriminative topics from collection of documents.

This paper [2] proposed the work which is used for extending the sequential mining approaches. They proposed the sequential pattern mining algorithm which may solves the problem of discovering the presence of frequent sequences in the given database. The disadvantage of this paper is it is difficult to engender and examine a number of intermediate sub sequences. And they suggested to use the apriori algorithm which used can be extensive to web content mining and web structure mining analysis.

This author had constructed a new [7] result patterns which are involved based on document topics, and has wide potential application scenarios, like real-time monitoring on abnormal behaviors of Internet users. Various mining problems and a group of algorithms being designed and combined to analytically solve the problems. The author conducted the testing on real-time applications like Twitter and Gmail to exhibit the proposed idea to prove it efficient and effective to unmask the aberrant behavior of Internet users. Since the paper emphasizes on web data mining, one can work to enhance the techniques to take it a notch up. They also suggested enhancing the mining algorithm to focus much on degree of parallelism, and research on-the-fly algorithms targeting at real-time document streams.

III. PROPOSED SYSTEM

In this system, users' abnormal or unusual behavior can be monitored using sequence of document streams from multiple web applications. Our system extracts users' activity on real time data set available on Twitter and Gmail. Using this technique we can monitor the user's topic pattern based on their session identification number on multiple web applications with single sign on email id and their session id. To extract the topic and to mine the users' activity we have used the data of inbox and sent box mail content of Gmail and twitter's tweet and individual chats. Using NLP processing user's different activities can be extracted and monitored effectively. It is worth noting that the ideas above are also applicable for another type of document streams, called browsed document streams, mining users' rare sequential topic pattern can better discover special interests and browsing habits of Internet users, and is thus capable to give effective and context-aware recommendation for them.

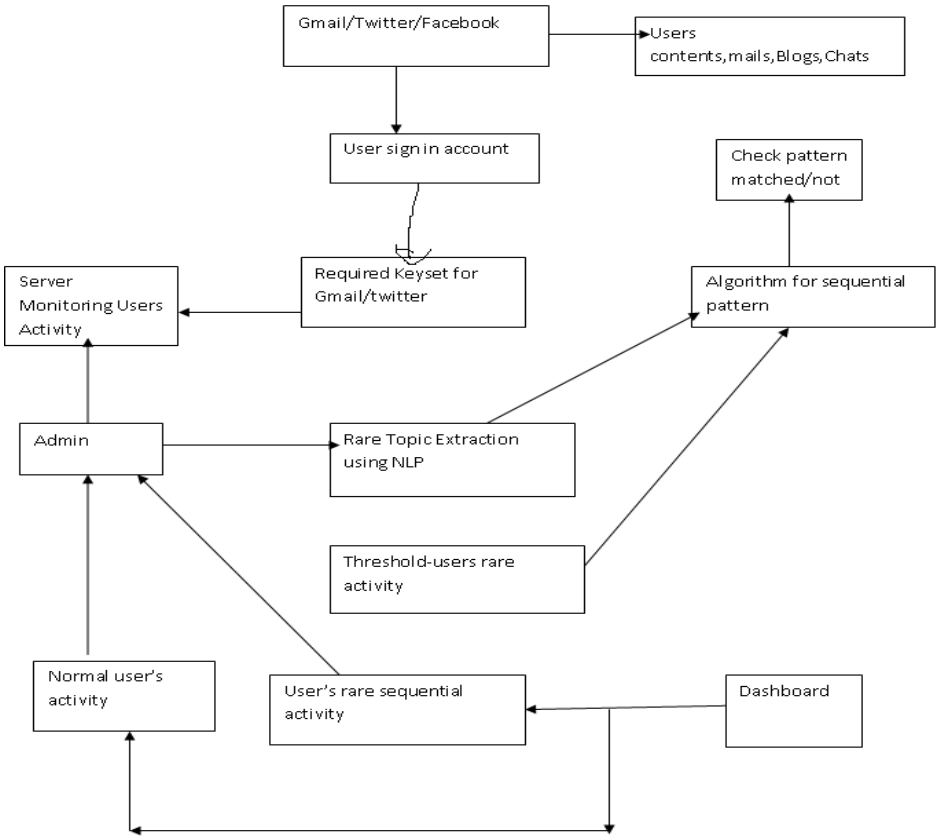


Fig. 1 Architecture diagram

While, this paper will concentrate on mining and identifying users' rare pattern and leave the applications for recommendation to future work. To tackle many challenges raised in mining rare pattern in document streams we initiate the existing techniques of mining cannot be directly applied to solve this problem because the input for this task is a textual stream.

A preprocessing phase is necessary to get exact descriptions of documents, and then to recognize repeated activities of Internet users by session identification. Second, in many applications the real-time requirements include both the accuracy and the efficiency of

mining algorithm. Third, rare patterns are completely different from normal patterns. And in the same way, unsupervised mining algorithms for this kind of patterns need rare to be planned in a manner different from existing frequent pattern mining algorithms.

IV. MODULES AND DESCRIPTION

A. Creating Datasets

The users have to register their email id and twitter key with our application. The email id and twitter key's id should be a single sign on Gmail and Twitter account. We build Stanford NLP algorithm to mine the user's activity. The data has been maintained and customized in the server. In this API we implements pos tagging, chunking processing, stemming, spell checking and word net connection.

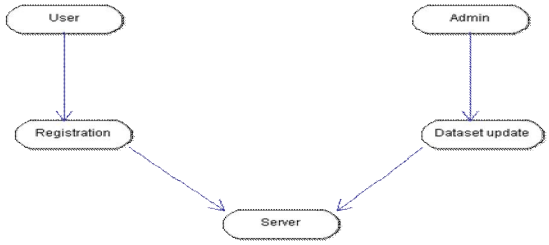


Fig. 2 Registration

B. NLP processing on Gmail and twitter content

The user's details can be extracted and monitored from the Gmail and Twitter to our local server database. Because of the huge amount of dataset we create threshold for retrieving data from the Social Medias. Using twitter key and email id the mail content and twitter content can be extracted using Java Mail API and Twitter4j API. The type of data set can be categorized like inbox, sent items, mail chats, Spam, drafts, user's tweets, twitter chats and micro blogs maintained in our local server database.

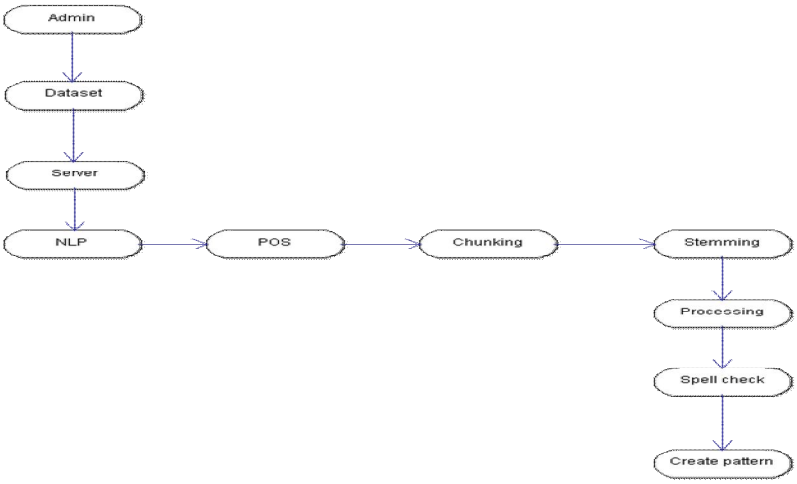


Fig. 3 NLP Processing

The Server will monitor all the extracted topics of the user's contents and creates a post tagging which are the parts of speech of the each content of the user's data set. Stemming process groups the similar types of words of the content like calling, call, called and callable, etc. Chunking process removes the common words filtering on the content like is, was, the, of, on, off etc.

C. Monitoring user's activity

The Server monitors every user's activity on Gmail and Twitter, Face book and yahoo. Single user activity on the two different web applications can be identified and extracted using single sign on email ids. The sequential topic extraction on sequence of

documents are extracted and grouped. Our application will grasp the individual topics as well as sequential relations of topics in successive documents published by a specific user.

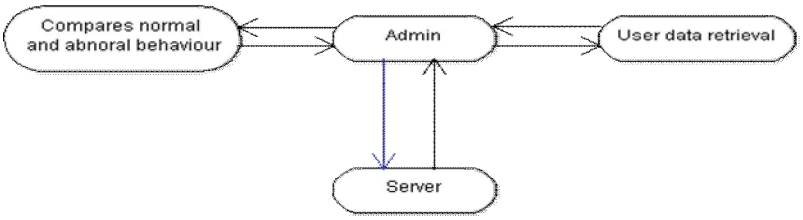


Fig. 4 Monitoring

For a document stream, some patterns may occur frequently and thus reflect common behaviours of involved users. Beyond that, there may still exist some other patterns which are globally rare for the general population, but occur relatively often for some specific user or some specific group of users. We call them User-aware Rare Sequential topic pattern.

D. Mining Rare Sequential Activity

If illegal behaviour where involved during monitoring and extraction of the users sequential topics .we can still expose them as long as they satisfy the properties of both global rareness and local frequentness. That can be regarded as important clues for suspicion and will trigger targeted investigations. Therefore, mining rare sequential topic patterns is a good means for real-time user behaviour monitoring on the Internet. We implement the aware recommendation on admin dashboard and highlight the rare user's activity and normal user's interest based on their social network behaviour.

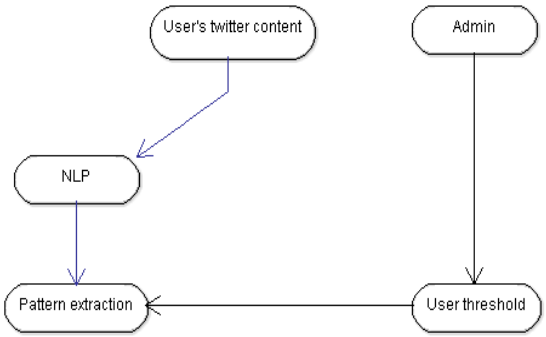


Fig. 5 Mining

E. Sequential Pattern Mining

There are two types of sequential data which is commonly used in data mining time-series and sequences. A sequence is an ordered list of nominal values (symbols) while time-series is an ordered list of numbers,. For example, Fig. 6 (left) shows a time-series representing amounts of money, while Fig. 6 (right) depicts a sequence of symbols (letters). Both time-series and sequences are used in various domains. For example, time-series are often used to symbolize data such as stock prices, temperature readings, and electricity consumption readings, while sequences are used to represent data such as sentences in texts (sequences of words), sequences of items purchased by customers in retail stores, and sequences of web pages visited by users.

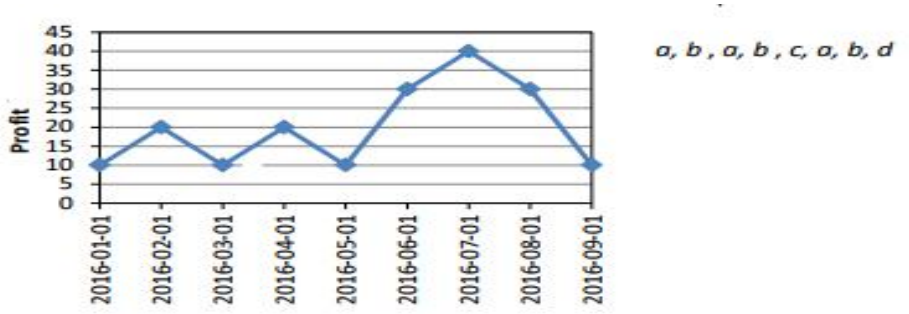


Fig. 6 Time and Sequence Data Set

IV. RESULT

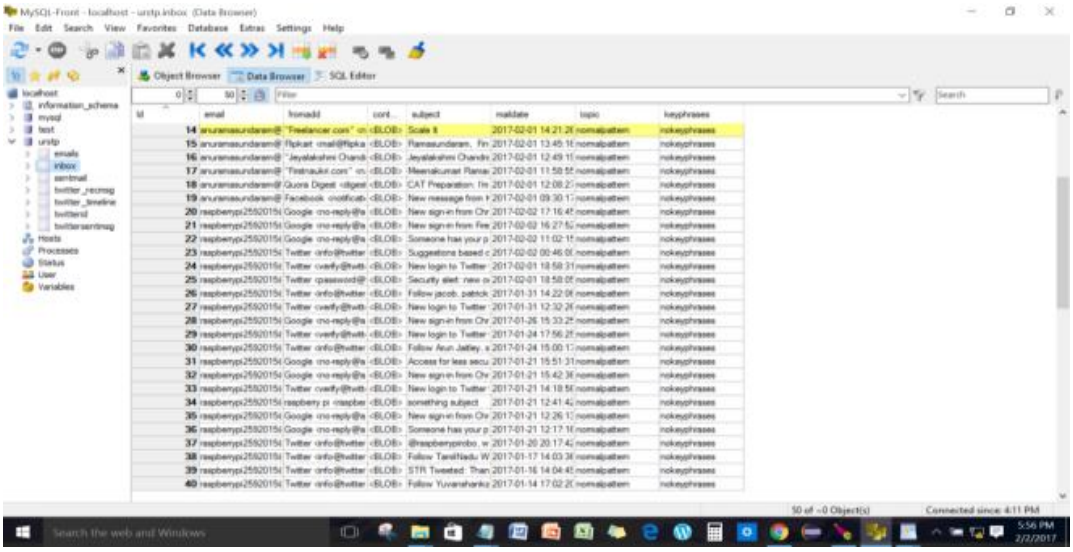


Fig. 7 Email Retrieved Messages

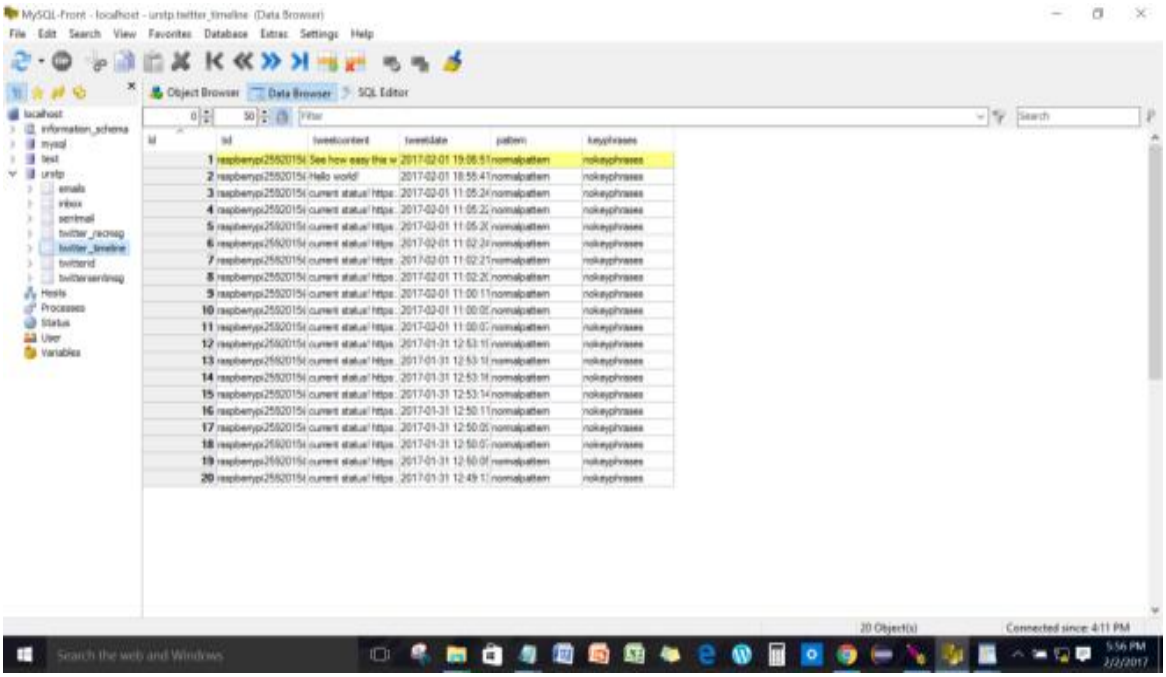


Fig. 8 Twitter Retrieved Messages

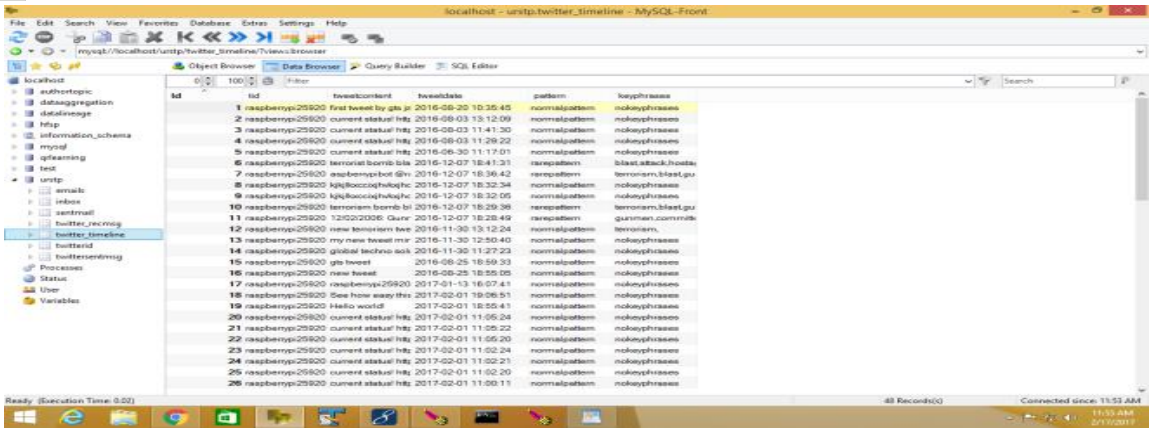


Fig. 9 Twitter Timeline Pattern Analysis

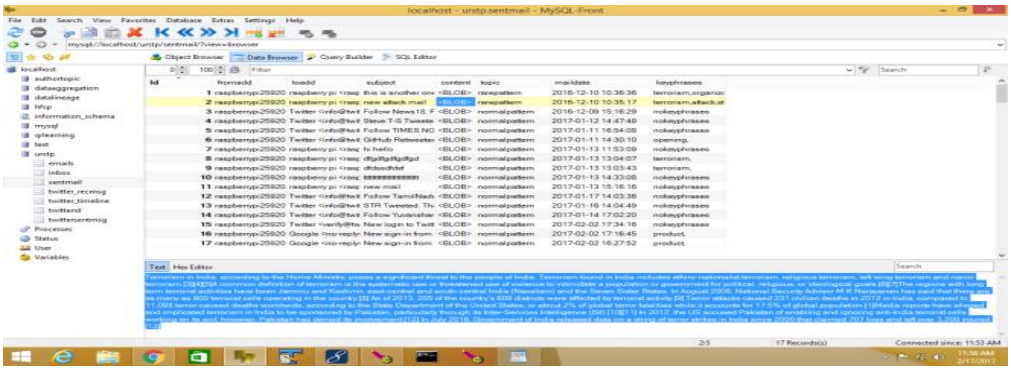


Fig. 11 Mail Pattern Domain Mapping

V. CONCLUSIONS

Thus mining users' activity with the datasets from more than one web application is an effective and efficient way of monitoring. This method monitors users' sequential activities which paves way for us to extract the required information without any ambiguity. NLP is one of the most efficient ways to extract the exact information from a piece of data. Sequential topic based monitoring associates' data with the topics as well as recognizes any sequential activity of the user thus identifying unique patterns. With these patterns we are able to distinguish between the normal users and abnormal/unusual users. If a users' profile contains any anti-social data or information above a certain threshold, his usage pattern will be notified to the admin as a rare pattern so that he/she can be tracked for further investigation.

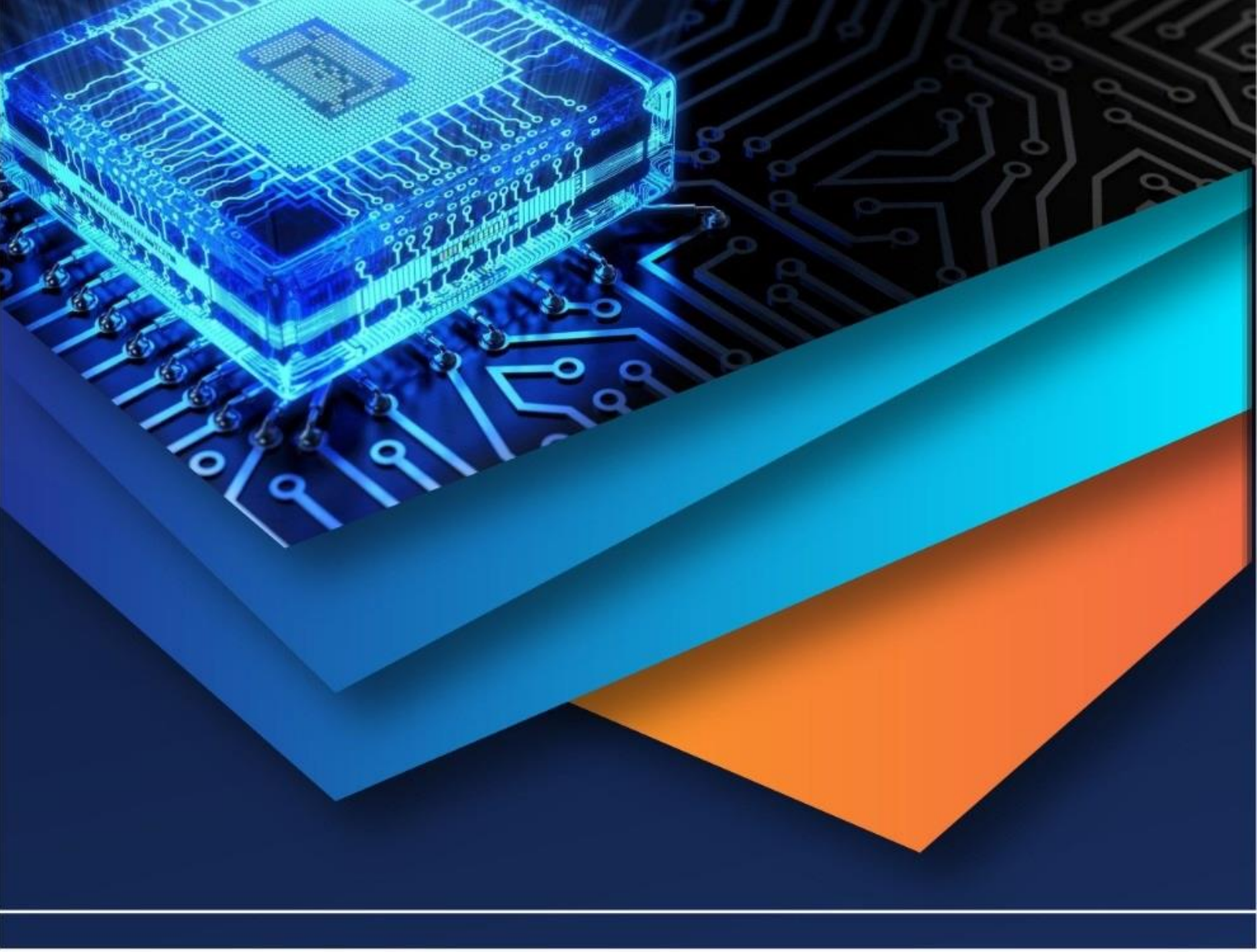
VI. FUTURE ENHANCEMENTS

It is a third party application as of now, in future user's behaviour will be automatically observed in Gmail/twitter with this application. Admin needs to login every time and check for abnormal behaviour in the current system, in future, admin will get alerts from the system if it encounters any abnormal behaviour and will also enable context-aware recommendation.

REFERENCES

[1] P Prof. B. Vikhe , S. P. Katore "Rare Sequential Topic Patterns in Document Stream"IJARIIE-ISSN(0)-2395-4396,VOL-2 Issuse-6 2016.
[2] R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc.IEEE Int.Conf. Data Eng., 1995, pp.3–14.
[3] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in Proc. 31st Int. Conf. Very Large Data Bases, 2005, pp. 181–192.
[4] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, "FreeSpan: frequent pattern-projected sequential pattern mining," in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2000, pp. 355–359.
[5] N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latent topic sequential patterns," in Proc. 6th ACM Conf. Recommender Syst., 2012, pp. 131–138.
[6] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999, pp. 50–57.

[7] M.Sangeegtha, D.Swathi,J.Priyanka, Shalini Yuvaraj "Prediction Of User Rare Sequential Topic Patterns Of Internet Users"international Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072 © 2017, IRJET | Impact Factor value: 5.181 | ISO 9001:2008 Certified Journal | Page 965.
[8] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. 1st Workshop Soc. Media Anal., 2010, pp. 80–88.
[9] Z. Hu, H. Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, "Discovery of rare sequential topic patterns in document stream," in Proc. SIAM Int. Conf. Data Mining, 2014, pp. 533–541.
[10] A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 497–504.
[11] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in Proc. ACM Int. Conf. Mach. Learn., 2006, vol. 148, pp. 577–584.
[12] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR '05, pp.
[13] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
[14] Guha & Garg, 2004] R. Guha and A. Garg, "Disambiguating People in Search," Technical report, Stanford University, 2004.
[15] R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of People in a Social Network," Proc. Int'l World Wide WebConf. (WWW '05), pp. 463-470, 2005.
[16] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle,"Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD, 2009, pp. 119–128.
[17] Christo Ananth, M.Muthamil Jothi, A.Nancy, V.Manjula, R.Muthu Veni, S.Kavya, "Efficient message forwarding in MANETs", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 1,Issue 1, August 2015,pp:6-9
[18] C. K. Chui and B. Kao, "A decremental approach for mining frequent itemsets from uncertain data," in Proc. 12th Pacific-Asia Conf. Adv. Know. Discovery Data Mining, 2008, pp. 64–75.
[19] W. Dou, X. Wang, D. Skaw, W. Ribarsky, and M. X. Zhou, "Lead Line: Interactive visual analysis of text data through event identification and exploration," in Proc. IEEE Conf. Vis. Anal. Sci. Technol., 2012, pp. 93–102.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY