International Journal of Trend in Scientific Research and Development, Volume 1(4), ISSN: 2456-6470
www.ijtsrd.com

Efficient Way to Identify User Aware Rare Sequential Patterns in Document Streams

Swati V. Mengje
ME Student (CSE)
Department of Computer Science & Engineering
H.V.P.M's COET, Amravati University

Prof. Rajeshri R. Shelke
Assi. Professor (CSE), Ph.D (pursuing)
Department of Computer Science & Engineering
H.V.P.M's COET, Amravati University

ABSTRACT

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 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. Here present solutions to solve this innovative mining problem through three phases: pre-processing 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.

KEYWORDS: Web mining, sequential patterns, document streams, rare events, pattern-growth, dynamic programming.

INTRODUCTION

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 behaviour 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. 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. 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 [9].

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.

Sequential Topic Patterns:

Keeping in mind the end goal to describe client practices in distributed record streams, we think about on the connections among points extricated from these archives, particularly the successive relations, and indicate them as Sequential Topic Patterns (STPs). Each of them records the total and rehashed conduct of a client when she is distributing a progression of reports, and are appropriate for deriving clients' inborn qualities and mental statuses.

Initially, contrasted with individual themes, STPs catch both mixes and requests of subjects, so can serve well as discriminative units of semantic relationship among records in vague circumstances. Second, contrasted with report based examples, theme based examples contain dynamic data of archive substance and are along these lines helpful in grouping comparative records and discovering a few regularities about Internet clients. Third, the probabilistic depiction of points keeps up and gathers the instability level of individual themes, and can along these lines achieve high certainty level in example coordinating for questionable information.

User-aware Rare Sequential Topic Patterns:

For an archive stream, a few STPs may happen oftentimes and in this manner reflect normal practices of included clients. Past that, there may in any case exist some different examples which are all inclusive uncommon for the overall public, however happen generally frequently for some particular client or some particular gathering of clients. We call them User-mindful Rare STPs (URSTPs). Contrasted with successive ones, finding them is particularly intriguing and huge. Hypothetically, it characterizes another sort of examples for uncommon occasion mining, which can portray customized and unusual practices for extraordinary clients [7].

There is no standard way to define user interactions formally. Knowing the ways users act in a multi-user interaction environment, like social networking websites, helps to identify the behavior pattern. Behavioral pattern can be used to analyze the user's activities. Social networks offer several activities which baffles analyst to identify privacy issues, like how user's profile information should be protected from other clients through these activities, moreover, activities in social networking websites are not transparent [12]. This study defines the behavior pattern from the process mining perspective to detect user's abnormal behaviors. For achieving this goal, these actions should be followed: creating a user's activities log file (process mining techniques runs on log files), extracting process models, defining the normal model and detecting abnormal activities. The study has been substantiated by presenting a case study that includes real and syntactic data sets of Facebook. There are some tools fitting the proposed approach, such as ProM and CPN [15]. The comprehensive dataset is needed to evaluate the approach; we use both syntactic and real datasets. The data in the syntactic set are replicated intellectually by Color Petri Nets (CPN) tool. The syntactic dataset is a combination of possible behaviors of the users rooted from the real dataset. Conventionally, data mining tools are used to produce behavior pattern from datasets of social networking website[2].

There are two problems in this approach. First, databases normally are huge and the social network databases are always growing. Most machine learning algorithms run with difficulties whenever they encounter gigantic datasets. Second, databases are dynamic where databases are often exposed and can be altered many times, so data mining systems cannot perform fit classification.

The alternative technique is process mining. Pattern discovery is conducted by process mining; it produces an explicit process model that goes through event logs to make a compatible model for dynamic behavior (Aalst *et al*., 2012).

This study develops a control system for user's activities monitoring to detect threats. The system performs like an intrusion detection system. By logging activities and comparing the actions of the users against predefined model, the system can detect suspicious activities. Knowing the threat in such a dynamic domain is difficult, while, in the intrusion detection system, the discovered intrusion can be as

simple as system boundary security intruded by malicious users [8].

Normally, the security border is distinct in such a system. But, in social network websites, the margin between what the users intend to preserve and to share to the public is blurred [4]. Abnormal behaviors can be detected when compared to normal behaviors. Normal behavior needs norms. Norm recognition is difficult because of the immaturity of social networks and radical variation of users. Norms are changing as time passes by. The heart of the system is norm recognition. After knowing the norms, the abnormal activity can be identified by measuring the norm deviation of the users. There are three algorithms that are commonly used for anomaly detection in process mining.

First, the algorithm based on sampling makes a sample population from log file based on sampling factor. Then, it models M as the normal model. If a trace out of population can be parsed by M, then it is normal, otherwise, it is abnormal. The algorithm exposes some deficiencies. Assuming the sampling model is contaminated by abnormal traces, the M model does not represent norms. With regard to sampling factor, failure is possible as the sampling model can include abnormal traces in different numbers. It is time-consuming as the algorithm runs for each trace. In addition, the traces are labeled abnormal might be parsed partially. Therefore, the algorithm reduces the measurability of abnormal trace recognition since there is no way to determine the parsing ability.

Second, the algorithm based on threshold (Bezerra and Waine, 2013) divides the log file into two sections: frequent and infrequent. The frequent log file is more numerous than the infrequent one. By randomly picking from the infrequent set and inclusion cost of anomalous candidate measurement, algorithm ensures that the deviation never exceeds the threshold. By dividing a log in terms of frequency, contamination problems are partially alleviated. Inclusion cost is a metric for evaluating the differences of process models vertices. Third, the algorithm based on iteration is similar to the threshold algorithm, although it picks all anomalous traces once. It works with the threshold to categorize the traces into anomalous or normal. Therefore, it saves more time. There is function called select () that returns trace with highest inclusion cost.

LITERATURE SURVEY

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.

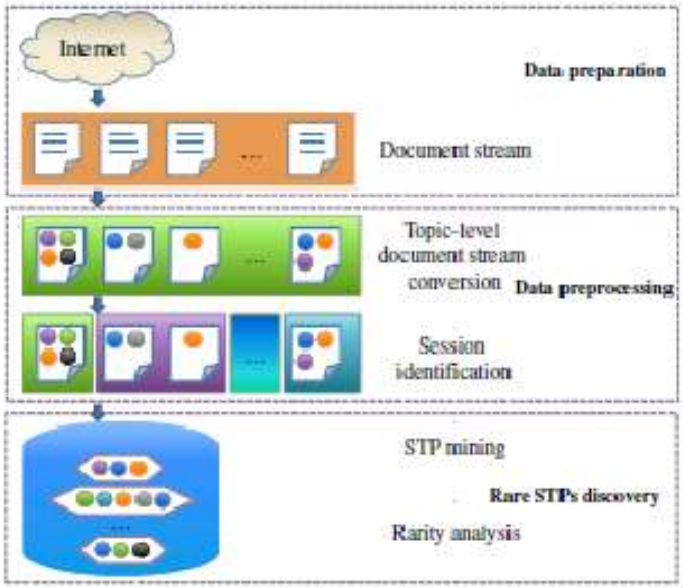

Fig.2. 1 Processing framework of URSTP mining.

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.2.1. 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 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. Textual documents made and disseminated on the Internet are constantly changing in different structures. The vast majority of existing works are given to subject demonstrating and the advancement of individual points, while consecutive relations of themes in progressive records distributed by a particular client are overlooked. The greater part of existing works investigated the development of individual themes to distinguish and foresee get-togethers and in addition client practices [22].

Sequential pattern mining was first introduced by Agrawal and Srikant [17] in the context of market basket analysis, which led to a number of other algorithms for frequent subsequence, including GSP [19], PrefixSpan [18] and SPAM [3]. Frequent sequence mining suffers from pattern explosion: a huge number of highly redundant frequent sequences are retrieved if the given minimum support threshold is too low. One way to address this is by mining frequent closed sequences, i.e., those that have no subsequence's with the same frequency, such as via the BIDE algorithm [21].

However, even mining frequent closed sequences does not fully resolve the problem of pattern explosion. We refer the interested reader of for a survey of frequent sequence mining algorithms[23].

In an attempt to tackle this problem, modern approaches to sequence mining have used the minimum description length (MDL) principle to find the set of sequences that best summarize the data. In fact, finding the most compressing sequence in the database is strongly related to the maximum tiling problem, i.e., finding the tile with largest area in a binary transaction database. SQS-Search (SQS) [20] also uses MDL to find the set of sequences that summarize the data best: a small set of informative sequences that achieve the best compression is mined directly from the database. SQS uses an encoding scheme that explicitly punishes gaps by assigning zero cost for encoding non-gaps and higher cost for encoding larger gaps between items in a pattern. While SQS can be very effective at mining informative patterns from text, it cannot handle interleaving patterns, unlike GoKrimp and ISM, which can be a significant drawback on certain datasets e.g. patterns generated by independent processes that may frequently overlap.

In related work, Mannila and Meek [15] proposed a generative model of sequences which finds partial orders that describe the ordering relationships between items in a sequence database. Sequences are generated by selecting a subset of items from a partial order with a learned inclusion probability and arranging them into a compatible random ordering. Unlike ISM, their model does not allow gaps in the generated sequences and each sequence is only generated from a single partial order, an unrealistic assumption in practice. There has also been some existing research on probabilistic models for sequences, especially using Markov models. Gwadera et al. use a variable order Markov model to identify statistically significant sequences [24].

PROPOSED WORK AND OBJECTIVES

The proposed method is outlined in Fig. 4.1 and having main component: pattern extraction, search the pattern and find occurrence of the words.

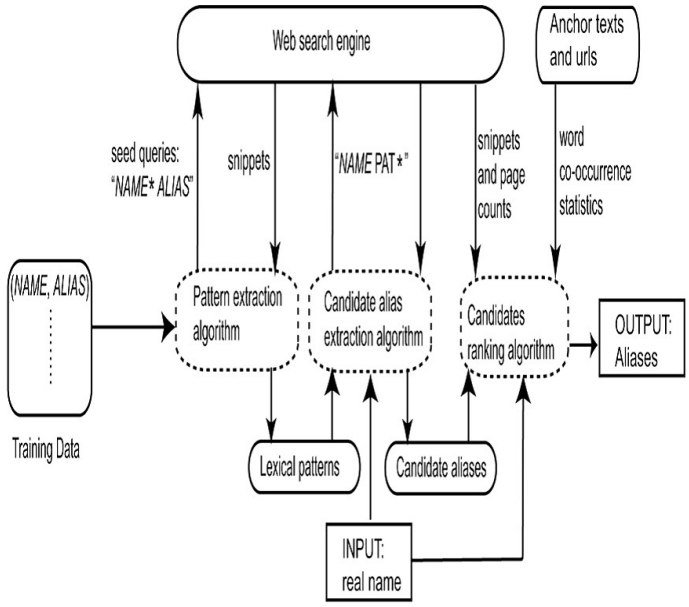

Fig. 4.1 Proposed Method

- The main objective to developed a system is to make easy for the user to search the file.
- By using pattern matching technique, we can able to access the file.
- Here we provide a facility to search the file by using topic search. Second is we are able to search by using description and also by using keyword.
- The file those contains some words like whose describe some abnormal behaviour, we separate that files.
- So that it makes easy to distinguish such files and efficiently we can retrieve the information.
- We are able to upload the text document, pdfs files. By matching the pattern it gives the count of the word that used in the document, it means it gives the occurrences of the word.

DESIRED IMPLICATIONS

It can be useful tool for analyzing the name of person from his alias name.

In this it can be predictable step for detection of any kind of cyber crime.

We will utilize both lexical patterns extracted from snippets retrieved from a web search engine as well as anchor texts and links in a web crawl. Lexical patterns can only be matched within the same document. In contrast, anchor texts can be used to identify aliases of names across documents. The use of lexical patterns and anchor texts, respectively, can be considered as an approximation of within document and cross-document alias references. By combining both lexical patterns based features and anchor text-based features, better performance in alias extraction will be achieved. It can only incorporate first order co-occurrences. An alias might not always uniquely identify a person. For example, the alias Bill is used to refer many individuals who has the first name William. The namesake disambiguation problem focuses on identifying the different individuals who have the same name. The existing namesake disambiguation algorithm assumes the real name of a person to be given and does not attempt to disambiguate people who are referred only by aliases. The knowledge of aliases is helpful to identify a particular person from his or her namesakes on the web. Aliases are one of the many attributes of a person that can be useful to identify that person on the web.

APPLICATIONS

A. Text Editor and Search Engines

Text editors, search engines and digital libraries need to perform pattern matching in a text or database. Most text editors use direct implementation of Boyer-Moore algorithm to implement find/replace command.

B. Computational Biology and Bioinformatics

String matching algorithms are widely used in DNA sequencing, finding close mutation, searching antimicrobial structures, developing local data warehouses for DNA, genes and proteins.

C. Network Intrusion Detection System

Modern intrusion and computer virus detection systems incorporate use of string matching algorithms of packets against signatures. Multiple patterns from a

virus database can be matched in parallel and compiled automaton stored for later use.

D. Musical Pattern Detection

Approximate string matching algorithms are used in modern music search technique to retrieve musical note from musical database.

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 behaviors 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. 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 future work.

In the future, we will refine the session identification process and the measures of user-related rarity, and improve the mining algorithms mainly on the degree of parallelism. Moreover, based on STPs, we will try to define more complex event patterns, such as imposing timing constraints on the sequential topics, and design corresponding mining algorithms. We are also interested in the dual problem, i.e., discovering patterns occurring frequent on the whole, but comparatively rare for specific users. What's more, we will apply our approach in more real-life mining problems and develop some practical tools.

Acknowledgment

I feel deeply indebted and thankful to all who opined for technical knowhow and helped in collection of market data also feel thankful to my guide Shelke mam and to all who directly and indirectly help me for and transactional information. A special thanks to my family members for constant support and motivation.

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. Mann and D. Yarowsky, "Unsupervised Personal NameDisambiguation," Proc. Conf. Computational Natural LanguageLearning (CoNLL '03), pp. 33-40, 2003.

[4] R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of 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 Advanced Social 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 Information Science 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] A. Bagga and B. Baldwin, "Entity-Based Cross-Document Coreferencing Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), pp. 79-85, 1998.

[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] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," Proc. Int'l Semantic Web Conf. (ISWC '05), 2005.

[13] 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.

[14] G. Salton and M. McGill, Introduction to Modern, Information Retreival. McGraw-Hill Inc., 1986.

[15] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, pp. 206-214, 1998.

[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. Agrawal and R. Srikant. Mining sequential patterns. In *ICDE*, pages 3–14, 1995.

[18] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *ICDE*, pages 0215–0215, 2001.

[19] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *EDBT*, pages 3–17, 1996

[20] N. Tatti and J. Vreeken. The long and the short of it: summarising event sequences with serial episodes. In *KDD*, pages 462–470, 2012.

[21] J. Wang and J. Han. BIDE: Efficient mining of frequent closed sequences. In *ICDE*, pages 79–90, 2004.

[22] C. Aggarwal and J. Han. *Frequent Pattern Mining*. Springer, 2014.

[23] H. T. Lam, F. Moerchen, D. Fradkin, and T. Calders. Mining compressing sequential patterns. *Statistical Analysis and Data Mining*, 7(1):34–52, 2014.

[24] H. Mannila and C. Meek. Global partial orders from sequential data. In *KDD*, pages 161–168, 2000.