**ISSN: 0975-766X**
**CODEN: IJPTFI**
***Available Online through*** ***Research Article***
**www.ijptonline.com**

# MINING ABNORMAL- USER'S TOPIC PATTERN ON DOCUMENT STREAM FOR MULTIPLE SINGLE SIGN-ON APPLICATIONS

**Aiswarya Chandran, Mrs. S, Priya**
M. Tech Final year, CSE, Computer Science and Engineering, SRM University, Chennai.
Assistant Professor (O.G), Computer Science and Engineering, SRM University, Chennai.
*Email: aiswarya.chandran093@gmail.com*

*Received on: 15.10.2016* *Accepted on: 12.11.2016*

**Abstract**

Over a decade mining, sequential patterns have been a focused theme in data mining. Finding the behavior of a sequential pattern are helpful in finding many analyzing applications like predicting next event has been vital. Sequential Topic patterns are used to characterize and detect abnormal and personalized behaviors of Internet users and express the issue of mining User-mindful Rare Sequential Topic Patternson the Internet in document streams. Discussing the surprising value of features that have unexpected occurrence characteristics, and briefly, explore on- line adaptive filtering to handle evolving events in the news. These patterns are usually good on the whole but frequent for certain users, so this is applicable in real-life scenarios. A set of algorithms to elucidate this advanced mining problem is portrayed through three main phases: carrying out preprocessing to identify probabilistic topics and to find sessions for distinct users, developing all the STP users with support values for every user by pattern growth and selecting User- rare sequential topic patterns by involving user- aware rarity analysis on obtained STPs. Hypothesis on both real (Twitter) and Gmail datasets discover that the approach can discover unique users and interpret URSTP's effectively, which significantly reflect users characteristics. We also use Natural Language Processing for extracting the document stream to analyze the relative sequential topic pattern.

**Keywords**: - Sequential patterns, document streams, rare events, pattern growth, dynamic programming

## 1. Introduction

Discovering frequent subsequences as patterns in a database, is an important problem with broad applications, including understanding purchasing patterns of consumers or customers, the studying of sequencing or time- related processes such

as disease treatments, DNA Analysis, natural calamitiesetc which leads the way for sequential pattern mining. All these domains concentrate on some specific content or topics which give a detailed picture of a user's characteristic in real life. To process this data, a lot of analysis of text mining concentrated in extracting topics from documents through various probabilistic models, such as classical PLSI, LDA and extensions have been carried out. Most of the text mining research focused on finding topics in documentstreams. From the stream involving both semantic and transient data by different thememodeling methods, the topics can be extracted [1, 2, 4, 5]. Apparently, there may be some correlations among these obtained topics in successive documents for a specific user, and these correlations could be described by Sequential Topic Patterns (STPs). [3] Some STPs frequently occur in a document stream and thus reflect common behaviors of users. Besides, there are still some others which are rare for the all inclusive community yet happen generally regularly for a fewspecificuser or some specific group of users. Mining these user-related rare STPs in document streams[11] is more interesting, compared to frequent ones. Theoretically, it defines a new kind of patterns for event mining, which could characterize these individual and personalized behaviors in a certain context. Practically, it can be applied in many real-lifescenarios.

**2. Sequential Topic Patterns**

A document stream is defined as a sequence DS= ((d1; u1; t1); (d2; u2; t2); . . . ; (dN; uN; tN)), where di(i= 1. . . ..N) is a document published by user ui at time ti on a specific website, and ti<=tj for all i <=j. Usually, one user cannot write two documents simultaneously, for anyspecific user, at most one document is published. Formally, if ti!=tj, then ui!=uj always hold.

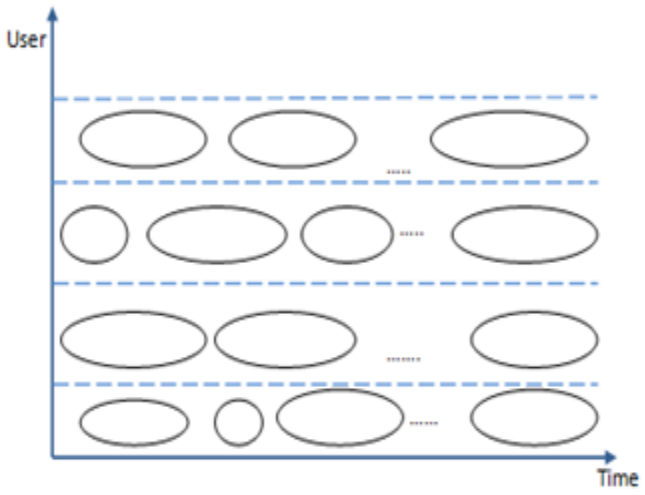


**Fig. 1.A sketch map of session identification.**

We pay attention to the correlations among successive documents published by the same user in a document stream. A fundamental but important correlations is the sequential relation among topics of these documents, which can be defined by sequential topic patterns, and abbreviated as STPs. When publishing documents to different websites they characterize users complete and personalized behavior.

A Sequential Topic Pattern (STP) a is defined as a topic sequence (z1; z2; . . . ; zn), where each ziET is a learnt topic. N=|a| denotes the number of topics contained in a, and is called the length of a. Pattern length n is called an n-STP.

**(Session)**:- For a specific user, there may exist multiple sessions in a document stream, which should be disjoint and consecutive.

A sketch map of session identification is shown in Fig. 1.Each ellipse represents a session, and all the sessions in each line constitute a document subsequence for a specificuser. When the session set of a topic-level document stream is obtained, we can find some concrete instances of an STP for each session.

**2.1. User-Aware Rare Sequential Topic Patterns**

They are globally rare for all sessions involving all users of a document stream; and locally and relatively frequent for the sessions associated with a specific user. Specifying this kind of STPs starts with the classical concept of "support" to describe the frequency.

The expected support is appropriate in measuring the frequency on uncertain sequences and can be computed by summing up the occurrence probabilities of alpha in all sequences. This is necessary because the session number here is nolonger a constant when we consider both the global frequency and the local frequency for different users. Usually, longer STPs tend to have lower support values.

Based on scaled support, we analyze these STPs in terms of their abilities in characterizing personalized and abnormal behaviors of Internet users and pick out some significant ones. First, each selected STP should be linked to a specific user, so it can be called a user aware STP. Second, it reflects the particularity on frequency, not only at user level but also at pattern level.

According to these ideas, we define two new measures, absolute rarity, and relative rarity. These regularize the scaled support of an STP for a user over other users and other STPs respectively and will be computed successively.

**3. Mining URSTP**

A novel approach to mining URSTPs in document streams. The main goal is to discover all the STP candidates in the document stream for all users and pick out important URSTPs related to particular users by user-aware rarity analysis. Designing a cluster of algorithms to unify the notations and many variables are denoted and stored in the key-value form is also done. Some more thresholds are used in preprocessing procedures, but preprocessing strategies have to be chosen with some common rules according to the characteristics of the input stream. Therefore we assume preprocessing as a separate and independent module, and thus do not regard the thresholds defined there as the input parameters of the whole mining problem.

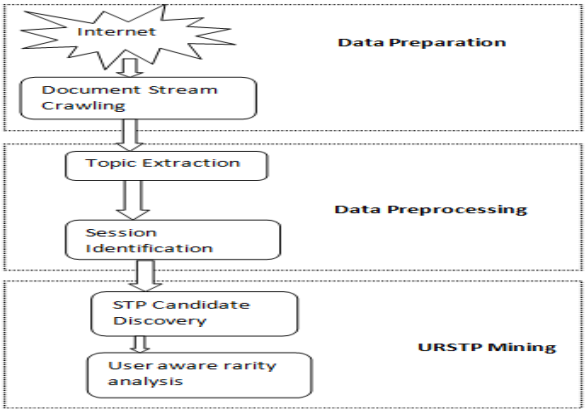


**Fig .2.Theprocessing framework of URSTP mining.**

After preprocessingas shown in Fig. 2, we obtain a set of user-session pairs. For each of them with a specific user, a new thread is invoked. These threads are executed in parallel relying on the hardware environment.

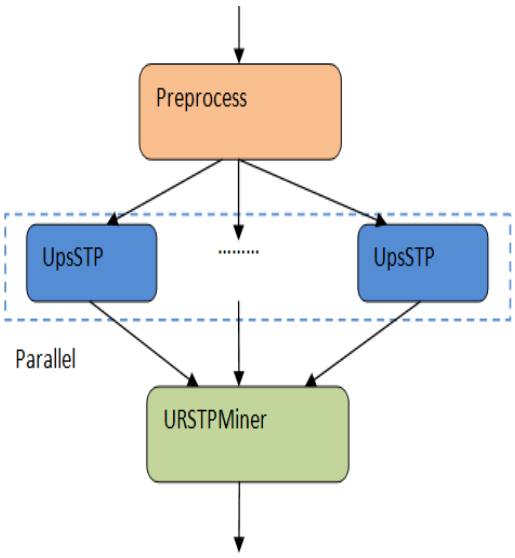


**Fig.3. Workflow of URSTP mining.**

When all of these finish, another subprocedure URSTPMiner calls to prepare the user-aware rarity analysis for these STPs together and get the output set User URSTP, which has all the pairs of users and their respective URSTPs as shown in Fig. 3 with values of relative rarity.

Many preprocessing procedures and mining algorithms are available in detail. Some of them are :

**Data Preprocessing**: This paves the way for topic extraction and session identification.

1) **STP Candidate Discovery by Pattern-Growth**: This mining algorithm renders the support values with the help of DP- Based algorithm and Approximation algorithm.
2) **User- Aware Rarity Analysis:** This will make the user-aware rarity analysis to pick out URSTPs, which gives personalized, abnormal and thus significant behaviors.

**4. Experiments**

The problem of mining URSTPs in document streams proposed is innovative, but the effectiveness of our approach in discovering personalized and abnormal behaviors, especially the reasonability of the URSTP definition, needs to be practically validated. Conducting interesting and informative experiments on message streams in Twitter datasets, shows that most of the users approach are actually special in real life.

We also evaluate the efficiency of the approach on synthetic datasets, and compare the two alternative subprocedures of STP candidate. Many users also want to discover useful data outside their own streams, such that interesting URL son Twitter posted by friends of friends, or relevant blogs in Google Reader that are subscribed by other friends and it also enables the users to send and read short text messages. We also study the social relationships in the user network in order to quantify the importance of each analyzed content.

In addition, we use the approximation algorithm Ups STP-a to replace Up sSTP, and carry out the two steps of mining for comparison. Applying URSTP Miner on STPs to mine user-aware rare ones to find special and abnormal behaviors of Internet users, which are intuitively in minority for the general population, so the effectiveness of our approach should be reflected by the quality of those URSTPs with topmost values of the relative rarity, as well as their associated users.

**4.1. Quality of Related users**

It is very difficult to get the exact ground truth of these users for the randomly crawled datasets. Here, making a reasonable assumption that is “verified” users in Twitter are more likely to have special and repeated behaviors than

ordinary users, so theyare regarded as approximate ground truth of special users. Twitter is more popular because of its wide spreading of instant messages (i.e., tweets), bursts of world news, entertainment gossip about celebrities, and discussions over recently released products are all spreading on Twitter massively. Text content is one of the most important elements of social networks.

Characterization of these contents in documents is a standard problemin addressing information retrieval and statistical natural language processing. In Twitter, hashtags, prefixing one or more characters with a hash symbol as "#hashtag", are a community-driven convention for adding additional context and metadata to tweets, making tweets semi-structured texts.

The weakly-checked information given by hashtags can build direct semantic relations between tweets so that the words in tweets have more complex topical relationships than in normal texts.

**The related works involve:-**

Topic models on flat text

Topic models on semi- structured text

Hashtag topic distribution is an important by-product of HGTM (Hash-tag graph based topic model). The goal is to get the capacity of HGTM to distinguish hashtags with different semantic domains. Hence, tweets also play a vital role in mining the abnormal users in a topic pattern[12].

**5. Extension of Sequential Pattern Mining**

Intensive studies are carried out during recent years; as there exists a great diversity of algorithms for sequential pattern mining. Motivated by the potential applications for the sequential patterns, numerous extensions on initial definition have been proposed which could be related to other types of time- related patterns or to the addition of time restrictions. Some extensions of these algorithms for special purposes such as multidimensional, closed, time interval, and constraint based sequential pattern mining are discussed in following section.

**5.1. Multidimensional Sequential Pattern Mining:**

Considering one attribute along with time stamps in pattern discovery process, while mining sequential patterns with multiple dimensions are considered as multiple attributes at the same time. Mining multiple dimensional sequential patterns was introduced by Helen Pinto and Jiawei Han [6] which can help render more informative and useful patterns.

For example to get a traditional sequential pattern from the supermarket database that after buying product "a" most people also buy product "b" in a defined time interval. Therefore, using multiple dimensional sequential pattern mining we can further find various groups of people have different purchase patterns.

Lets take another example of M.E. students as they always buy product "b" after they buy product "a", while the sequential rule weakens other groups of students. Hence, we can see that multi- dimensional sequential pattern mining provides more precise and accurate information for further decision support.

**5.2. Discovering Time-interval Sequential Pattern:**

Eventually sequential topic patterns tells us what items are frequently clubbed together and in what systematic order they are aligned. Hence, they cannot provide information about the time span between items for further decision support. In other words, although we know what items will be bought after the preceding items, wedo not have any idea as to when the next purchase will happen.

Y. L. Chen, M. C. Chiang, and M. T. Kao [7] has come up with a solution for this problem which is to generalize the mining problem into discovering time-interval sequential patterns, that tells not only the sequence of items but also the time intervals between successive items.

A discontinuity constraint imposes a threshold on the separation of two consecutive elements of an identified sequence. This type of constraints is crucial for the applicability of these methods to numerousproblems, especially those with long sequence.

**5.3. Closed Sequential Pattern Mining:**

The sequential pattern mining algorithms evolved so far have good performance in databases which consists of short or small frequent sequences. But in an unsuitable manner, when mining long frequent sequences, or when using very low support thresholds, the performance of such algorithms often scales down dramatically. An alternative but a similar powerful solution instead of mining the complete set of frequent subsequence, we mine frequently closed subsequence only, i.e., those that do not contain any super-sequence with the same support. This mining technique will come up with a significantly less number of discovered sequences than the traditional methods while safeguarding the same expressive power since the whole set of frequent subsequences together with their supports, will be derived easily from the mining results [8].

**5.4. Discovering Constraint Based Sequential Pattern:**

However efficiency of mining the entire set of sequential patterns has been improved substantially, but in many cases sequential pattern mining still encounters or will face tough challenges in both effectiveness and efficiency. Similarly, there will be a huge amount of sequential patterns in a large database. A user will be often interested in only a small subset of such patterns.

Showcasing the complete set of sequential patterns will make the mining result difficult to understand and hard to use. To overcome this problem Jian Pei, Jiawei Han and Wei Wang [9] have presented the problem systematically by pushing various constraints deep into sequential pattern mining using pattern growth methods.

**6. Future Direction:**

This area will be focused on future research by improving the efficiency of the algorithms either with new structures andnew representations or by managing the database in the main memory. So inaccordance to these criteria's sequential pattern mining is scrutinized into two major groups such as Apriori Based and Pattern Growth based algorithms. By comparative analysis of various mining algorithms, it is clear that PrefixSpan Algorithm is more efficient with respect to running time, space utilization and scalability and it can be more efficient if we use DISC (Direct Sequent Comparison) Strategy [10] with PrefixSpan Algorithm in the pruning step, says we can eliminatenonfrequent sequences according to the other sequences with the same length. But eventually there are various research challenges in this field of data mining.

**7. Conclusion**

Mining URSTPs in published document streams on Internet is an important 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.

The experiments on 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 UR STPs from Internet document streams, which can well capture users' personalized and abnormal behaviors and characteristics.

We can regard readers of documents as personalized usersand make context-awarere commendation for them. Developing some practical tools for real life tasks of user behavior analysis on the Internet is one way of handling this

scenario. Widespread content analysis and social media mining has become an important task within tweets for uncovering topics. Users have provided hashtags as a powerful and valuable data source in the vast amount of tweets on the web.

One effective alternative of utilizing user-contributed hashtags for tweet topic modeling is to handle both sparseness and noise in tweets. This approach is highly scalable and can be used in a number of real-world applications, such as hashtag recommendation, short text retrieval, and event detection.

**References:**


1. D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.
2. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 113–120.
3. 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.
4. A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 497–504.
5. 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.
6. Helen Pinto Jiawei Han Jian Pei Ke Wang, ―Multidimensional Sequential Pattern Mining, In Proc. 2001 Int. Conf. Information and Knowledge Management (CIKM'01), Atlanta, GA, Nov. 2001 pp. 81–88.
7. Chen, Y.L., Chiang, M.C. and Kao, M.T, ―Discovering time interval sequential patterns in sequence databases, Expert Syst. Appl., Vol. 25, No. 3, 2003, pp. 343–354.
8. Yan, X., Han, J., and Afshar, R., ―CloSpan: Mining closed sequential patterns in large datasets, In Third SIAM International Conference on Data Mining (SDM), San Fransico, CA, 2003, pp. 166–177.
9. Jian Pei, Jiawei Han, Wei Wang, ―Constraint-based sequential pattern mining: the pattern growth methods, J IntellInfSyst, Vol. 28, No.2, 2007, pp. 133 –160.
10. Ding-Ying Chiu, Yi-Hung Wu, Arbee L.P. Chen "An Efficient Algorithm for Mining Frequent Sequences by a New Strategy without Support Counting" Data engineer, 2004, proc. of 20 International conference, pp.375-386, 2004.

11. 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".
12. Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng "Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs".