# Mining User - Aware Rare Sequential Topic Pattern in Document Streams

A.Mary
*Assistant Professor, Department of Computer Science And Engineering*
*Alpha College Of Engineering, Thirumazhisai, Tamil Nadu, India*

A.Anbarasi
*UG Scholar, Department of Computer Science And Engineering*
*Alpha College Of Engineering, Thirumazhisai, Tamil Nadu, India*

G.Bhavana
*UG Scholar, Department of Computer Science And Engineering*
*Alpha College Of Engineering, Thirumazhisai, Tamil Nadu, India*

**Abstract- Textual documents created and distributed on the Internet are ever changing in various forms. Most of existing works are devoted to topic modeling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. In this paper, in order to characterize and detect personalized and abnormal behaviors of Internet users, we propose Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet. They are rare on the whole but relatively frequent for specific users, so can be applied in many real-life scenarios, such as real-time monitoring on abnormal user behaviors. We present a group of algorithms to solve this innovative mining problem through three phases: preprocessing to extract probabilistic topics and identify sessions for different users, generating all the STP candidates with (expected) support values for each user by pattern-growth, and selecting URSTPs by making user-aware rarity analysis on derived STPs. Experiments on both real (Twitter) and synthetic datasets show that our approach can indeed discover special users and interpretable URSTPs effectively and efficiently, which significantly reflect users' characteristics.**

**Keywords – Web mining, sequential patterns, document streams, rare events, pattern-growth, dynamic programming**

## I. INTRODUCTION

Document streams are created and distributed in various forms, such as news streams, emails, micro blog articles, chatting messages, research paper archives, web forum discussions, and so forth. In order to characterize user behaviors in published document streams, we study on the correlations among topics extracted from these documents, especially the sequential relations, and specify them as Sequential Topic Patterns(STPs).First, compared to individual topics, STPs capture both combinations and orders of topics, so can serve well as discriminative units of semantic association among documents in ambiguous situations. Second, compared to document- based patterns, topic based patterns contain abstract information of document contents and are thus beneficial in clustering similar documents and finding some regularities about Internet users. Third, the probabilistic description of topics helps to maintain and accumulate the uncertainty degree of individual topics, and can thereby reach high confidence level in pattern matching for uncertain data. For a document stream, some STPs may occur frequently and thus reflect common behaviors 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 specific user or some specific group of users. We call them as User-aware Rare STPs (URSTPs). Compared to frequent ones, discovering them is especially interesting and significant. Theoretically, it defines a new kind of patterns for rare event mining, which is able to characterize personalized and abnormal behaviors for special users. Practically, it can be applied in many real-life scenarios of user behavior analysis, as illustrated in the following example. It is worth noting that the ideas above are also applicable for another type of readers, so compared to statistical methods, mining URSTPs can better discover special interests and browsing habits of Internet users, and is thus capable to give effective and context-aware recommendation for them. To solve this innovative and significant problem of mining URSTPs in document streams, Many new technical challenges are raised. First, the input of the task is a textual stream, so existing

techniques of sequential pattern mining for probabilistic databases cannot be directly applied to solve this problem. Second, in view of the real-time requirements in many applications, both the accuracy and the efficiency of mining algorithms are important and should be taken into account, especially for the probability computation process. Third, different from frequent patterns, the user aware rare pattern concerned here is a new concept and a formal criterion must be well defined, so that it can be effectively characterize most of personalized and abnormal behaviors of Internet users, and can adapt to different application scenarios. To the best of our knowledge, this is the first work that gives formal definitions of STPs as well as their rarity measures, and puts forward the problem of

- Mining URSTPs in document streams, in order to characterize and detect personalized and abnormal behaviors of Internet users.
- We propose a framework to pragmatically solve this problem, and design corresponding algorithm to support it.
- We validate our approach by conducting experiments on both real and synthetic datasets.

## II. PROPOSED ALGORITHM

In our proposed system, Users rare and sequential activities can be monitored using sequence of document streams on multiple web application. We proposed our system to extract the user's activity on real time web application data set on Twitter and Gmail. Using our technique can monitor the user's sequential topic pattern based on their session identification on multiple applications with single sign on email id and their session id. We used the documents of inbox and send box mail of Gmail contents and twitter's tweet and individual chats to extract the topic and mining the user's activity. We extract the topic of document stream content using Stanford Natural Language Processing. Using this NLP processing and Monitoring dynamic 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, where Internet users behave as readers of documents instead of authors. In this case, STPs can characterize complete browsing behaviors of readers, so compared to statistical methods, mining URSTP scan better discover special interests and browsing habits of Internet users, and is thus capable to give effective and context-aware recommendation for them. While, this paper will concentrate on published document streams and leave the applications for recommendation to future work .To solve this innovative and significant problem of mining URSTPs in document streams, many new technical challenges are raised and will be tackled in this paper. First, the input of the task is a textual stream, so existing techniques of sequential pattern mining for probabilistic databases cannot be directly applied to solve this problem. A preprocessing phase is necessary and crucial to get an abstract and probabilistic descriptions of documents by topic extraction, and then to recognize complete and repeated activities of Internet users by session identification. Second, in view of the real-time requirements in many applications, both the accuracy and the efficiency of mining algorithms are important and should be taken into account, especially for the probability computation process. Third, different from frequent patterns, the user aware rare pattern concerned here is a new concept and a formal criterion must be well defined, so that it can effectively characterize most of personalized and abnormal behaviors of Internet users, and can adapt to different application scenarios. And correspondingly, unsupervised mining algorithms for this kind of rare patterns need to be designed in a manner different from existing frequent pattern mining algorithms.

- **User's registration and Creating dataset for user rare topics.**
- **NLP processing on Gmail and Twitter Contents.**
- **Monitoring user's activity using Gmail and Twitter dataset.**
- **Mining rare user sequential activities.**

*User's registration and creating dataset for user rare topics:*

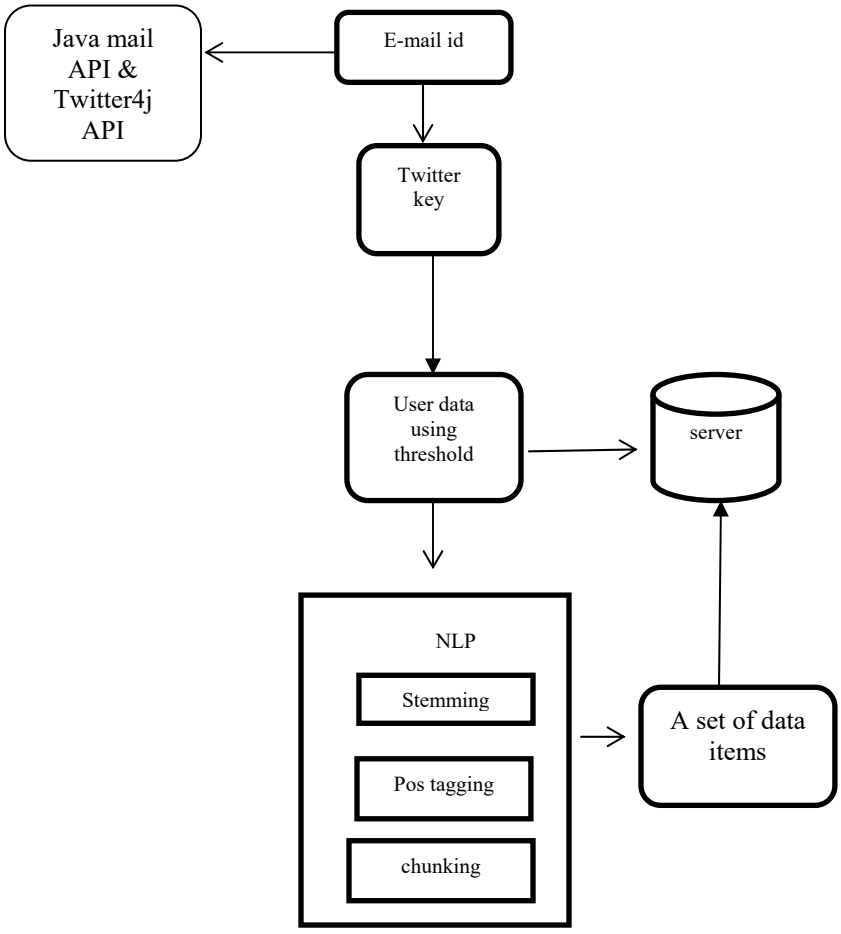
In this module the users have to register their email id and twitter key with our application. The email id and regarded twitter key's id should be a single sign on Gmail and Twitter account. Our application using users details make threshold for every users account by admin process. The data set of user's sequential topic extraction has to provide to the application. We build Stanford NLP algorithm to mining the user's activity. The data has been maintained and customized in the server. The user's details are stored in server database in the encrypted format because of the security purpose. To implement the effective rare topic extraction on sequential of document stream of the user's activity we used deserved data set of data mining process using Stanford NLP. In this API we implements POS tagging, chunking processing, stemming, spell checking and word net connection. We can feasibly extract the content of the user's rare topics using above mentioned NLP processing.

*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 data set we create threshold based data retrieving from the Social Medias content. Before proceeding to the content retrieving has been make sure of single sign on id for Twitter and Gmail. Using twitters 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, user's tweets, twitter chats and micro blogs maintained in our local server database. These social media contents are mined and extracted using Stanford NLP processing. The extracted topics of the user's contents are monitored in the server. POS tagging create 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, etc.

*Monitoring user's activity using Gmail and Twitter dataset*

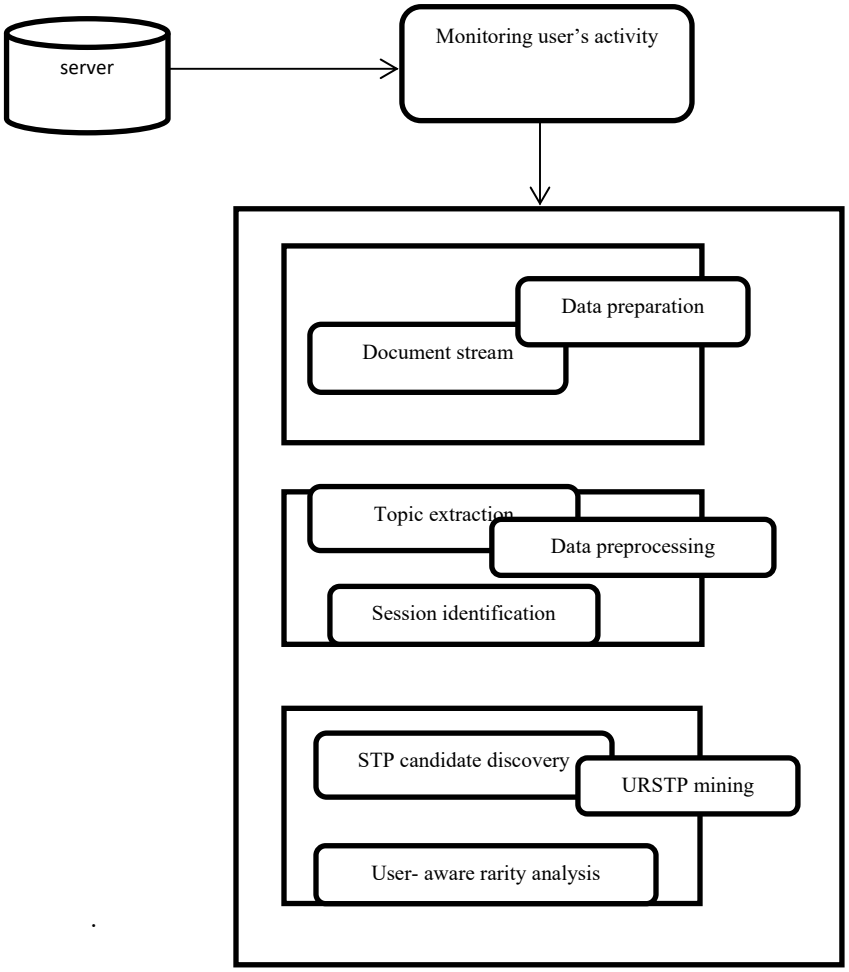
The Server monitors every user's activity on Gmail and Twitter. Single user activity on the two different web applications can be identified and extracted using single sign on email ids. The sequential topic extraction on sequence of documents are extracted and grouped. The evolution of individual topics, while sequential relations of topics in successive documents published by a specific user can be grasped by our application. For a document stream, some STPs may occur frequently and thus reflect common behaviors 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 STPs (URSTPs). Compared to frequent ones, discovering them is especially interesting and significant. Theoretically, it defines a new kind of patterns for rare event mining, which is able to characterize personalized and abnormal behaviors for special users. Practically, it can be applied in many real-life scenarios of user behavior analysis.

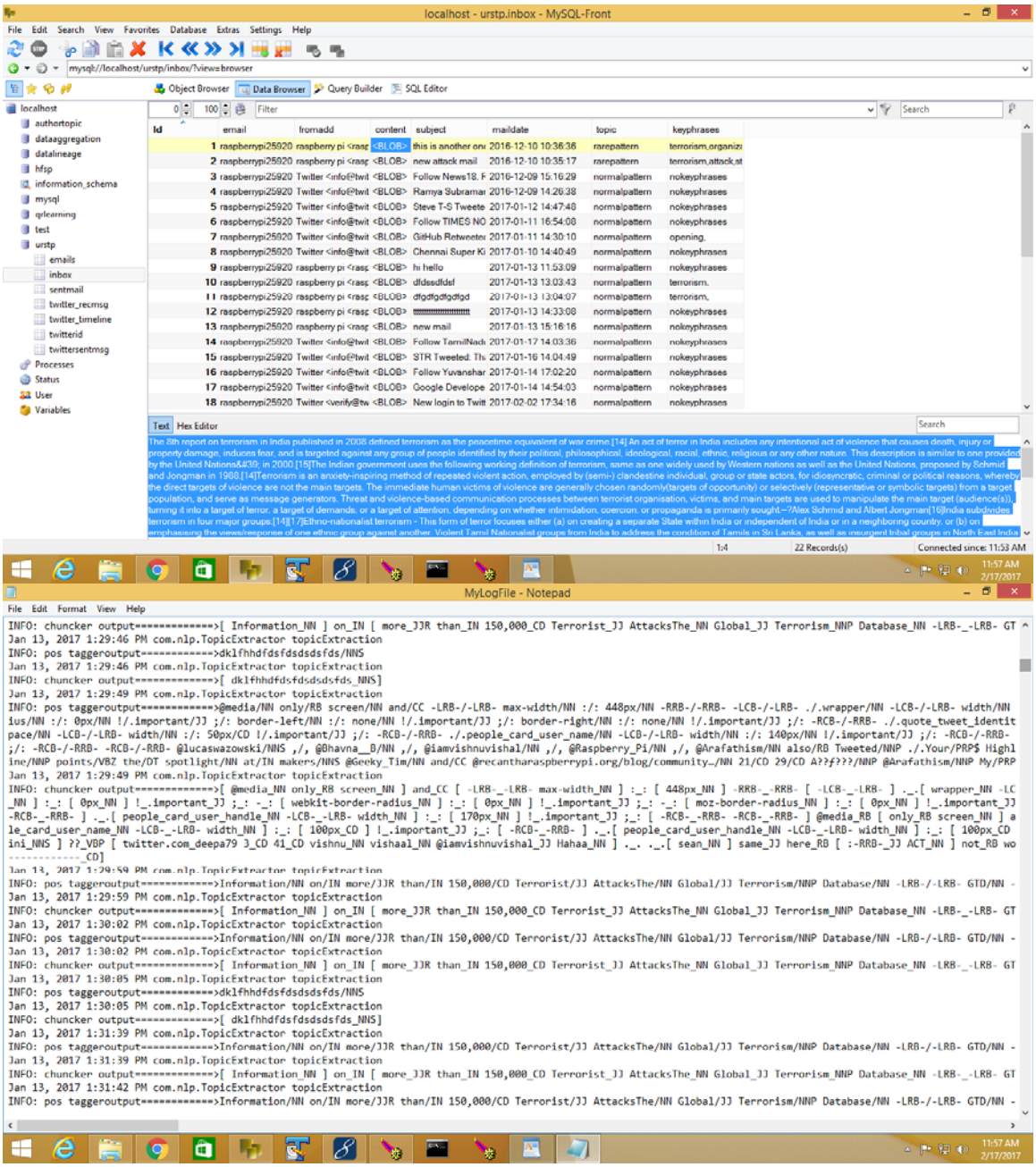
*Mining rare user sequential activities :*

While monitoring and extraction of the users sequential topics , if illegal behaviors are involved, detecting and monitoring them is particularly significant for social security surveillance. We can still expose them by URSTPs, 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 URSTPs is a good means for real-time user behavior monitoring on the Internet. The ideas above are also applicable for another type of document streams, called browsed document streams, where Internet users behave as readers of documents instead of authors. In this case, STPs can characterize complete browsing behaviors of readers, so compared to statistical methods, mining URSTPs can better discover special interests and browsing habits of Internet users, and is thus capable to give effective and context-aware recommendation for them. We implement the aware recommendation on admin dashboard. We highlight the rare user's activity and normal user's interest based on their social network.




.

## III. EXPERIMENT AND RESULT




## IV.CONCLUSION

Mining URSTPs in published document streams on the Internet is a significant and challenging problem. It formulates a new kind of complex event patterns based on document topics, and has wide potential application scenarios, such as real-time monitoring on abnormal behaviors of Internet users. In this paper, several new concepts and the mining problem are formally defined, and a group of algorithms are designed and combined to systematically solve this problem. The experiments conducted on both real (Twitter) and synthetic datasets demonstrate that the proposed approach is very effective and efficient in discovering special users as well as interesting and interpretable URSTPs from Internet document streams, which can well capture users' personalized and abnormal behaviors and characteristics. As this paper puts forward an innovative research direction on Web data mining, much work can be built on it in the future. At first, the problem and the approach can also be applied in other fields and scenarios. Especially for browsed document streams, we can regard readers of documents as

personalized users and make context-aware recommendation for them. Also, we will refine the measures of user-aware rarity to accommodate different requirements, improve the mining algorithms mainly on the degree of parallelism, and study on-the-fly algorithms aiming at real-time document streams. Moreover, based on STPs, we will try to define more complex event patterns, such as imposing timing constraints on sequential topics, and design corresponding efficient mining algorithms. We are also interested in the dual problem, i.e., discovering STPs occurring frequently on the whole, but relatively rare for specific users. What's more, we will develop some practical tools for real life tasks of user behavior analysis on the Internet.

## REFERENCES


[1] N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," ACM Comput.Surv., vol. 43, no. 1, pp. 3:1–3:41, 2010.

[2] A. K. McCallum. (2002). MALLET: A machine learning for language toolkit. [Online]. Available: http://mallet.cs.umass.edu

[3] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme Mining, 2002, pp. 418– 425.pattern mining on weblogs," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 533–542.

[4] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in Proc. ACM Int. Conf. Mach. Learn., 2007, pp. 633–640.

[5] C. H. Mooney and J. F. Roddick, "Sequential pattern mining approaches and algorithms," ACM Comput. Surv., vol. 45, no. 2, pp. 19:1–19:39, 2013.

[6] M. Muzammal, "Mining sequential patterns from probabilistic databases by pattern-growth," in Proc. BNCOD, 2011, pp. 118–127.

[7] M. Muzammal and R. Raman, "On probabilistic models for uncertain sequential pattern mining," in Proc. 6th Int. Conf. Adv. Data Mining Appl., 2010, pp. 60–72.

[8] M. Muzammal, "Mining sequential patterns from probabilistic databases," in Proc. 5th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2011, pp. 210–221.

[9] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining sequential patterns by prefixprojected growth," in Proc. IEEE Int. Conf. Data Eng., 2001, pp.215-224.

[10] M. Seno and G. Karypis, "SLPMiner: An algorithm for finding frequent sequential patterns using length-decreasing support constraint," in Proc. IEEE 11th Int. Conf. Data Mining, 2002, pp. 418– 425.