User-Aware Rare Sequential Topic Patterns in Data Mining Document Streams

N.Akhila[1], Dr.G.V.Ramesh Babu[2]
[1] *Student, Dept. of MCA, SVU.College of C M & C's*
[2] *Assistant Professor, Dept. of MCA, SVU.College of C M & C's, Tirupati,A.P*

***Abstract*- Textual documents created and distributed on the Internet are ever changing in various forms. 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. Most of the existing works are devoted in this topic modeling and they system evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. We present of a group of algorithms then to solve this innovative mining of problem through three phases these are preprocessing to extract topics and identify they 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*- Sequential patterns, rare events, pattern-growth, dynamic programming, document streams, Web mining.**

I. INTRODUCTION

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], 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. ocument 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 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

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 behaviors 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.

II. RELATED WORK

Topic mining in document collections has been extensively studied in the literature. TopicDetection and Tracking (TDT) task [3], [9], [35] 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], [34], like Twitter-LDA [39]. In many real applications, document collections generally carry temporal information and can thus be considered as document streams. Various dynamic topic modeling methods have been proposed to discover topics over time in document streams [6], [18], [33], [38], and then to predict offline social events [8], [11], [23]. However, these methods were designed to construct the evolution model of individual topics from a document stream, rather than to analyze the correlations among multiple topics extracted from successive documents for specific users. 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], [25]. 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 lengthdecreasing support constraints. Nevertheless, the obtained.

III. PRELIMINARIES

At first, we define documents in a usual way. Definition 1 (Document). A textual document d in a document collection D consists of a bag of words from a fixed vocabulary V = {w1,w2,⋯ ,w|V |}. It can be represented as {c(d,w)}w∈V , where c(d,w) denotes the occurrence number of the word w in d.

Given a document collection D and a topic number K, latent topics of these documents can be learnt through probabilistic topic models like LDA [7] and Twitter-LDA [39], and comprise a set T. Each topic is defined as follows. Definition 2 (Topic). A semantically coherent topic z in the text collection D is represented by a probabilistic distribution of words in the given vocabulary V . It is denoted as {p(w|z)}w∈V , which satisfiesPw∈V p(w|z) = 1. In this way, each document can be represented by a probabilistic mixture (proportion) of these K independent topics, which form a structured topic-level document. Definition 3 (Topic-Level Document). Given an original document d ∈ D and a topic set T, the corresponding topic-level document tdd is defined as a set of topic probability pairs, in the form of tdd = {(z,p(z|d))}z∈T. Here, the probabilities are obtained through some topic model and satisfy Pz∈T p(z|d) = 1. The superscript d can be omitted when the original document is not cared.

Actually, we can select some representative topics from T to approximately describe the document, which will be discussed in the preprocessing procedure in the next section.

3.1 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. Definition 4 (Document Stream). A document stream is defined as a sequence DS = h(d1,u1,t1),(d2,u2,t2),⋯ ,(dN,uN,tN)i, 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, so we can assume that at any time point, for any specific user, at most one document is published. Formally, if ti = tj, then ui 6= uj always hold. Obviously, each document stream can be transformed into a topic-level document stream of the form TDS = h(td1,u1,t1),(td2,u2,t2),⋯ ,(tdN,uN,tN)i, by extracting topics for each document according to Definition 3. In this paper, we pay attention to the correlations among successive documents published by the same user in a document stream. A kind of fundamental but important correlations is the sequential relation among topics of these

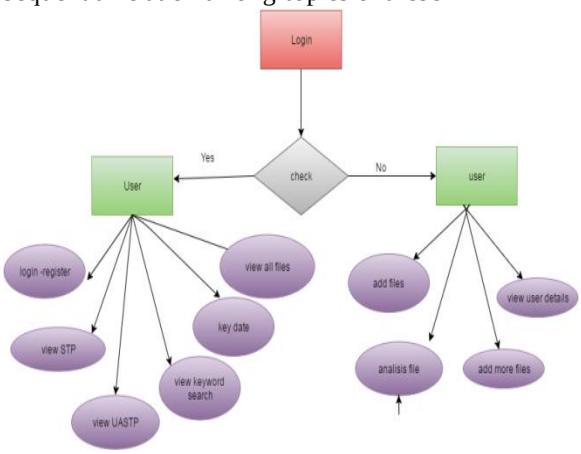

Fig. 1. Sketch map of session identification.

IV 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 topiclevel document stream and then divided into many sessions to identify complete userbehaviors. Finallyand 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, User Sess

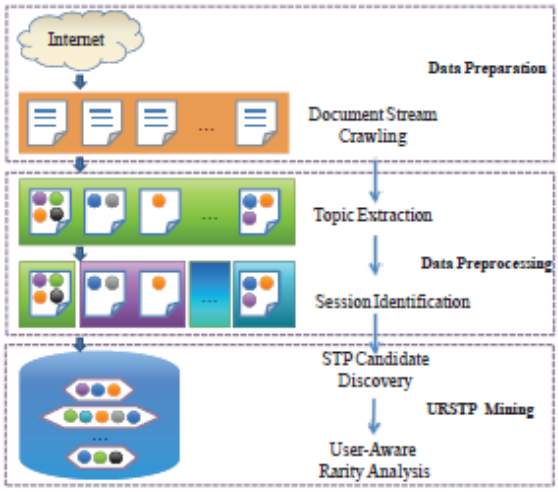

Fig. 2. Processing framework of URSTP mining.

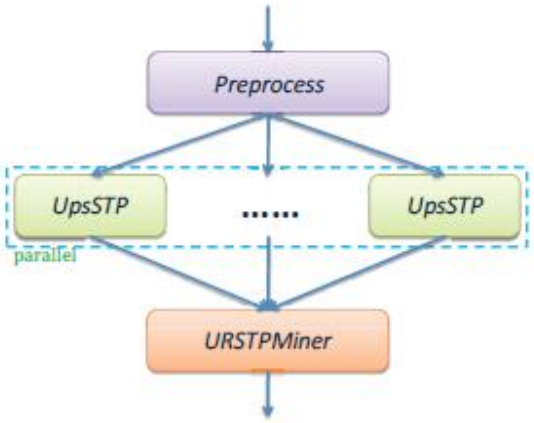

Fig. 3. Workflow of URSTP mining.

4.1 Data Preprocessing

4.1.1 Topic Extraction

In order to obtain a topic-level document stream, we at first employ the classical probabilistic topic models like LDA [7], [22] and Twitter-LDA [39] to get the topic proportion of each document and the word distribution of each learnt topic, with a predefined topic number K. For each document, the generated topic proportion may contain some topics with low probability. They cannot reflect the content of the document with high confidence, so can be excluded from the topic-level representation to reduce the complexity of later computations. To this end, we select some representative topics to get an approximate topic-level document. The input of this process is the topic proportion of a document d of the form {p(z|d)}z∈T satisfying Pz∈T p(z|d) = 1, while the output is a topic-level document of the form {(z1,p1),(z2,p2),⋯ ,(zK′,pK′)}. It satisfies that K′ ≤

K, and for all i = 1,...,K′, zi ∈ T and pi = p(zi|d) hold, which implies PK′ i=1 pi ≤ 1. There are two main selection strategies as follows. The pseudocodes are omitted here due to the page limit.

1) Topic Probability Threshold. It selects all the probabilities more than or equal to a predefined threshold htp. Formally, for all i = 1,...,K′, pi ≥ htp holds, and for all z ∈ T − {z1,⋯ ,zK′}, p(z|d) < htp holds. 2) Probability Summation Threshold. After sorting the probability values of the K topics in the nonincreasing order, it selects them according to the order as many as possible such that their summation is less than or equal to a predefined threshold hps. Formally, PK′ i=1 pi ≤ hps holds, and for all z ∈ T−{z1,⋯ ,zK′},PK′ i=1 pi+p(z|d) > hps holds.

4.1.2 Session Identification

Since each session should contain a complete publishing behavior of an individual user, we need to at first divide the document stream according to different users, which is an easy job as the author of each document is explicitly given in the input stream. The result for each user u is a subsequence of the topic-level document stream restricted to that user, i.e., TDSu = h(td1,u,t1),(td2,u,t2),⋯ ,(tdN,u,tN)i. After that, we also need to partition the subsequence to identify complete and repeated activities as consecutive and non-overlapped sessions. They constitute a session set Su = {s1,s2,⋯ ,sm} satisfying TDSu = s1 ◦ s2 ◦ ⋯ ◦ sm, where ◦ is the concatenation operator.

V. 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 subprocedures of STP candidate discovery to demonstrate the tradeoff between accuracy and efficiency.

VI. 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.

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, "On-line 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 spatio-temporal 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.

[22] A. K. McCallum. (2002) MALLET: A machine learning for language toolkit. [Online]. Available:

[23] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in Proc. WWW'06, 2006, pp. 533–542.

[24] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in Proc. ACM ICML'07, 2007, pp. 633–640.

[25] 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.

[26] M. Muzammal, "Mining sequential patterns from probabilistic databases by pattern-growth," in Proc. BNCOD'11, 2011, pp. 118– 127.

[27] M. Muzammal and R. Raman, "On probabilistic models for uncertain sequential pattern mining," in Proc. ADMA'11, 2010, pp. 60–72

[28] "Mining sequential patterns from probabilistic databases," in Proc. PAKDD'11, 2011, pp. 210–221.

[29] 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 ICDE'01, 2001, pp. 215–224.

[30] M. Seno and G. Karypis, "SLPMiner: An algorithm for finding frequent sequential patterns using length-decreasing support constraint," in Proc. IEEE ICDM'02, 2002, pp. 418–425.