www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 8 August 2017, Page No. 22291-22295
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i8.10

A Review on Rare Sequential Topic Patterns

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

[1]Shivaji University, Bharati Vidyapeeth's College of Engineering,
Kolhapur, Maharashtra, India
patilbhakti9@gmail.com
[2]Bharati Vidyapeeth's College of Engineering, Shivaji University,
Kolhapur, Maharashtra, India
sachintakmare@gmail.com
[3]Bharati Vidyapeeth's College of Engineering, Shivaji University,
Kolhapur, Maharashtra, India
pramodkharade@gmail.com
[4]Bharati Vidyapeeth's College of Engineering, Shivaji University,
Kolhapur, Maharashtra, India
rahulmirajkar982@gmail.com

Abstract:
Textual documents are created in various forms. Existing work is devoted to topic modeling and the evolution of individual topics, but sequential relations of topics in successive documents created by specific users are ignored. In this paper we introduce Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams that characterize and detect personalized and abnormal behaviors of users. Such an innovative problem of mining will be solved by using three phases: extraction of probabilistic topics using preprocessing 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. Finally our result reflects users' characteristics.

Keywords: dynamic programming, pattern-growth, rare events, sequential patterns**.**

1. Introduction

Document streams are created and distributed in various forms such as chatting messages, email, news etc. The content of these documents describes some specific topic. Mining these pieces of information, text mining focused on extracting topics from document collections and document streams using various probabilistic topic models, such as LDA [1] and their extensions [2], [3], [10], [11].

Existing works analyzed the individual topics detect and predict social events and user behaviors by using the extracted topics in document streams [4].Some researches concentrates on correlations among different topics of successive documents published by a specific user, so some hidden but important information to present personalized behaviors has been neglected.

We study the correlations between topics, which are extracted from documents, mostly the sequential relations, to characterize user behaviors in published document streams and specify them as Sequential Topic Patterns (STPs).

For a document stream, some STPs may occur frequently and other patterns which are globally rare for some specific user or group of users which are known as User-aware Rare STPs (URSTPs). Frequent patterns reflect common behaviors. But discovering URSTPs is interesting and significant. it defines a new kind of patterns for rare event mining, which is able to characterize personalized and abnormal behaviors for special users.

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. , 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 [4], FreeSpan [5] and SPADE [6]. They discovered frequent sequential patterns whose support values are greater than a user-defined threshold, and were extended by SLPMiner [7] 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.

By taking these issues into consideration, problem of mining URSTPs is defined more formally and systematically, and we will try to focus on published document streams;

[1]. The formula to compute the relative rarity of an STP for a user is modified to become fully user-specific and more accurate;

[2]. The preprocessing strategies including topic extraction and session identification are presented using several heuristic methods.

[3]. We are trying to present dynamic programming based algorithm that exactly compute the support values of derived STPs, which provides a trade-off between accuracy and efficiency.

2. Literature Survey

Blei, Ng, Jordan have proposed a generative probabilistic model to solve the problem of modeling text document and other collections of discrete data. A generative probabilistic model referred as Latent Dirichlet Allocation (LDA) is a three level hierarchical Bayesian model which models each item as an infinite mixture with underlying set of topic probabilities. An explicit representation of document is represented by topic probabilities. They presented efficient approximate inference techniques methods and an EM algorithm for empirical Bayes parameter estimation. Also they presented results of document modeling, text classification and collaborative filtering and compared with a mixture of unigrams model and the probabilistic LSI model [1].

A textual document carries important events and which are temporal over time. So Dou, Wang, Skau, Ribarsky and Zhou have proposed an interactive visual analytics system LeadLine that automatically identify meaningful events in news and social media data and provides exploration of the events. It integrates topic modeling, event detection, and named entity recognition techniques to automatically extract information. LeadLine provides a concise summary in case of large-scale text documents through events. It also supports the construction of simple narratives. The results of LeadLine indicates that it can not only accurately identify meaningful events given a text collection, but can also contribute to users' understanding of the events through interactive exploration [8].

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu have re-examined the sequential pattern mining problem and proposed FreeSpan (frequent pattern-projected sequential pattern mining) that integrates the mining of frequent sequences with that of frequent patterns and then use projected sequence databases. FreeSpan mines the complete set of patterns by reducing efforts of candidate subsequence generation [5].

J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu have proposed a novel approach for sequential topic patterns called Prefix-Span (Prefix-projected Sequential pattern mining) explores prefix projection in sequential pattern mining. It also reduces the efforts of candidate subsequence generation and reduces the size of projected databases which leads to efficient processing[4].

Zaki have proposed a SPADE (Sequential Pattern Discovery using Equivalence classes) algorithm which overcomes a problem of repeated database scans and use complex hash structures which have poor locality and produces a fast discovery of Sequential Patterns. SPADE divides the original problem into smaller sub-problems; these sub-problems can be solved independently. SPADE makes only three database scan [6].

Seno & Karypis have proposed a SLPMiner that finds all frequent sequential patterns which satisfies a length-decreasing support constraint. SLPMiner combines an efficient database-projection-based approach for sequential pattern discovery with three effective database pruning methods that reduces the search space. Experimental evaluation shows that

SLPMiner, by effectively exploiting the length-decreasing support constraint, is up to two orders of magnitude faster, and its runtime increases as the average length of the sequences and the discovered frequent patterns increases. The pruning methods are not specific to SLPMiner but almost all of them can be incorporated into other algorithms for sequential pattern discovery [7].

Zhao, Jiang, Weng, He, Lim, Yan and Li have compare the content of Twitter with a traditional news medium. They use a Twitter-LDA model to discover topics from a representative sample of the entire Twitter and then use text mining techniques to compare these Twitter topics with topics from New York Times by considering topic categories and types. They also study the relation between the proportions of opinionated tweets and retweets and topic categories and types [9].

3. Problem Definition

3.1 Preliminaries

At first, we define some basic definitions.

- Definition 1(Document):

A textual document d in document collection D is made up of different words from a fixed vocabulary V.

- 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):

A topic- level document is a set of topic-probability pairs with probability 1.

- Definition 4 (Document Stream):

A document stream is defined as a sequence (DS) $=\{(d_1,u_1,t_1),(d_2,u_2,t_2), \ldots, (d_n,u_n,t_n)\}$ consists of a document d_i published by user u_i at time t_i on a specific website, and $t_i \leq t_j$ for all $i \leq j$.

- Definition 5 (Sequential Topic Pattern):

A Sequential Topic Pattern (STP) α is defined as a topic sequence $\{z_1, z_2, \ldots, z_n\}$ where each $z_i \in T$ is a learnt topic.

- Definition 6 (Session):

A session s is defined as a subsequence of TDS (topic level document stream) associated with the same user, i.e., $s=\{(td_1,u_1,t_1),(td_2,u_2,t_2),\ldots,(td_n,u_n,t_n)\}$.

- Definition 7 (User-Aware Rare STP)

A topic-level document stream TDS, a scaled support threshold hss, and a relative rarity threshold hrr, an STP α is called a User-aware Rare STP (URSTP) if and only if both scaled support of α is less than or equal to scaled support threshold and relative rarity of user is greater than or equal to relative rarity threshold hold for some user u.

4. Mining Rare STP

In this section we propose a framework to mining user aware rare sequential topic patterns in document streams. The main preprocessing framework is shown in fig. below. Framework consists of four phases.

At first, we collect text documents such as news from blog, e-mail, tweeter tweet etc. and employ them as input to our framework. In preprocessing phase, the original textual stream is transformed into topic level document stream. Then for identification of complete user behaviors, topic level document stream is again divided into many sessions.

Finally, we discover all the STP candidates in the document stream for all users, and then find out significant Rare STPs related to specific users by user-aware rarity analysis.

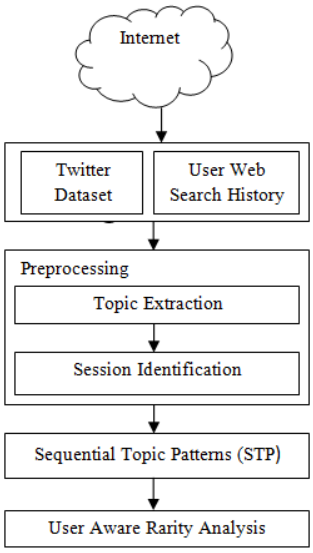

Figure 1: Preprocessing framework of Rare STP mining

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 subprocedure 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.

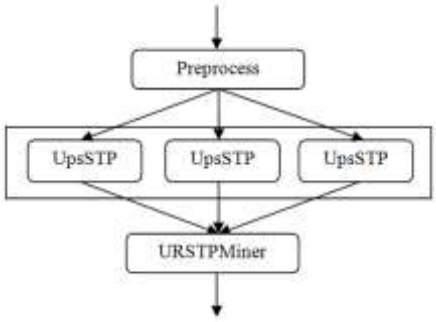

Figure 2: Rare STP Mining Workflow

After completion of all threads execution, another subprocedure 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.

- **STP Candidate Discovery by Pattern-Growth**

The STP candidate discovery procedure is executed in parallel for each user. It finds all STPs occurring in the document stream associated with a specific user, paired with the expected support values of these STPs. STPs can be finding using two different ways. First way is a DP-based algorithm that derives all STPs for the user and exactly computes the support values of them. Then, for efficiency improvement of our approach, we also give an approximation algorithm to estimate the support values for all STPs. Both algorithms are designed in the manner of pattern-growth.

- **User-Aware Rarity Analysis**

After discovering all the STP candidates for all users, we
will use next procedure to make the user-aware rarity analysis so that it can find out Rare STPs, which provides personalized, abnormal, and significant behaviors. It transforms the set of user-STP pairs into a set of user-URSTP pairs, with the set of user-session pairs and two thresholds, the scaled support threshold and the relative rarity threshold.

5. 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] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., Vol.3, pp. 993-1022, 2003.

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

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

[4] 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.

[5] 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.

[6] M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Mach. Learn., vol. 42, no. 1-2, pp. 31–60, 2001.

[7] 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.

[8] 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.

[9] 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.

[10] Z. Zhang, Q. Li, and D. Zeng, "Mining evolutionary topic patterns in community question answering systems," IEEE Trans. Syst., Man, Cybern. A, vol. 41, no. 5, pp. 828–833, Sep. 2011.

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

Sachin B. Takmare is working as Assistant Professor in Computer Science and Engineering Department of Bharati Vidyapeeth's College of Engineering, India with teaching experience of about 10 years.

Bhakti G. Patil received the B.E. degree in Computer Science from Shivaji University. She is currently pursuing M.E. in Computer Science and Engineering from Shivaji University, India.

Pramod A. Kharade is working as Assistant Professor in Computer Science and Engineering Department of Bharati Vidyapeeth's College of Engineering, India.

Rahul P. Mirajkar is working as Assistant Professor in Computer Science and Engineering Department of Bharati Vidyapeeth's College of Engineering, India.