INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume3, Issue4)

Available online at: www.ijarnd.com

User behavioral prediction

Veena Jayan[1], Alma Mary Margret[2]

[1]Student, Cochin College of Engineering and Technology, Valanchery, Kerala
[2]Assistant Professor, Cochin College of Engineering and Technology, Valanchery, Kerala

ABSTRACT

People use the Internet for different purposes e.g. social networking, blogging etc. with respect to their context. This leads to a dynamic change in creation and distribution of document streams over the Internet. This would challenge the topic modeling and evolution of individual topics. In this paper, we have proposed Sequential Topic Patterns (STPs) mining over the published user-aware document streams and formulate the problem of mining User- Aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet in order to find rare users. They are generally rare and infrequent over the Internet. For URSTPs mining we need to perform three phases: pre-processing to extract topics, generating STPs, determining URSTPs by rarity analysis of STPs. The experiment can be performed on both real times and synthetic data-sets. In the proposed work, we have focused on synthetic datasets.

Keyword: *Sequential patterns, Document streams, Rare sequential patterns, Pattern-growth, Dynamic programming.*

1. INTRODUCTION

Day by day the world is becoming more and more ubiquitous due to the dramatic increase in the popularity of the Internet services viz. social networking, e-commerce websites, e-learning websites etc. This generates and spreads the huge number of document streams over the Internet. So for determining the particular user's characteristic from its document stream is crucial. Data mining is the first and essential step in the process of knowledge discovery in this context. Various data mining methods are available such as association rule mining, sequential pattern mining, closed pattern mining and frequent item set mining to perform different knowledge discovery tasks from document streams. In real time scenario, we come across the micro-blog such as Twitter etc. where the users are spontaneously publishing their statuses. These messages are real-time and report what user is feeling and doing. So it can reveal users characteristics. However, it's difficult to guess the real intension or mindset of users behind it, but both content information and temporal relations are required for analyzing the user's characteristics.

There are some users which can use the Internet for abnormal purposes viz. online fraud, hijacking activity, spreading terrorism etc. Their behavior is undesirable for society and hence detecting such rare users become very essential. We formulate the problem of URSTPs mining for finding such abnormal and rare users. It is worth noting that the ideas above are also applicable to 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 an effective and context-aware recommendation for them. While this paper will concentrate on published document streams. In order to find rare users from their published document streams, we study the correlations among topics extracted from their document streams, especially the sequential relations, and specify them as Sequential Topic Patterns (STPs). Some of these STPs are frequently common for all the users but there are some patterns which are rare and infrequent. These RareSTPs (RSTPs) over the user-aware document streams constitute the URSTPs which are used to find the rare users. Sequential Pattern Mining is the method of finding interesting sequential patterns among the large databases.

It also finds out frequent sub sequences as patterns from a sequence database. Enormous amounts of data are continuously being collected and stored in many industries and they are showing interests in mining sequential patterns from their database. Sequential pattern mining has broad applications including web-log analysis, client purchase behavior analysis and medical record analysis. Sequential or sequence pattern mining is the task of finding patterns which are present in a certain number of instances of data [6]. The identified patterns are expressed in terms of sub sequences of the data sequences and expressed in an order that is the order of the elements of the pattern should be respected in all instances where it appears. If the pattern is considered to be frequent if it appears in a number of instances above a given threshold value, usually defined by the user, then it is considered to be frequent. There may be a huge number of possible sequential patterns in a large database. Sequential pattern mining identifies whether any

relationship occurs in between the sequential events [3]. The sequential patterns that occur in particular individual items can be found and also the sequential patterns between different items can be found. The number of sequences can be very large, and also the users have different interests and requirements. If the most interesting sequential patterns are to be obtained, usually a minimum support is pre-defined by the users. In this paper, we focus on the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in a sequence of sets. Mining sequential patterns have become an important data mining task with broad applications. For example, supermarkets often collect customer purchase records in sequence databases in which a sequential pattern would indicate a customer's buying habit [9]. Sequential pattern mining is commonly defined as finding the complete set of frequent subsequences in a set of sequences. Much research has been done to efficiently find such patterns.

2. EXISTING SYSTEM

Most of the existing works analyzed the evolution of individual topics to detect and predict social events as well as user behaviors. Many mining algorithms have been proposed based on support, such as *PrefixSpan*, *FreeSpan*, and *SPADE*. They discovered frequent sequential patterns whose support values are not less than a user-defined threshold and were extended by *SLPMiner* to deal with length decreasing support constraints. 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 or pattern-growth. 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 [12]. Furthermore, the algorithms on deterministic databases are not applicable for document streams, as they failed to handle the uncertainty in topics.

3. PROPOSED SYSTEM DESIGN

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).To solve the 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 identification. 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 the personalized and abnormal behaviors of Internet users, and can adapt to different application scenarios [10]. 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.

To the best of our knowledge, this is the first work that gives formal definitions of STPs as well as their rarity measures and puts forward the problem of mining URSTPs in document streams, in order to characterize and detect personalized and abnormal behaviors of Internet users. We propose a framework to pragmatically solve this problem, and design corresponding algorithms to support it. At first, we give preprocessing procedures with heuristic methods for topic extraction and session identification. Then, borrowing the ideas of pattern-growth in an uncertain environment, two alternative algorithms are designed to discover all the STP candidates with support values for each user [15]. That provides a trade-off between accuracy and efficiency. At last, we present a user-aware rarity analysis algorithm according to the formally defined criterion to pick out URSTPs and associated users. We validate our approach by conducting experiments on both real and synthetic datasets. These types of systems can be used in integration with social networking, various blogs, forums etc. To find the rare and abnormal users to maintain social harmony. This can be used to minimize cybercrimes by keeping track on abnormal and rare users.

4. IMPLEMENTATION

Implementation consists of several modules such as document stream crawling, STP candidate Discovery, RSTPs Mining, Topic Extraction.

4.1 Document Stream Crawling

It involves mainly three phases as document streams crawling, pre-processing to transform into topic level document stream and mining RSTPs over user-aware document streams [2]. It crawls the textual documents and acts as input stream for topic extraction.

4.2 STP candidate Discovery

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. In this module, we pay attention to the correlations among successive documents published by the same user in a document stream [1]. The sessions contain the topic level document streams for different users. In this step, the STPs are identified for the particular user.

4.3 RSTPs Mining

This step deals with RSTPs mining. These are very rare and infrequent patterns which are used to detect rare users. In this module, we propose a novel approach to mining URSTPs in document streams. 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 [8]. 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.

4.4 Topic Extraction

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. In the admin part, the user search history options are provided with the categories, they used and search keywords with their number of searches made [11]. From this, the user behavior can be analyzed and the analysis is made life through the sequential topics the users have accessed on their perspective login.

5. SYSTEM DESIGN

5.1 USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases [13]. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

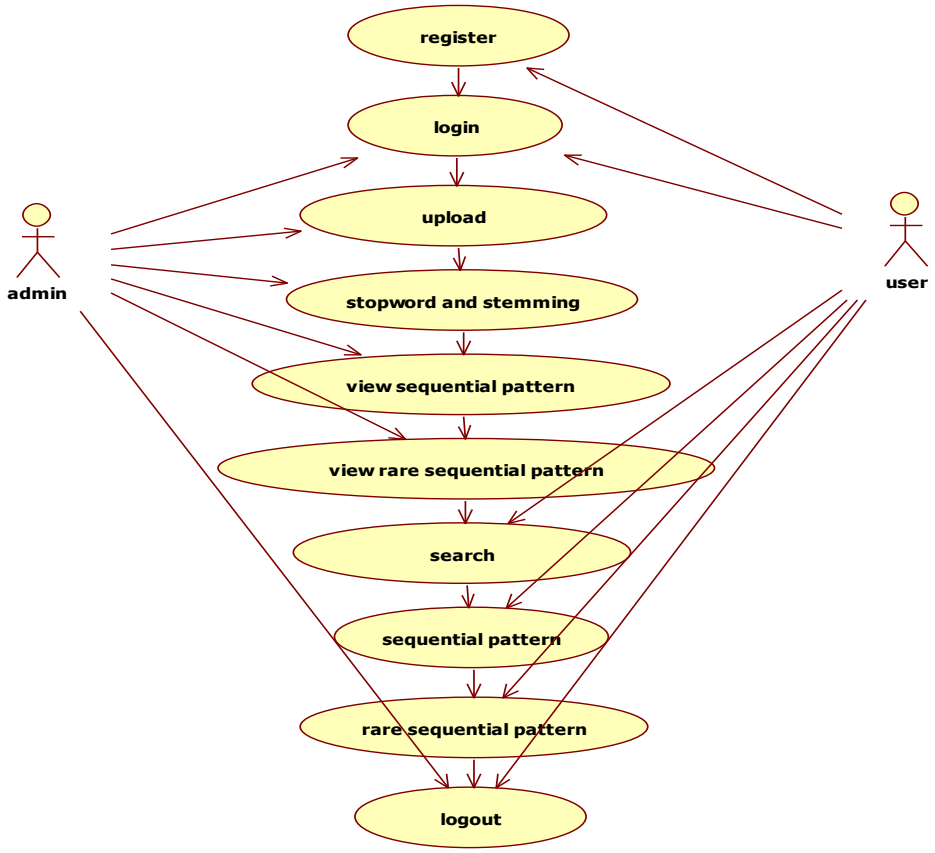

Fig 5.1 Usecase diagram

5.2 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

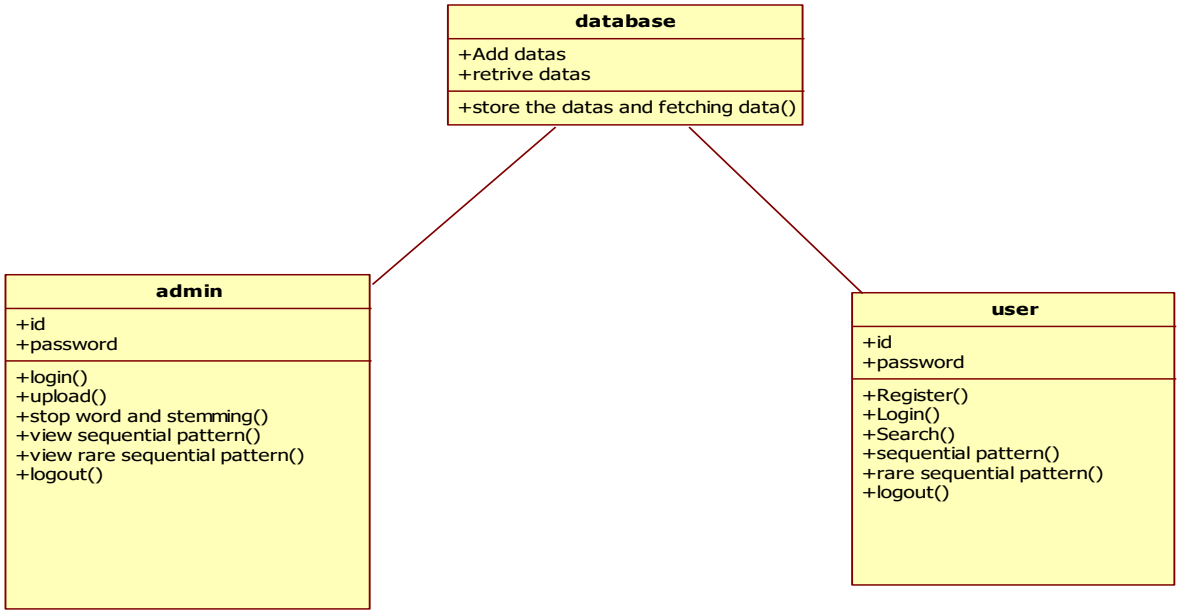

Fig 5.2 Class Diagram

5.3 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order [8]. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

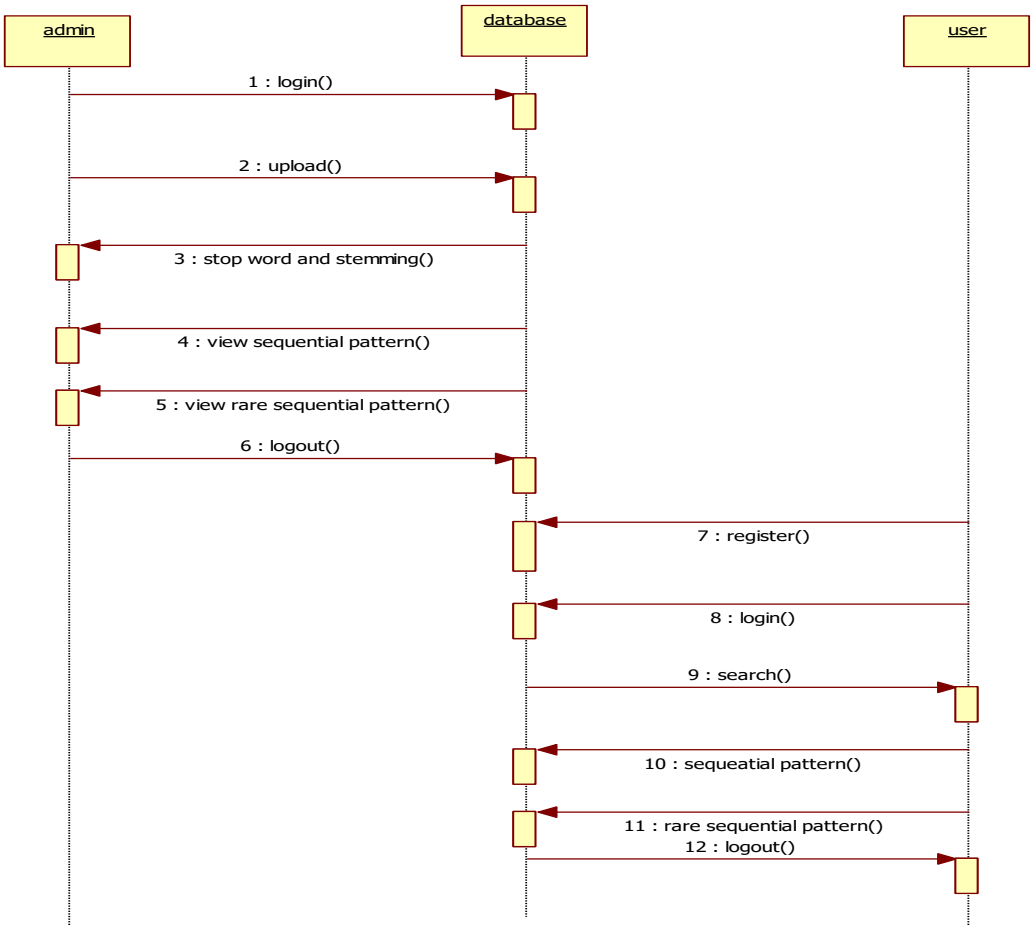

Fig 5.3 Sequence diagram

5.4 DEPLOYMENT

Component diagrams are used to describe the components and deployment diagrams shows how they are deployed in hardware. UML is mainly designed to focus on the software artifacts of a system. However, these two diagrams are special diagrams used to focus on software and hardware components.

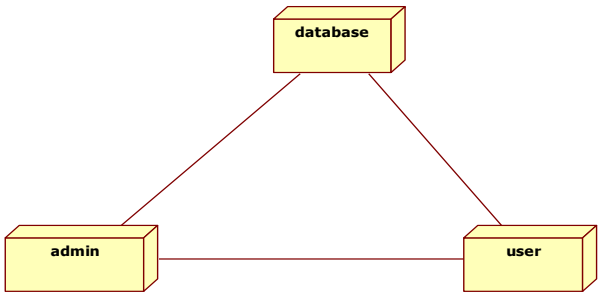

Fig 5.4 Deployment diagram

5.5 DATA FLOW DIAGRAM

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.DFD shows how the information moves through the system and how it is modified by a series of transformations [12]. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

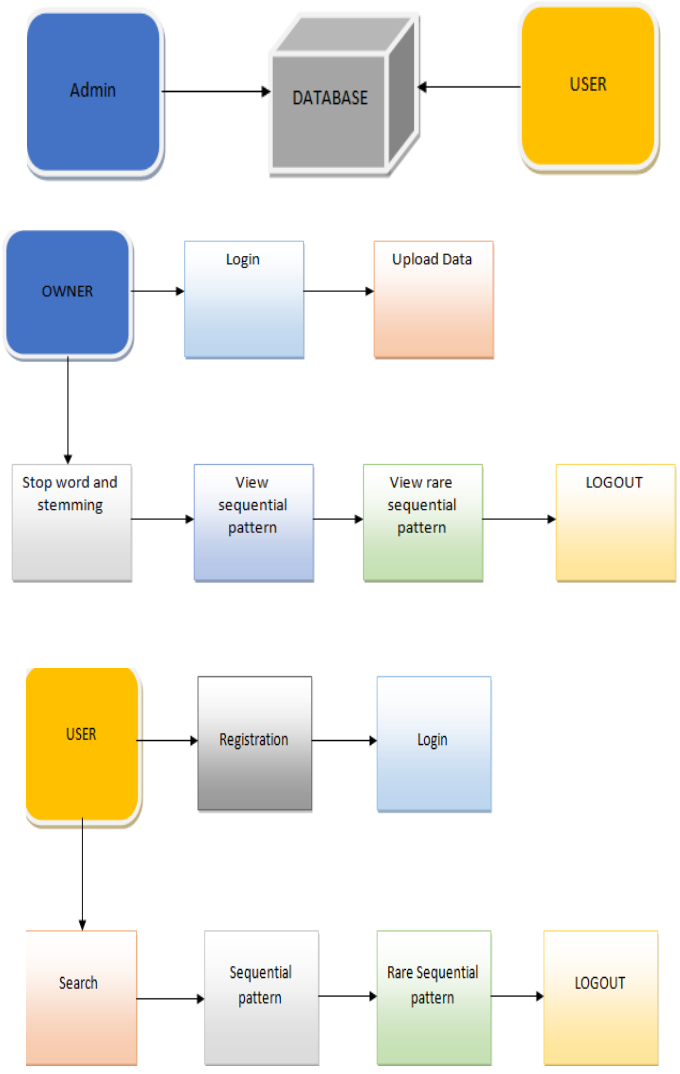

Fig 5.5 Data flow diagram

6. CONCLUSION

Knowledge discovery by various data mining techniques in documents streams is crucial. Topics are extracted in document streams and by topic modeling the sequential correlation is established to determine Sequential Topic Patterns (STPs). There are very rare uncommon patterns called Rare Sequential Topic Patterns called RSTPs. Mining RSTPs over user-aware document stream (URSTPs) is challenging task as users published the document streams dynamically. In order to find rare users from its published document streams over the Internet is difficult. So by mining RSTPs from the published user-aware document streams (URSTPs) we can find rare users. The future work consists of using predefined dictionaries for RSTPs designating abnormal users. If the comparison of discovered RSTPs by existing system to that of dictionaries' entries exceeds some threshold then system admin can block such users. In addition to this future work will consist of characterizing user's behavior by mining RSTPs over its browsed/surfed document streams and designing recommendation system.

7. ACKNOWLEDGEMENT

First and foremost I take immense pleasure in thanking the Management and respected principal, Mr. SAKKARIYA.T, for providing me with the wider facilities. I express my sincere thanks to Ms. ALMA MARY MARGRET, My guide nd Head of Department of Computer Science and Engineering, CCET for giving me the opportunity to present this project research and for timely suggestions. I wish to express my deep sense of gratitude to the PG Coordinator Mrs. JASEELA JASMIN T.K Asst professor, Department of Computer Science and Engineering, who coordinated in right path.

8. REFERENCES

[1] Jiaqi Zhu, Member, IEEE, Kaijun Wang, Yunkun Wu, Zhongyi Hu and Hogan Wang, Member, IEEE, Mining User-Aware Rare Sequential Topic Patterns in Document Streams, IEEE Transactions on Knowledge and Data Engineering, 2016
[2] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, Frequent pattern mining with uncertain data, in Proc. ACM SIGKDD09,pp.29-38, 2009.
[3] T. Bernecker, H.P. Kriegel, M. Renz, F. Verhein and A. Zuefle, Probabilistic frequent itemset mining in uncertain databases, in Proc. ACM SIGKDD09, pp.119-128, 2009.
[4] 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 V AST12, pp. 143-152, 2012.
[5] 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.
[6] C. K. Chui and B. Kao, A decremental approach for mining frequent itemsets from uncertain data, in Proc. PAKDD08, pp. 64-75, 2008.
[7] 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.
[8] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation,J. Mach. Learn. Res, vol. 3, pp. 993-1022, 2003.
[9] 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 VAST12, pp. 93102, 2012.
[10] G. P. C. Fung, J. X. Yu, P. S. Yu and H. Lu, Parameter-free bursty events detection in text streams, in Proc. [11] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal and M. Hsu, FreeSpan: frequent pattern-projected sequential pattern mining, in Proc. ACM SIGKDD00,pp. 355-359, 2000.
[12] A. K. McCallum and MALLET: machine learning for language toolkit, [Online]. Available: http://mallet.cs.umass.edu.
[13] 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 Adv. Inform. Retr. LNCS 6611, Springer, pp. 338-349, 2011.
[14] 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 ICDE01, pp. 215-224, 2001.
[15] Z. Zhao, D. Yan and W. Ng, Mining probabilistically frequent sequential patterns in large uncertain databases, IEEE Trans. Knowl. Data Eng., vol. 26, no. 5, pp. 1171-1184, 2014.

