iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** II **Month of publication:** February 2017

DOI: http://doi.org/10.22214/ijraset.2017.2019

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue II, February 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Study on Efficient Way to Identify User Aware Rare Sequential Pattern Matching in Document Stream

Swati V. Mengje[1], Prof. R R Shelke[2]
[1]ME Computer Science & Engineering Department, HVPM COET, AMRAVATI
[2]Asst. Professor, Computer Science & Engineering Department, HVPM COET, AMRAVATI

Abstract: ***As we know internet is the source of large number textual document those are created by users and distributed in various forms. Most of existing works are done on topic modelling 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 behaviours 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 behaviours. 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. Twitter is the best real time example, from that we able to discover the users abnormal behaviour. This approach gives the effective and efficient way to find out rare pattern in document string.***

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

I. INTRODUCTION

The main goal of mining is to discover the useful and unexpected patterns in database. There is lot of work in the field of data mining about pattern mining. In this, we will be interested by a specific type of database called sequence databases. Sequence database contains some sequences. Here we find out sequential pattern in document stream. Sequential pattern mining has wide applications on the client purchase behaviour analysis, web-log analysis and medical record analysis. We find out the pattern that is frequently used by the user. This technique is useful to find out the users abnormal behaviour on the internet. Sequence database sequence pattern mining is the task of finding patterns which are present in a certain number of instances of data. 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 huge number of possible sequential patterns in a large database. Sequential pattern mining identifies whether any relationship occurs in between the sequential events. 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. By using the minimum support, sequential patterns which are not so important is taken out and hence the mining process will be more efficient [2].

In this paper, we focus on the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in sequence of sets. Mining sequential patterns has 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. Sequential pattern mining is commonly defined as finding the complete set of frequent subsequences in a set of sequences [1]. Much research has been done to efficiently find such patterns. But to the best of our knowledge, no research has examined in detail what patterns are actually generated from such a definition. In this paper, we examined the results of the support framework closely to evaluate whether it in fact generates interesting patterns.

©IJRASET: All Rights are Reserved

Volume 5 Issue II, February 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

II. LITERATURE REVIEW & RELATED WORK ON STRING PATTERN MATCHING

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 , LDA [7] and their extensions integrating different features of documents [5] as well as models for short texts like Twitter-LDA . In many real applications, document collections generally carry temporal information and can thus be considered as document streams. Various dynamic topic modelling methods have been proposed to discover topics over time in document streams [6], and then to predict offline social events [8]. 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. The concept support 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* [15], *FreeSpan* [12] 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. Topic mining has been extensively studied in the literature. Topic Detection and Tracking (TDT) task [3] aimed to detect and track topics (events) in news streams with clustering-based techniques. Many generative topic models were also proposed, such as Probabilistic Latent Semantic Analysis (PLSA) [11], Latent Dirichlet Allocation (LDA) [5] and their extensions.

In many real applications, text collections carry generic temporal information and therefore can be considered as a text stream. To obtain the temporal dynamics of topics, various dynamic topic modeling methods have been proposed to discover topics over time in document streams [6]. However, these methods were designed to extract the evolution model of individual topics from a document stream, rather than to analyze the relationship among extracted topics in successive documents for specific users.

Sequential pattern mining has been well studied in the literature in the context of deterministic data, but not for topics with uncertainty. The concept support is the most popular criteria for mining sequential patterns. It evaluates frequency of a pattern and can be interpreted as occurrence probability of the pattern.

Many methods have been proposed to solve the problem of sequential pattern mining based on *support*, such as PrefixSpan [16], FreeSpan [9] and SPADE. These methods were designed to discover frequent sequential patterns whose supports are not less than a user-defined threshold minsupp. However, the obtained patterns are not always interesting, because those rare but significant patterns are pruned for their low supports. Furthermore, the frequent sequential pattern mining from deterministic databases is completely different from the STP mining that handles uncertainty of topics. Few researches addressed the problem of sequential pattern mining on uncertain data. Muzammal and Raman [10] proposed a method to discover frequent sequential patterns from probabilistic databases and evaluated the frequency of a pattern based on the expected support. However, the data model cannot be applied to topic sequences. In addition, they focused on the frequent pattern mining and failed to discover interesting rare patterns for some users.

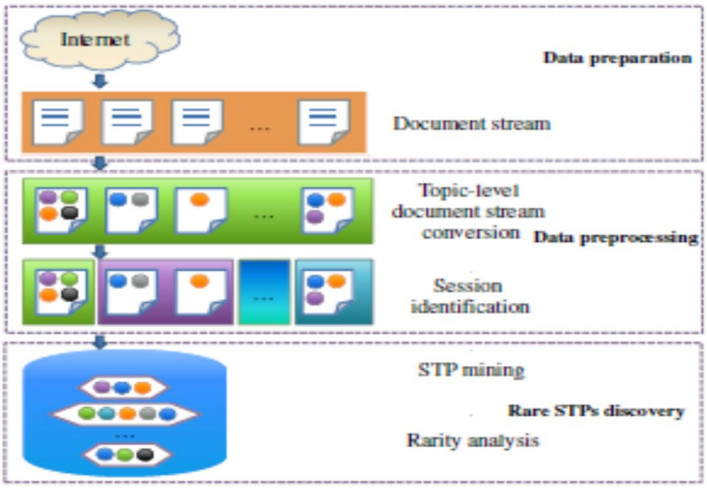


Fig. 1. Processing framework of URSTP mining.

©IJRASET: All Rights are Reserved

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. 1. 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 behaviours. 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.

III. ANALYSIS OF PROBLEM

Here analyzed the problem related to pattern matching. Given string T(text) and P(pattern), the pattern matching problem consists of finding a substring of T equal to P. Algorithms are designed and analyzed the problem. [13]. In information retrieval, to improve recall of a web search on a person name or any user entered sting a search engine can automatically expand a query using aliases of the name or string [14]. In our previous example, a user who searches for *Amitabh Bacchan* might also be interested in retrieving documents in which *Bacchan* is referred to as *BigB*. Consequently, we can expand a query on real name using his alias name *BigB*. The semantic web is intended to solve the entity disambiguation problem by providing a mechanism to add semantic metadata for entities. However, an issue that the semantic web currently faces is that insufficient semantically annotated web contents are available. Automatic extraction of metadata can accelerate the process of semantic annotation. For named entities, automatically extracted aliases can serve as a useful source of metadata, thereby providing a means to disambiguate an entity.

IV. RESEARCH CHALLENGES

Today several methods are available for efficiently discovering sequential patterns according to the initial definition. Such patterns are widely used for a large number of applications. But still there are various research challenges in this field of data mining. Some of the research challenges [16] are:

(1)Finding the complete set of patterns and satisfying the minimum support (frequency) threshold is a complex task. When the database is large, distributed sequential pattern mining is used for mining process which helps to increase the scalability.

(2)The ability to incorporate various kinds of user-specific constraints is a complex process. To add other useful constraints to the RFM patterns, for example, the constraint that the number of repetitions in a sequence must be no less than a given threshold.

(3)Constraints like frequency and Monetary constraints are difficult to be studied and checking their effect with respect to execution time, memory usage and scalability is also difficult.

(4)Algorithm should handle large search space. Repeated scanning of the database during the mining process must be reduced as much as possible.

(5)Various methods are used by which early candidate sequences are pruned and search space partitioning will be possible for efficient mining of patterns.

(6)There are many interesting problems especially in the development of specialized sequential pattern mining methods for particular applications such as DNA sequence mining[1] that may identify faults which in turn allows various insertions, deletions, and mutations in DNA sequences, and handling industry or engineering sequential process analysis are interesting issues for future research. The mining of multi-level time-interval sequential patterns are performed by using fuzzy time value.

V. CONCLUSION

Mining user-related rare Sequential Topic Patterns (STPs) in document streams on the Internet is an innovative challenging problem. It formulates a new kind of patterns for uncertain complex event detection and inference, and has wide potential application fields, such as personalized context aware recommendation and real-time monitoring on abnormal user behaviours on the Internet. Due to the continuous addition of large amount of data in the databases, the idea of sequential pattern mining is becoming popular. Various algorithms have been developed that are used for mining the sequential patterns in the data. These algorithms have proved to be more effective for smaller databases, but when the size of the database is increased, their performance may decline. Hence these methods have to be improved in order to perform the mining processes in a better way.

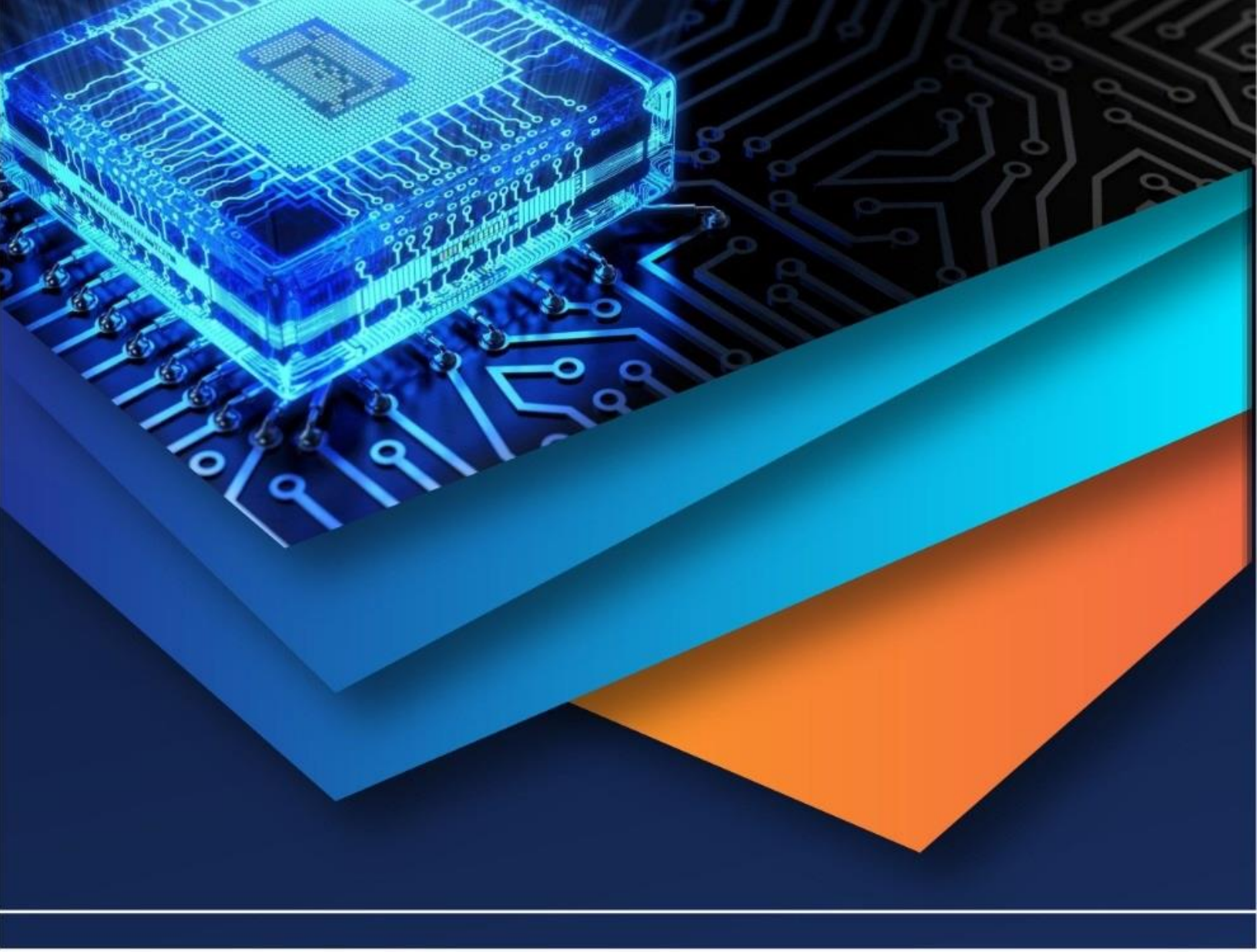
REFERENCES

[1] Guha & Garg, 2004] R. Guha and A. Garg, “Disambiguating People in Search,” Technical report, Stanford University, 2004.
[2] J. Artiles, J. Gonzalo, and F. Verdejo, “A Testbed for PeopleSearching Strategies in the WWW,” Proc. SIGIR ’05, pp. 569-570,2005.
[3] G\and D. Yarowsky, “Unsupervised Personal NameDisambiguation,” Proc. Conf. Computational Natural LanguageLearning (CoNLL ’03), pp. 33-40, 2003.

©IJRASET: All Rights are Reserved

[4] R. Bekkerman and A. McCallum, "Disambiguating Web Appearancesof People in a Social Network," Proc. Int'l World Wide WebConf. (WWW '05), pp. 463-470, 2005.
[5] P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04),2004.
[6] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H.Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An AdvancedSocial Network Extraction System," Proc. WWW '06, 2006.
[7] C. Galvez and F. Moya-Anegon, "Approximate Personal Name-Matching through Finite-State Graphs," J. Am. Soc. for InformationScience and Technology, vol. 58, pp. 1-17, 2007.
[8] T. Hokama and H. Kitagawa, "Extracting Mnemonic Names of People from the Web," Proc. Ninth Int'l Conf. Asian Digital Libraries(ICADL '06), pp. 121-130, 2006.
[9] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
[10] M. Muzammal and R. Raman, "On probabilistic models for uncertain sequential pattern mining," in Proc. ADMA'11, 2010, pp. 60–72.
[11] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," Proc.Conf. Empirical Methods in Natural Language (EMNLP '04), 2004.
[12] S. Sekine and J. Artiles, "Weps2 Evaluation Campaign: Overview of the Web People Search Attribute Extraction Task," Proc. Second Web People Search Evaluation Workshop (WePS '09) at 18th Int'l World Wide Web Conf., 2009.
[13] G. Salton and M. McGill, Introduction to Modern, Information Retreival. McGraw-Hill Inc., 1986.
[14] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, pp. 206-214, 1998.
[15] Chetna Chand, Amit Thakkar, Amit Ganatra- Sequential Pattern Mining: Survey and Current Research Challenges, International Journal.
[16] 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.l of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012
[17] R. R. Shelke , Dr. V. M. Thakare, Dr. R. V. Dharaskar, "Study of Data Mining Approach for Mobile Computing Environment", International Journal on Computer Science and Engineering(IJCSE) , ISSN : 0975-3397, Vol. 4, 12 Dec 2012, pp.1920-1923.
[18] R. R. Shelke , Dr. R. V. Dharaskar, Dr. V. M. Thakare, "Data Mining in Wireless Environment-An Overview", proc. Of National Conference on Innovative Paradigms in Engineering & Technology, S. B. Jain college Engineering and Technology, Nagpur, feb 2013, pp.362-366.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)