NATIONAL CONFERENCE ON EMERGING TRENDS IN ENGINEERING AND TECHNOLOGY.
NCETET-2017, 7st March 2017, BVCOE, Kolhapur.
ISSN: 2231-5063.

A SURVEY ON RARE SEQUENTIAL TOPIC PATTERNS

Ms. Bhakti G. Patil, Mr. Sachin B. Takmare

P.G. Student, Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, Kolhapur, India.

Assistant Professor, Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, Kolhapur, India.

Abstract: Many text documents are created and distributed in many different formats. Most of the work is dedicated to topic modeling and creation of individual topics where the sequential relation between the topic in sequential documents published by same user is ignored. In this paper, we are tried to characterize and detect personalized and abnormal behaviors of users, for that first we propose Sequential Topic Patterns (STPs) and then formulate the problem of mining User-aware Rare Sequential Topic Patterns(URSTPs) in documents. Patterns are rare on the whole but relatively frequent for specific users so can applied for real-time monitoring on abnormal user behaviors. We are using a group of algorithms to solve such an innovative mining problem using three phases: preprocessing to extract topics and identify sessions for different users, generating all the STP candidates with (expected) support values for each user by pattern-growth, and choosing URSTPs by making user-aware rarity analysis on derived STPs. Experiments on both real (Twitter) and synthetic datasets can be performed to discover special users and interpretable URSTPs effectively and efficiently, which reflect users' characteristics.

Keywords: pattern growth, dynamic programming, rare patterns, sequential patterns.

I. INTRODUCTION

Textual Document streams are created and scattered such streams may include news, emails, chatting messages etc. Content of such streams specifies some particular topic which reflect events and users characteristics. To mine such streams, text mining researches focused on extracting topics from document collections and document streams using different probabilistic topic models, such as PLSI [1], LDA [2] .

These extracted topics in document Streams are used for the creation of individual topics to detect and predict social events and user behaviors [3], [4]. Some researches concentrates on the correlations between different topics created successively and documents are published by a same user.

For the characterization of user behaviors in published document streams, we study on the correlations between extracted topics from these documents, mainly the sequential relations, and refer them as Sequential Topic Patterns STPs). Each document stream gives the complete and repeated behavior of a user when that user is publishing a series of documents, then it is suitable to deduce users characteristics and psychological statuses.

For a document stream, some STPs may occur frequently and reflects common behaviors of users. Away from that, there may still exist some globally rare patterns for the general population, but occur for some specific user or some specific group of users. We refer them as User-aware Rare STPs (URSTPs). Compared to frequent ones, discovering them is especially interesting and significant. Practically, it can be applied in many real-life scenarios of user behavior analysis.

The innovative and significant problem of mining URSTPs in document streams, we try to follow some steps. First, the input of the task is a textual stream. Then a preprocessing phase is necessary and crucial to get abstract and probabilistic descriptions of documents by topic extraction, and then to recognize complete and repeated activities of users by session identification. Second, in case of real-time applications, both the accuracy and the efficiency of mining algorithms are important and should be taken into account. Third, the user aware rare pattern can effectively characterize most of personalized and abnormal behaviors of users.

Sequential pattern mining is an important problem in data mining. The concept of support is the most popular measure for evaluating the frequency of a sequential pattern, and is defined as the number or proportion of data sequences containing the pattern in the database. Many mining algorithms have been proposed based on support, such as PrefixSpan [6], FreeSpan [7] and SPADE [8]. They discovered frequent sequential patterns whose support values

are greater than a user-defined threshold, and were extended by SLPMiner [5] to deal with length-decreasing support constraints. The obtained patterns are not always required for our purpose, because those rare but significant patterns representing personalized and abnormal behaviors are pruned due to low supports. Furthermore, the algorithms on deterministic databases are not applicable for document streams, as they failed to handle the uncertainty in topics.

II. LITERATURE SURVEY

Blei, Ng, Jordan have proposed a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is considered as a finite mixture over an underlying set of topics. Every topic is in turn considered as an infinite mixture over an underlying set of topic probabilities. Text modeling point of view, the topic probabilities specify an explicit representation of a document. They present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation and they report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model [2].

Blei, Lafferty has proposed approach is to use state space models on the natural parameters of the multinomial distributions that represent the topics. Variational approximations based on Kalman filters and nonparametric wavelet regressions are developed to carry out approximate posterior inference over the latent topics. In addition to giving quantitative, predictive models of a sequential corpus, dynamic topic models provide a qualitative window into the contents of a large document collection. The models are demonstrated by analyzing the OCR'ed archives of the journal Science from 1880 through 2000 [9].

PrefixSpan explores prefix projection in sequential pattern mining. PrefixSpan mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. Moreover, prefix-projection substantially reduces the size of projected databases and leads to efficient processing. PrefixSpan outperforms both The Apriori-based GSP algorithm and another recently proposed method, FreeSpan, in mining large sequence Databases [6].

FreeSpan integrate mining of frequent sequences with that of frequent patterns and use projected sequence database to confine the search and the growth of subsequent fragments. FreeSpan mines the complete set of but reduce the efforts of candidates' subsequence generation. It examines substantially small numbers combinations of subsequences and runs faster than Apriori-base GSP algorithm [7].

SPADE algorithm for fast discovery of Sequential Patterns. The existing solutions to this problem make repeated database scans, and use complex hash structures which have poor locality. SPADE utilizes combinatorial properties to decompose the original problem into smaller sub-problems that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. All sequences are discovered in only three database scans. Experiments show that SPADE outperforms the best previous algorithm by a factor of two, and by an order of magnitude with some pre-processed data. It also has linear scalability with respect to the number of input-sequences, and a number of other database parameters. Finally, they discuss how the results of sequence mining can be applied in a real application domain [8].

III. EXISTING SYSTEM

The existing mostly concentrates on extracting topics from a document stream using text mining technique as well as some systems just concentrates on the correlations between the topics. But it ignores the sequential topic pattern mining which is an essential part for the real time monitoring.

IV. PROPOSED SYSTEM

A] Block diagram:

It consists of three phases as shown in fig. 1. At first, textual documents, in our case documents streams are tweets are collected using the twitter archiver tool and used a document stream as the input. Then, as preprocessing procedures, the 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 discover all the STP candidates in the document stream for all users, and further pick out significant URSTPs associated to specific users by user-aware rarity analysis.

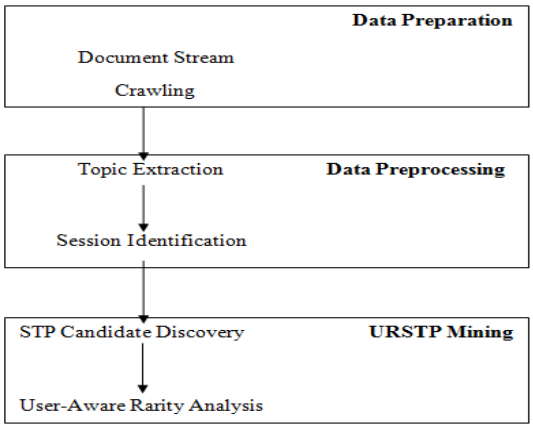

Fig. 1: Block diagram to formulate URSTP

After preprocessing, we will obtain a set of user-session pairs. For each of them with a specific user u, a new thread is started and a pattern-growth based subprocedure will be recursively invoked to find all the STP candidates for u, paired with their support values, and add the combined pair to the set User STP. These threads can be executed in parallel relying on the hardware environment. When all of them finish, another subprocedure will be called to make user-aware rarity analysis for these STPs together and get the output set, which contains all the pairs of users and their corresponding URSTPs with values of relative rarity.

V. CONCLUSION

Mining Rare Sequential Topic Patterns in document streams is a significant and challenging problem. It formulates complex event patterns based on document topics, which can be used in many application scenarios. We proposed a framework which is very effective and efficient that capture users' personalized and abnormal behaviors and characteristics.

REFERENCES

1. T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999, pp. 50–57.
2. D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
3. 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.
4. 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.
5. 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.
6. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining sequential patterns by prefix projected growth," in Proc. IEEE Int. Conf. Data Eng., 2001, pp. 215–224.
7. 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.
8. M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Mach. Learn., vol. 42, no. 1-2, pp. 31–60, 2001.
9. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 113–120.