ISSN 23499842(Online), Volume 2, issue 1 March 2017
International Journal of Innovative Trends and Emerging Technologies

A REVIEW ON MINING USER-AWARE RARE USING SEQUENTIAL TOPIC PATTERNS IN DOCUMENT STREAMS

B.Yuvaraj[1]and Dr. C. Sureshgnanadhas[2]
[1]*Research Scholar, Department of Computer Science and Engineering, Manonmaiam Sundaranar University, Tamil Nadu,*
[2] *Professor& Head, Department of Computer Science and Engineering, Vivekanandhan College of Engineering for Women, Elayampalayam, Thiruchencode, Tamil Nadu.*
E-Mail :[1] byuvarajb@gmail.com [2]sureshc.me@gmail.com

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.

Index terms: Web mining, sequential patterns, document streams, rare events, pattern-growth, dynamic programming.

1. Introduction
DOCUMENT streams are created and distributed in various forms on the Internet, such as news streams, emails, micro-blog articles, chatting messages, research paper archives, web forum discussions, and so forth. The contents of these documents generally concentrate on some specific topics, which reflect offline social events and users' characteristics in real life. To mine these pieces of information, a lot of researches of text mining focused on extracting topics from document collections and document streams through various probabilistic topic models, such as classical PLSI [15], LDA [7] and their extensions [5], [6], [16], [18], [19], [24].

Taking advantage of these extracted topics in document streams, most of existing works analyzed the evolution of individual topics to detect and predict social events as well as user behaviors [8], [11], [12], [23]. However, few researches paid attention to the correlations among different topics appearing in successive documents published by a specific user, so some hidden but significant information to reveal personalized behaviors has been neglected.

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). Each of them records the complete and repeated behavior of a user when she is publishing a series of documents, and are suitable for inferring users' intrinsic characteristics and psychological statuses. Firstly, 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. Secondly, 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 regularity about Internet users. Thirdly, 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 some specific user or some specific group of users. We call them Useraware 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 reallife scenarios of user behavior analysis, as illustrated in the following example.

Scenario 1 (Real-time monitoring on abnormal user behaviors).
Recently, micro-blogs such as Twitter are attracting more and more attentions all over the world. Micro-blog messages are real-time, spontaneous reports of what the users are feeling, thinking and doing, so reflect users' characteristics and statuses. However, the real intentions of users for publishing these messages are hard to reveal directly from individual messages, but both content information and temporal relations of messages are required for analysis, especially for abnormal behaviours without prior knowledge. What's more, if illegal behaviors are involved, detecting and monitoring them is particularly significant for

social security surveillance. For example, the lottery fraud behaviors via Internet usually accord with the following four steps, which are embodied in the topics of published messages: (1) make award temptations; (2) diddle other users' information; (3) obtain various fees by cheating; (4) take illegal intimidation if their requests are denied. STPs happen to be able to combine a series of inter-correlated messages, and can thus capture such behaviors and associated users.Furthermore, even if some illegal behaviors are emerging, and their sequential rules have not been explicit yet, 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.

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 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. 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. Firstly, 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 abstract and probabilistic descriptions of documents by topic extraction, and then to recognize complete and repeated activities of Internet users by session identificati on. Secondly, 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. Thirdly, 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.

2 Related Work

Topic mining in document collections has been extensively studied in the literature. Topic Detection and Tracking (TDT) task [3], [9] aimed to detect and track topics (events) in news streams with clustering-based techniques on keywords. Considering the co-occurrence of words and their semantic associations, a lot of probabilistic generative models for extracting topics from documents were also proposed, such as PLSI [15], LDA [7] and their extensions integrating different features of documents [5], [19], [24], as well as models for short texts [16], like Twitter-LDA.

Sequential pattern mining is an important problem in data mining, and has also been well studied so far. In the context of deterministic data, a comprehensive survey can be found in [21]. The concept support [25] 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 target database. Many mining algorithms have been proposed based on support, such as PrefixSpan [29], FreeSpan [13] and SPADE [36]. They discovered frequent sequential patterns whose support values are not less than a user-defined threshold, and were extended by SLPMiner [30] to deal with length decreasing support constraints. Nevertheless, the obtained patterns are not always interesting 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.

For uncertain data, most of existing works studied frequent item set mining in probabilistic databases [1], [10], but comparatively fewer researches addressed the problem of sequential pattern mining. Muzammal et al. focused on sequence-level uncertainty in sequential databases, and proposed methods to evaluate the frequency of a sequential pattern based on expected support, in the frame of candidate generate-and-test [28] or patterngrowth [26]. Since expected support would lose the probability distribution of the support, a finer measure frequentness probability was defined for general item sets [4], [32], [37], and used in mining frequent sequential patterns for sequencelevel and element -level uncertain databases [20], [27], [40]. However, these works did not consider where the uncertain databases come from and how the probabilities in the original data are computed, so cannot be directly employed for our problem which takes document streams as input. Moreover, they also focused on frequent patterns and thus cannot be utilized to discover rare but interesting patterns associated with special users This paper is an extension of our previous work [17], and has significant improvements on the following aspects:

- The problem of mining URSTPs is defined more formally and systematically and the application field focuses on published document streams;
- The formula to compute the relative rarity of an STP for a user is modified to become fully user-specific and more accurate;
- The preprocessing strategies including topic extraction and session identification are presented in detail, where several heuristic methods are discussed;
- Besides improving the approximation algorithm given in [17] which discovers STP candidates

with estimated support values, this paper presents a dynamic programming based algorithm to exactly compute the support values of derived STPs, which provides a trade-off between accuracy and efficiency;

- Experiments are conducted for new algorithms on more real Twitter datasets and more generalized synthetic datasets, and quantitative results for the real case are given to validate our approach.

3. PROBLEM DEFINITION

In this section, we give some preliminary notations, define several key concepts related to STPs, and formulate the problem of mining URSTPs to be handled in this paper

1. Preliminaries

At first, we define documents in a usual way

2. Sequential Topic Patterns

On the Internet, the documents are created and distributed in a sequential way and thus compose various forms of published document streams for specific websites. In this paper, we abbreviate them as document streams.

3. User-Aware Rare Sequential Topic Patterns

Most of existing works on sequential pattern mining focused on frequent patterns, but for STPs, many infrequent ones are also interesting and should be discovered. Specifically, when Internet users' publish documents, the personalized behaviors characterized by STPs are generally not globally frequent but even rare, since they expose special and abnormal motivations of individual authors, as well as particular events having occurred to them in real life. Therefore, the STPs we would like to mine for user behavior analysis on the Internet should be distinguishing features of involved users, and thus satisfy the following two conditions.

4. MINING URSTP

In this section, we propose a novel approach to mining URSTPs in document streams. The main processing framework for the task is shown in Fig. 2. It consists of three phases. At first, textual documents are crawled from some micro-blog sites or forums, and constitute a document stream as the input of our approach. 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.

In order to fulfil this task, we design a group of algorithms. To unify the notations, many variables are denoted and stored in the key-value form. For example, Uteruses represents the set of user-session pairs, and each of its elements is denoted as hu : Sui , in which the user u is the key of the map and its value Su is a set containing all the sessions associated with u.

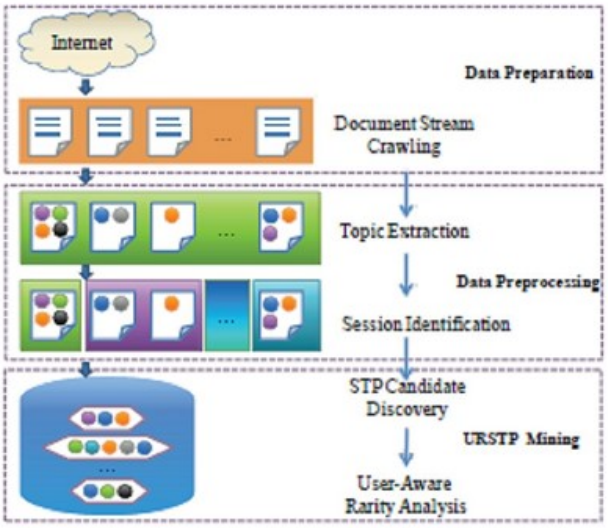

Fig. 2. Processing framework of URSTP mining

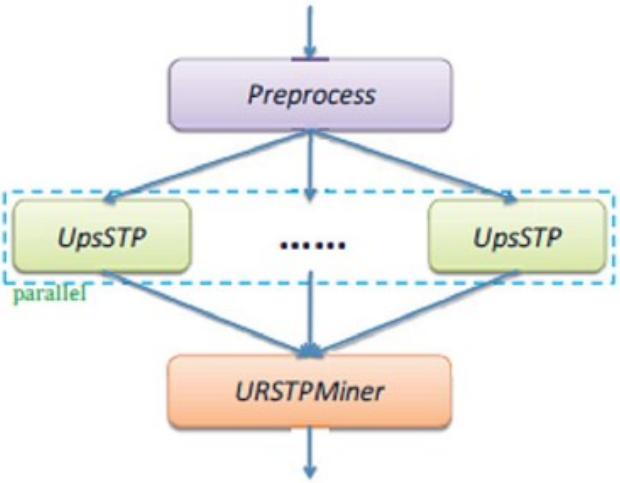

Fig. 3. Workflow of URSTP mining

5. Experiments

Since the problem of mining URSTPs in document streams proposed in this paper is innovative, there are no other complete and comparable approaches for this task as the baseline, but the effectiveness of our approach in discovering personalized and abnormal behaviors, especially the reasonability of the URSTP definition, needs to be practically validated. In this section, we conduct interesting and informative experiments on message streams in Twitter datasets, to show that most of users discovered by our approach are actually special in real life, and the mined URSTPs can indeed capture personalized and abnormal behaviors of Internet users in an understandable way. In addition, we also evaluate the efficiency of the approach on synthetic datasets, and compare the two

alternative sub procedures of STP candidate discovery to demonstrate the tradeoff between accuracy and efficiency.

6. Conclusions And Future Work

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] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in Proc. ACM SIGKDD'09, 2009, pp. 29–38.
[2] R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc.IEEE ICDE'95, 1995, pp. 3–14.
[3] J. Allan, R. Papka, and V. Lavrenko, "Online new event detection and tracking," in
Proc. ACM SIGIR'98, 1998, pp. 37–45.
[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD'09, 2009, pp. 119–128.
[5] D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.
[6] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM ICML'06, 2006, pp. 113–120.
[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in Proc. IEEE VAST'12, 2012, pp. 143–152.
[9] K. Chen, L. Luesukprasert, and S. T. Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," IEEE Trans. Knowl. Data Eng., vol. 19, no. 8, pp. 1016–1025, 2007.
[10] C. K. Chui and B. Kao, "A decremental approach for mining frequent itemsets from uncertain data," in Proc. PAKDD'08, 2008, pp. 64–75.
[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 VAST'12, 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. VLDB'05, 2005, pp. 181– 192.
[13] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, "FreeSpan: frequent pattern-projected sequential pattern mining," in Proc. ACM SIGKDD'00, 2000, pp. 355–359.
[14] N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latent topic sequential patterns," in Proc. ACM RecSys'12, 2012, pp. 131–138.
[15] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. ACM SIGIR'99, 1999, pp. 50–57.[16] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. ACM SOMA'10, 2010, pp. 80–88.
[17] 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 SDM'14, 2014, pp. 533–541.
[18] A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in Proc. ACM ICML'06, 2006, pp. 497–504
[19] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in Proc. ACM ICML'06, vol. 148, 2006, pp. 577–584.
[20] Y. Li, J. Bailey, L. Kulik, and J. Pei, "Mining probabilistic frequent spatiotemporal sequential patterns with gap constraints from uncertain databases," in Proc. IEEE ICDM'13, 2013, pp. 448–457.
[21] 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.