User-Aware Rare Sequential Topic Patterns in Document Streams

SHAIK NAHIDA BEGUM[1], SHAIK FAAMIDA[2], SHAIK DARWIN[3], RAMINENI CHAITANYA SREE[4], P. SIVA PRASAD[5]
[1,2,3,4,5] *Computer Science of Engineering, Vasireddy Venkatadri Institute of Technology, Guntur*

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.

I. 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 to pics, 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], [33], [34], [38], [39]. 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 regularities 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.

II. RELATED WORK

Topic mining in document collections has been extensively studied in the literature. Topic Detection 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 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 is 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 pattern-growth [26]. Since expected support would lose the probability distribution of the support, a finer measure frequentness probability was defined for general itemsets [4], [32], [37], and used in mining frequent sequential patterns for sequence-level 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.

In the aspect of sequential patterns for topics, Haririet al. [14] presented an approach for context-aware music recommendation based on sequential relations of latent topics. The topic set of each song is at first determined by a threshold on the topic probabilities obtained from LDA. Then, frequent topic-based sequential patterns occurring among playlists are discovered to predict the next song in the current interaction. Nevertheless, the topic sets here are deterministic, so the uncertainty degree of topics is lost due to the approximation in the threshold-based filtering. In addition, the target is not a published document stream, and the globally rarity was not taken into account to find personalized and uncommon patterns. 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 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 cases are given to validate our approach.

A. EXISTING SYSTEMS:

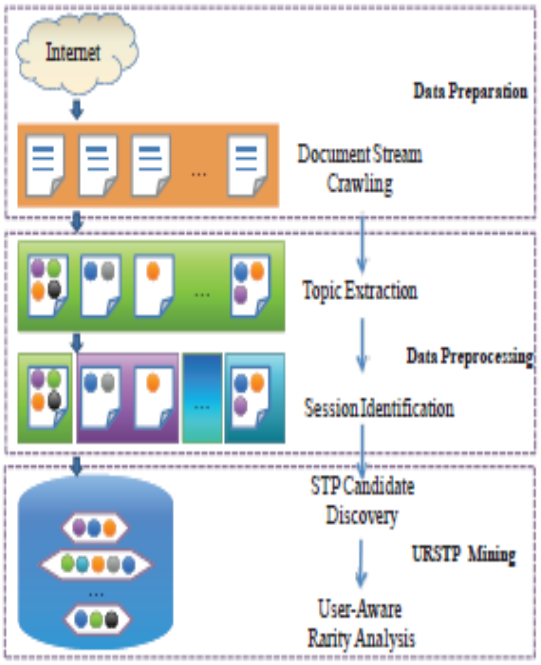
Most of existing works are devoted to topic modeling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. Taking advantage of these extracted topics in document streams, most of existing works analyzed the evolution of individual topics to detect and predict social events as well as user behaviors. 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. 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. 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.

B. PROPOSED SYSTEM:

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. 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. Topic mining in document collections has been extensively studied in the literature. Topic Detection and Tracking (TDT) task aimed to detect and track topics (events) in news streams with clustering-based techniques on keywords. 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.

III. 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. 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 pic 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, User_ Sess



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. All the structures of such sets of pairs used in our algorithms are summarized. The workflow of our approach is presented in Algorithm 1 gives the pseudo-code of the main procedure.

The input includes an original document stream DS = h(d1, u1, t1), (d2, u2, t2), ・ ・ ・ , (dN, uN, tN)i, a scaled support threshold hss and a relative rarity threshold hrr. As discussed later, there are still some thresholds used in preprocessing procedures,

but since preprocessing strategies will be chosen with some common rules according to the characteristics of the input stream, we think 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.

Algorithm 1. Main(DS, hss, hrr)
1: User Sess ← Preprocess(DS);
2: User STP ← ?;
3: for all hu : Sui ∈ User Sess do
4: Start a new thread;
5: STP Suppu ← UpsSTP(?, Su,?, Su);
6: User STP ← User STP ∪ {hu : STP Suppui};
7: User URSTP ← URSTPMiner(User STP,User Sess, hss, hrr);
8: return User URSTP;

After preprocessing, we 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 sub procedure UpsSTP is recursively invoked to find all the STP candidates for u, paired with their support values, and add the combined user-STP 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 URSTPMiner will be called to make user-aware rarity analysis for these STPs together and get the output set User URSTP, which contains all the pairs of users and their corresponding URSTPs with values of relative rarity.

IV. EXPERIMENTS

We collect two Twitter datasets as real document streams, a general dataset and a special sports-related dataset. To get the general dataset, we start from a famous user "SteveNash", crawl 150 latest tweets and 50 randomly selected active friends of him through Twitter's Rest API, and put these users in a waiting queue. Here, the activeness is determined by the total tweet number (not less than 150) and friend number (not less than 50). Then, this process is repeated for the users in the queue until 2000 users are collected, which realizes a breadth-first user traversal. The direct and indirect friends of the seed user spread over various kinds of fields, so the topics of these tweets are diversified. After removing those users with too high or too low publishing rates as well as very short and non-English tweets, the dataset contains 1950 users and 183960 tweets. The special dataset is obtained in a similar way, except that the seed user becomes a sports journalist "WojVerticalNBA". Most of his friends are closely connected to sports, such as journalists, players and commentators. To control the tweet contents, we remove the users irrelevant to sports according to the descriptions in their profiles. Consequently, the topics of tweets in this dataset focus on sports, but the subtopics are various and reflect user's characteristics and roles. The dataset contains 955 users and 94943 tweets. In the preprocessing phase, we use a public package of the Twitter-LDA model [39] in Github developed by the SMU Text Mining Group, with the topic number K = 15 and K = 10, respectively for the two datasets. In this model, each tweet is assumed to talk about only one topic, so each derived topic-level document just contains a single topic with probability 1. Although later computations are still feasible, the uncertainty degree is totally lost. We recover it by recording the topic values of each tweet at 10 iteration points after the burn-in period (1000 iterations), and computing the proportion of them to get probabilistic topics. We find 60% of tweets involve a unique topic, and others follow biased distributions. That implies convergence and coincides with the characteristics of short tweets. Then, the Topic Probability Threshold with value 0.3 is adopted to select representative topics, and sessions are identified through the Time Interval Heuristics with the threshold set to 5 hours. Afterwards, STP candidates for all users are discovered by calling the subprocedure UpsSTP with five parallel threads. Here, we restrict the STP length in between 2 and 4, as longer STPs are generally insignificant and hard to interpret. Then, we apply URSTPMiner on these STPs to mine user-aware rare ones. Notice that our target is 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 of those URSTPs with topmost values of the relative rarity, as well as their associated users. To this end, we set hss = 0.05 and hrr = 0.3 to get relatively strong conditions, and evaluate on a small but representative result set. In addition, we also use the approximation algorithm UpsSTP-a to replace UpsSTP, and carry out the two steps of mining for comparison. Very similar results of URSTPs are obtained and omitted here due to the page limit. All the experiments were conducted on a

desktop with Intel(R) Core(TM) 3GHz i5 CPU and 6GB RAM. The algorithms are implemented in Java, and run in command linewith Java 1.7.0 79 on Ubuntu 12.04.

V. 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. 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 reallife 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, "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.

[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," J. 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.