International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

# Mining Users Rare Sequential Topic Patterns from Tweets based on Topic Extraction

**Bhakti Patil[1], Sachin Takmare[2], Rahul Mirajkar[3], Pramod Kharade[4]**

[1]*Student, Dept. of Computer Science & Engineering, Bharati Vidyapeeth's College of Engg, Kolhapur, Maharashtra, India.*
[2,3,4] *Professor, Dept. of Computer Science & Engineering, Bharati Vidyapeeth's College of Engg, Kolhapur, Maharashtra, India.*

---

**Abstract -** *Twitter is an online news and social networking service where users post and interact with messages, "tweets" spontaneously. Most of the existing works are dedicated to discovering the abstract "topics" that occur in a collection of documents and creation of discrete topic. It means when a specific user publishes successive documents then successive relation between topics is totally ignored. In this paper, a different approach for detecting users' Sequential Topic Patterns is proposed which consequentially characterizes and detects personalized and abnormal behaviors of users and then we prepare the problem of Mining Users Rare Sequential Topic Patterns(URSTP) from Tweets. URSTPs are rare for all users but relatively frequent for some specific users, so this approach 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 such innovative mining problem using different phases such as preprocessing to extract probabilistic topics, identifying sessions for different users, generating all the STP candidates and selecting URSTPs by making user-aware rarity analysis on derived STPs. Experiment show that our approach can significant to find special users and interpretable URSTPs, which significantly indicate users' characteristics.*

***Key Words***: **Sequential topics, Web mining, Topic Extraction, Keyword Extraction, frequent patterns, clustering.**

## 1. INTRODUCTION

Social networking service such as facebook, Twitter, LinkedIn creates an environment where user could spend a lot of time on it and use it for different purposes. Based on this interaction between users, we have a huge amount of data for each individual user. Documents of such services focus on some particular topic. Topic provides users characteristics. Text mining is one and only way to mine the piece of information for extracting topics. Generally some probabilistic topic models such as LDA [1], classical PLSI[5] and their extensions[3],[4],[6],[7],[8],[9] are used for topic extraction.

In the literature most of the researchers concentrates on adaptation of single topic to identify and imagine social events and user behaviors [10], [11], [12]. Some researchers studied relation between the different topics of successive documents published by same user successively where some hidden but important information behaviors has been neglected which uncovers personalized behaviors of that user.

In this paper we mainly concentrates on relation mainly between the extracted sequential topics refer them as Sequential Topic Patterns (STP) that indirectly reflects user behaviors. For a document stream some STPs may occur frequently and so it reflects common behaviors of involved users. But away from that, there may still exists some other patterns which are infrequent for the general population, but occur relatively frequent for some specific user or some specific group of users. We refer them User-aware Rare STPs (URSTPs). Compared to frequent patterns, discovering rare patterns is interesting and important. Basically, it formulates a new problem for rare event mining, so that it is possible to characterize personalized and abnormal behaviors for special users' behavior.

In our case STPs can characterize complete browsing behaviors of readers. Then compared with statistical methods, mining URSTPs can better to find special interests and browsing habits of users, and is thus capable to give effective and context-aware recommendation for them. Our approach will concentrate on published document streams.

Solving such important problem of mining URSTPs in document streams, new technical provocations are raised and will be solved in this paper. First, the input of the approach is a text stream, so existing techniques of probabilistic databases cannot be directly applied to solve this problem. A preprocessing phase is required and important thing to get conceptual and probabilistic descriptions of documents by topic extraction, and then to identify complete and repeated liveliness of users by session identification. Second, in case of the real-time requirements in many applications, both the precision and the effectiveness of mining algorithms are important, especially for the probability computation process. Third, unlike from frequent patterns, the user aware rare pattern can effectively characterize most of personalized and abnormal behaviors of users and can applied 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.

## 2. LITERATURE REVIEW

Topic mining in document collections has been extensively studied in the literature. Topic Detection and Tracking (TDT) task covers detection and tracking of topics (events) in news based on keywords. A lot of probabilistic generative models for extracting topics from documents were also proposed, such as LDA [1], PLSI and their extension[2] also models for short texts like Twitter-LDA [3].

LDA is a three-level hierarchical Bayesian model .Each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is modeled as an infinite mixture. Instead of text modeling, the topic probabilities provide an explicit representation of a document. Blei, Ng, and Jordan have presented an efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation.

Li and McCallum proposed pachinko allocation model (PAM) that captures arbitrary and sparse correlations between topics using a directed acyclic graph (DAG) which is not possible in case of LDA. The child node of the DAG shows individual words in the vocabulary, while each interior node represents a correlation among its children, which may be words or other interior nodes (topics).

Zhao, Jiang, Weng, He, Lim, Yan, and Li have compared the tweets with a traditional news medium by using unsupervised topic modeling. They discover topics through Twitter-LDA model then they compare Twitter topics with topics of news using text mining techniques. Also they concentrate on relation between tweets and retweets.

In case if real application, the content of document collection is temporal so various dynamic topic models have been proposed to discover topics over time in document streams and then offline social events are predicated. One of that is dynamic topic model proposed by Blei and Lafferty in which it uses state space to represent the topics. Approximate posterior inference over the latent topics carried out by variational approximations based on Kalman filters and nonparametric wavelet regression. Dynamic topic models provide a qualitative perspective into the contents of a large document collection.

The important problem in data mining is mining Sequential pattern. The frequency of a sequential pattern is evaluated by using support. The mining algorithms like PrefixSpan, FreeSpan, SPADE have been proposed based on support. These algorithms find outs frequent sequential patterns with support values are not less than a user-defined threshold, and then used by SLPMiner to deal with length-decreasing support constraints. Muzammal et al. concentrated on sequence-level uncertainty in sequential databases, and proposed methods to calculate the frequency of a sequential pattern based on expected support, using generate-and-test or pattern-growth. This paper is an extension of our previous work.

## 3. PRELIMINARIES

At first, we define some basic concepts in a usual way.

**Definition 1 (Document)**

A text document d in a document collection D consists of a many number of words from a fixed vocabulary V = {$w_1, w_2,$ ....., $w_{|v|}$}. Document can be represented d = {c (d, w)} where w∈v, c denotes the occurrence number of the word w in d.

**Definition 2 (Topic)**

A topic z in the text collection D is represented by a probabilistic distribution of words in the given vocabulary V.

**Definition 3 (Topic-Level Document)**

Given an original document d∈D and a topic set T, the corresponding topic-level document $td^d$ is defined as a set of topic-probability pairs, in the form of{(z,p(z|d))} where z∈T.

**Definition 4 (Document Stream)**
A document stream is defined as a set that consists of sequence of document number, a document published by user ui at time ti on a specific website, and ti ≤tj for all i ≤ j.

**Definition 5 (Sequential Topic Pattern)**

A Sequential Topic Pattern (STP) $\alpha$ is defined as a topic sequence of topics i.e. [$z_1, z_2$, ....., $z_n$] where topic z $\epsilon$ T.

**Definition 6 (Session)**

A session s is defined as a subsequence of topic level document stream associated with the same user, i.e. it is a set of topic level document with its associated time for different user.

**Definition 8 (Support of STP)**

It is defined as a probability of topics with respect to sessions.

**Definition 9(User-Aware Rare STP)**

An STP a is called a User-aware Rare STP (URSTP) if and only if both scaled support is less than or equal to scaled support threshold and relative rarity greater than or equal to relative rarity threshold hold for some user u.

## 4. MINING USERS RARE SEQUENTIAL TOPIC PATTERNS

At first we have list of users' tweets collected using some API. The one that we used was... Topic detection from the whole number of documents needs some pre-processing initially. At the first step, we will remove stop words and repeated posts. Stop words are such as ["at"," the"," how" etc.]. We now have list of cleaned twitter posts or we have list of cleaned documents. Tweets = [list of tweets of all users]. Each tweet has a list of posts. Therefore, for each user/tweet we are removing repeated words and stop words. This new list of tweets will be the input of keyword extraction algorithm to extract keywords. Fig.-1 shows different topic extraction methods.

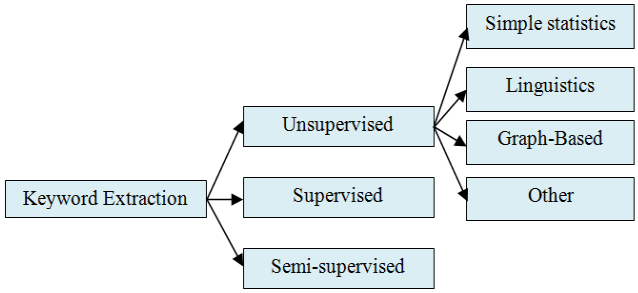


**Fig -1**: Keyword Extraction Methods

After that with cosine similarity, we cluster keywords. It means that we cluster posts, which are similar to each other. This approach is trained to work as unsupervised topic detection. Now, new tweets will be the input to our next step topic extraction. So, the output will be the index of the lookup table, which gives the list of topics as shown in Fig –2.

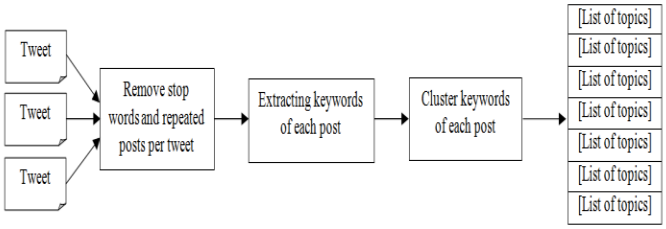


**Fig -2**: Overview of Keyword Extraction Process

Naïve Bayes classifiers which is a simple probabilistic classifiers dependent on using Bayes' theorem with strong (naive) independence assumptions between the features. We use such a baseline method for text categorization. This popular method solves problem of judging documents as belonging to one category or the other such as sports or politics, etc. with word frequencies. Naive Bayes classifiers requires a number of parameters linear in the number of variables i.e. features/predictors.

Naive Bayes model assigns class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All naive Bayes classifiers suppose that the value of a particular feature is independent of the value of any other feature, given the class variable. A naive Bayes classifier considers each of these features to contribute independently to the probability, regardless of any possible correlations .It works in case of a small number of training data and estimates the parameters required for classification. Using Bayesian probability terminology,

Posterior = (likelihood * prior) / evidence

We use Apriori algorithm to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation. Each transaction is a set of items (an itemset). Given a threshold C, the algorithm identifies the item sets which are subsets of at least C transactions in the database.

In this method frequent subsets are carried out one item at a time and groups of candidates are trying out against the data. Candidate item sets are counted by using breadth-first search and a Hash tree structure. It generates candidate item sets of length K from item sets of length k-1. Then it minimizes the candidates which have an infrequent sub pattern.

Now we propose an approach to mining rare sequential topic patterns in document streams. The main processing framework is shown in Fig.-3.

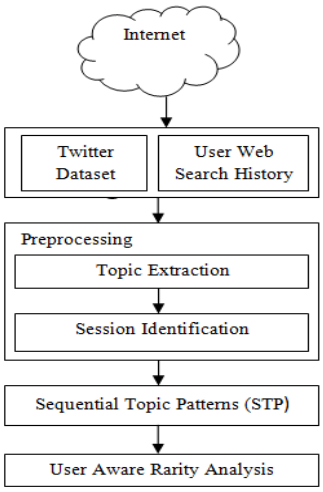


**Fig -3**: Processing framework of URSTP mining

It consists of three main phases. At first, text documents are collected from some micro-blog sites or forums (in our case we crawled tweets from Twitter using Twitter API ), and use a document stream as the input of our approach. Then, in preprocessing phase, we first remove useless symbols such as "@", "#", URL in the input tweet and stop words. Then original stream is transformed to a topic level document stream and then divided into many sessions to identify complete user behaviors. Finally and most importantly, we find out all the STP candidates in the document stream for all

users, and finally pick out important URSTPs associated to specific users by user-aware rarity analysis.

## 5. RESULT ANALYSIS

In this section, we apply our mining URSTP mechanism to find out rare topics. In order to simulate the proposed architecture, we implemented approach by using minimum 1GB RAM and 60GB (or above) hard disk. The results are carried out with different file size. As the file size increases, the time required for execution increases in both naïve Bayes algorithm and Apriori algorithm. But as compare to Apriori Algorithm, Naïve Bayes requires less time. The results are shown in Fig. 1 and 2.

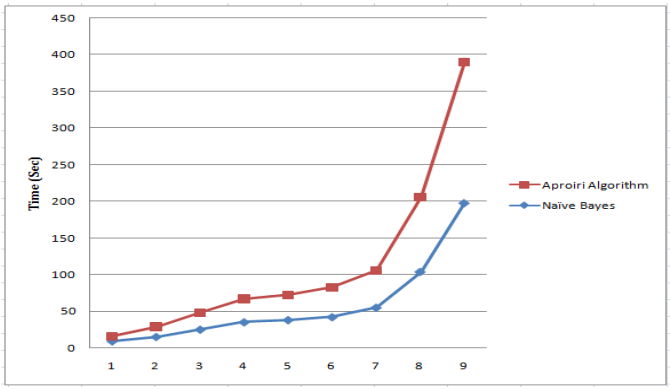


**Chart -1**: Time costs of the two algorithms

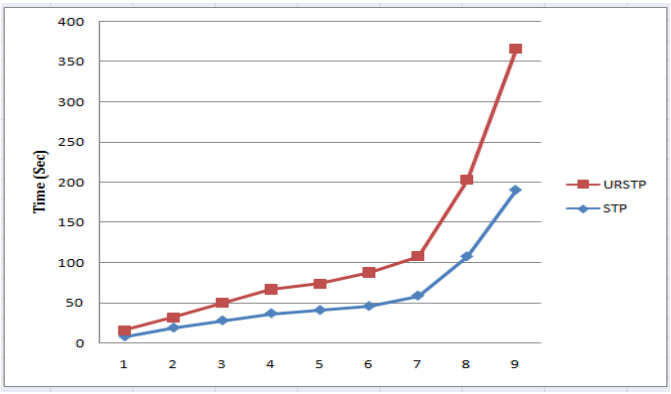


**Chart -2**: Performance Time Difference

Fig. 1 we compare, the time required to execute Naïve Bayes Algorithm and Apriori Algorithm of file size ranging from 1KB to 1000KB.

In Fig. 2 we compare, the time required for Sequential Topic Pattern and User Aware Rare Sequential Topic Patterns to find out the same from 1 KB to 1000 KB file size.

## 6. CONCLUSIONS

Twitter Tweets capability of social networking sites is really high. In order to tackle this ability of social networking sites, we propose some new methods. At first we extract users' posts through API then we extract appropriate topics depending on certain keywords. It is then shown that by creating clusters based on keywords which are helpful in easier detection of topics from users' tweets. These extracted topics are then advantageous to real-time monitoring on abnormal behaviors of users. Also we proposed an effective approach to discover special users and interesting URSTPs from document streams i.e. user tweet, which captures users' personalized and abnormal behaviors and characteristics of users'. We are interested in the dual problem, i.e., discovering sequential topic patterns occurring frequently on the whole, but relatively rare for specific users.

## REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[2] 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.

[3] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in Proc. 33rd Eur. Conf. Adv. Inf. Retrieval, 2011, pp. 338–349.

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 113–120.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999, pp. 50–57.

[6] D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.

[7] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. 1st Workshop Soc. Media Anal., 2010, pp. 80–88.

[8] 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.

[9] A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 497–504.

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

[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in Proc. IEEE Conf. Vis. Anal. Sci. Technol., 2012, pp. 93–102.

[12] 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.