International Journal of Pure and Applied Mathematics
Volume 117 No. 21 2017, 37-41
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

An approach to cyber-terrorism detection in document streams using data-mining techniques.

Abijith Sasikumar[1]
abijithsasikumar16@gmail.com
Department of Computer Science & Engineering
Sathyabama University, Chennai

Adnan Koroth[2]
adnankoroth@gmail.com
Department of Computer Science & Engineering
Sathyabama University, Chennai

Mercy Paul Selvan[3]
mercypaulselvan@gmail.com
Department of Computer Science & Engineering
Sathyabama University, Chennai.

***Abstract*— Most of the existing work have topic modelling related methodologies and this does not consider individual topics and their relationships, whereas act of terrorism connected documents are on the net are always changing in its variety. So as to characterize cyber-crimes and to observe customized and abnormal activities of web users, this work introduces a new concept of STPs and find out an algorithm to mine URSTPs in document streams. For specific users this can be a rare on the entire however comparatively frequent, created to be applied in several real-life eventualities, this paper formulates a new of algorithms to make this mining with important applications possible. This include three different phases: pre-processing stage in which different topics are been found out and distinct session based on different users, STPs are created and URSTPS are found out by applying rarity analysis in the STPs found.**

***Index Terms*— Web mining, sequential patterns, document streams, rare events, pattern-growth, cyber-terrorism.**

I INTRODUCTION

Textual documents are occurring on a distributed form on the internet. The works that are existing are more concentrated towards topic modelling based approaches. In these approaches individual topics are given more importance and the sequential relation between different topics occurring on documents uploaded by user are not given much importance. By this work a new concept is formulated which can be helpful in finding out abnormal behavior of users. The new concept formulated in this paper is the URSTPs or the rare sequential patterns which is mined from the STPs. These rare STPs which are rare in global but while considering specific users these might be frequent in occurrence. This makes these rare patterns important in finding personalized behavior and can be applied to different other scenarios too. This paper proposes a three phased solution to find out abnormal behavior of users mainly related to cyber terrorism.

II LITERATURE SURVEY

Mining Sequential Patterns: Generalizations and Performance Improvements [1], we are given an outsized info of client transactions, wherever every dealings consists of customer-id, dealings time, and therefore the things bought within the dealings. We introduce the matter of mining serial patterns over such databases. They have used three algorithms to resolve this drawback, and through empirical observation judge their performance using synthetic information. two of the planned algorithms, AprioriSome and AprioriAll, have comparable performance, albeit AprioriSome performs somewhat higher once the minimum range of consumers that has got to support a serial pattern is low. Scale-up experiments show that each AprioriSome and AprioriAll scale linearly with the quantity of client transactions. They even have wonderful scale-up properties with regard to number of dealings per client and therefore the number of things during a transaction. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling [2], with the large quantity of digitized textual materials currently offered on the net, it's nearly not possible for individuals to extract all the data in a very timely manner. To alleviate the problem, in this paper they suggest a unique approach for extracting hot topics from disparate sets of textual documents uploaded within a given period of time. Our technique consists of two steps. First, the hot terms are been

extracted based on time. Second, on the basis of the extracted hot terms, key sentences are found out and so will be classified into clusters that represent hot topics by the use of multi-dimensional sentence vectors. The results of our empirical tests show that this approach is considerably simpler in characteristic hot topics than existing ways. The SPADE algorithm to mine frequent sequences [3], n this paper we introduces SPADE, a brand new algorithmic rule for quick discovery of sequential Patterns. The present solutions to the current problem create repeated database scans, and use complicated hash structures that have poor locality. SPADE utilizes combinatorial properties to decompose the initial problem into smaller sub-problems that may be independently solved in main-memory by the use of efficient lattice search techniques, and by using simple join operations. All sequences are discovered in just 3 database scans. Experiments show that SPADE outperforms the simplest previous algorithmic rule by an element of two, and by an order of magnitude with some pre-processed information. It additionally has linear scalability with reference to the quantity of input-sequences, and variety of different database parameters. Finally, we discuss how the results of sequence mining will be applied during a real application domain. Sequential pattern mining -- approaches and algorithms [4], the existing sequential pattern mining algorithms are been analyzed in this paper. It presents a classifying study of sequential pattern-mining algorithms into five intensive categories. First, on the idea of Apriori-based rule, second on Breadth first Search-based strategy, third on Depth first Search strategy, fourth on sequential closed-pattern rule and five on the idea of progressive pattern mining algorithms. At the end, a comparative analysis is finished on the idea of vital key options supported by varied algorithms. This study provides an improvement in the understanding of the approaches of sequential pattern mining. An Intelligent Analysis of internet Crime information using data mining [5], concerning national security has magnified considerably. A scenario is related to extract the attributes and relations within the websites and reconstruct the scenario for crime mining. By the use of clustering / classification based model to anticipate crime trends. The data mining techniques are used here to analyse the net data.

III EXISTING SYSTEM

Most of the existing systems are based on individual topic mining and the data lacks information about the correlation between topics. Many mining techniques are implemented for sequential topic mining which mines the frequent sequential patterns. The data obtained from these mining techniques lacks the information about rare sequential patterns which are significant in finding out abnormal and personalized user behavior. When it comes to the document streams most of the algorithm fails. Many of the introduced mining algorithms are designed for deterministic database and fail for document streams.

IV PROPOSED SYSTEM.

In the proposed system the documents are being extracted topic wise and the correlations between the topics are been noted, especially the sequential relation. The sequential relations mined so is saved as the sequential topic patterns. A powerful unsupervised algorithm is used to mine the rare sequential topic patterns which are more significant in finding out the personalized and abnormal behavior of users. This mining algorithm can be used effectively in a lot of applications.

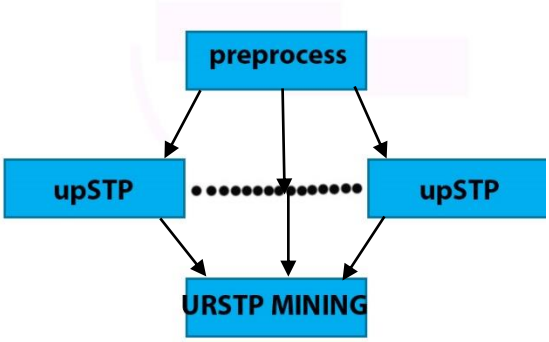

Fig. 1. Work flow

Text streams are taken as the input and a preprocessing stage is carried in which different topics are extracted and sessions are found out. From the data being extracted the STPs or the sequential topic patterns candidates are generated corresponding to each user. Now by the application of rarity analysis on the identified STPs to find out rare STPs or the URSTP.

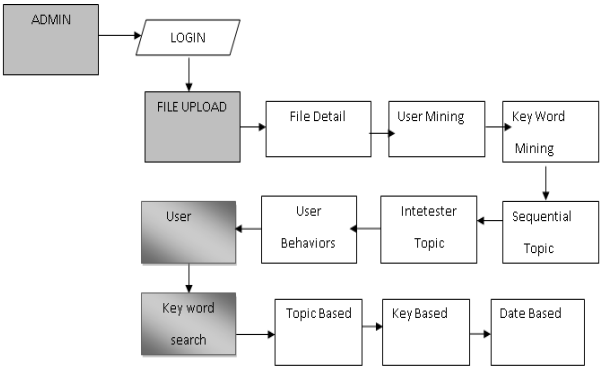

Fig. 2. Data flow diagram

1. EXPERIMENTAL RESULTS

The working of the system is on an uploaded document stream. Admin has given privileges to upload text files to the system and the users can search over this document stream for specific keyword. For example consider the keyword ‘’cloud’’ searched by a random user. The document stream has the keyword cloud in a textual document with text about cloud computing, the information about the file in which the keyword appears, the topic name given and the number of times the keyword is repeated will be provided as an output for the user by the system. In the case of the keyword “cloud”, the keyword has been repeated 35 times inside the document stream. On the same time the admin can monitor all the users searching for keywords through the document streams. The admin gets information about the keyword being searched by the user, how frequent the keyword is being searched etc. the admin can view the full history of all users and the information regarding their search. All the document related information will also be given with the, line count, character count and word count of the text inside the file.

File name	Line count	Word count	Character count
Cloudcomputing.txt	91	3528	23704
Bigdata.txt	64	2482	16832
Networking.txt	560	21700	128030
Social.txt	27	576	3769
Android.txt	50	458	3162

Table. 1. Textual document details

user	keyword	Keyword count
adnan	Of	1
abijith	Kill	1
magii	Overlay	3
pavithra	networking	60
rubini	cloudcomputing	19

Table. 2. User search monitoring

V CONCLUSION

The new proposed system rectifies the inability of the existing algorithms to work in text streams and to mine the significant URSTPs. The mining techniques can be used in a lot of applications like personalized and abnormal behavior of users. The system efficiently runs user-rarity analysis algorithm to mine URSTPs associated with users. With this the users are monitored for any abnormal behavior.

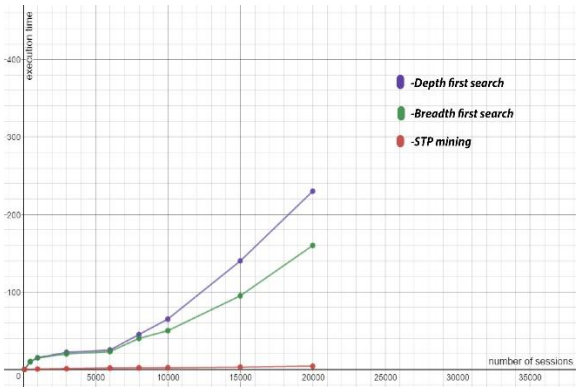

Fig. 3. Execution time variation of different algorithms with respect to increasing sessions.

VI FUTURE WORK

The mining of URSTP is an innovative approach, and has a lot of applications in real-time scenario like user behavior monitoring and document checking. The problem can be efficiently applied in lot of other scenarios like recommendation systems. The problem put forward a new direction of research in the field of web data mining. By improving the parallelism of thee algorithms used here, we can in future mine much more interesting patterns, with much more significance than the currently mined patterns.

VII REFERENCE

I. Jiaqi Zhu, *Member*, *IEEE*, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang, *Member*, *IEEE*, “Mining User-Aware Rare Sequential Topic Patterns in Document Streams”, IEEE Transactions on Knowledge and Data Engineering, 2016.

II. Kuan-Yu Chen, Luesak Luesukprasert, Seng-cho Timothy Chou, “Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling”, IEEE Transactions on Knowledge and Data Engineering, 2007.

III. Carl H Mooney, John F Roddick, “Sequential Pattern Mining: Approaches

and Algorithms", ACM Computing Surveys, 2013.

IV. R. Agrawal and R. Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements", IBM Almaden Research Center, 1995.

V. Mohammed j. zaki."SPADE: An Efficient Algorithm for Mining Frequent Sequences", Computer Science Department, Rensselaer Polytechnic Institute, 2001.

VI. Anshu Sharma, Shilpa Sharma , "An Intelligent Analysis of Web Crime Data Using Data Mining ", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012