RESEARCH ARTICLE OPEN ACCESS

Mining User-aware Rare STPs and Frequent Events from Social media

Salmath Amina KP [1], Farzin Ahammed T [2]
Department of Computer Science and Engineering
AWH Engineering College, KTU University
Kuttikkatoor, Kozhikode, India

ABSTRACT
Many users of social media publishing many posts over Internet through social media like Facebook and Twitter. They share different views about different stories each day. Sequential pattern mining is exploring idea for finding the most frequently using patterns from these data sets. 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 rare behaviors of Internet users, we propose Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (UaRSTPs) in social media on the Internet. They are rare on the whole but relatively frequent for specific users. We present a group of algorithms to solve this innovative mining problem through some phases: preprocessing to extract probabilistic topics and identify clusters for different users, generating all the STP candidates with support values, selecting UaRSTPs by making user-aware rarity analysis on derived STPs and extracting frequent events from twitter.

***Key words*:-** STP, UaRSTP, Pre-processing, Cluster, Events, Probabilistic topic

I. INTRODUCTION

This data mining is a computational process of discovering patterns in large. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The problem of mining sequential patterns over data is solved by this paper. A new algorithm for mining sequential patterns algorithm is especially efficient when the sequential patterns in the database are very long. Here we introducing a novel depth-first search strategy that integrates a depth-first traversal of the search space with effective pruning mechanisms.

Mining data streams for knowledge discovery is important to many applications, including Web click stream mining. The algorithm SWSS (Sequential pattern mining with the weighted sliding window model in SPAM) to mine frequent sequential patterns based on the weighted sliding windows model. This algorithm provides more space for users to specify which sequences they are more interested in.

Recently, data mining communities have focused on a new data model, where data arrives in the form of continuous streams. Real life datasets contain both frequent and infrequent items. The infrequent items can be called as rare items and these are of importance. Rare itemset problem occurs when the frequency of an item varies largely. For a very high minimum support, frequent patterns involving rare items are missed. So, a lower minimum support is set to find frequent patterns involving both rare and frequent patterns. But this may cause item explosions. So, we have proposed approaches to efficiently extract items.

Document streams are generated in various forms on the Internet, such as news streams, micro blog articles, instant messages, research paper archives, web forum discussion threads, and so forth. These document streams generally concentrate on specific topics.

Some STPs occur frequently in a document stream and thus reflect common behaviors of users. Besides, there are still some others which are rare for the general population, but occur relatively often for some specific user or some specific group of users. Compared to frequent ones, mining these user-related rare STPs are more interesting. Theoretically, it defines a new kind of patterns for event.

In order to characterize and detect personalized and abnormal behaviors of Internet users, propose Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (UaRSTPs) in document streams on the Internet. Each STP 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 user’s personalized and abnormal behaviors and characteristics.

II. RELATED WORKS

The problem of mining sequential patterns [1] over data is solved by this paper. Elements of a sequential pattern need not be simple items. The algorithm split the problem of mining sequential patterns into number of phases, Sort phase, Litemset phase, Transformation phase, Sequence phase, Maximal phase.

A new algorithm for mining sequential patterns algorithm is especially efficient when the sequential patterns in the database are very long. Here introducing a novel depth-first search strategy that integrates a depth-first traversal of the search space with effective pruning mechanisms.Finding sequential patterns in large transaction databases is an important data mining problem. The problem of mining sequential patterns and the support-confidence framework were originally proposed by Agrawal and Srikant.

Mining data streams for knowledge discovery is important to many applications, including Web click stream mining. There is a work, developed by Chuan XU, Y Chen and R. Bie used weighted sliding window. The algorithm SWSS [9](Sequential pattern mining with the weighted sliding window model in SPAM) to mine frequent sequential patterns based on the weighted sliding windows model. This algorithm provides more space for users to specify which sequences they are more interested in recently; data mining communities have focused on a new data model, where data arrives in the form of continuous streams. Many applications can generate great amount of data streams in real time, such as online transaction flows in retail chains, web click-streams in web applications, performance measurement in network monitoring, and ATM transaction records in banks, etc.

Tweet streams provide a variety of real-life and real-time information on social events that dynamically change over time. Although social event detection has been actively studied, how to efficiently monitor evolving events from continuous tweet streams remains open and challenging. One common approach for event detection from text streams is to use single-pass incremental clustering. As one of the most popular online social networking services, Twitter has been witnessing a burst of growth in the numbers of both users and posts. The quickly-updated tweets cover a wide variety of events that happen around the world every day. These events reveal valuable information on breaking news, hot discussions, public opinions, and so on. Moreover, these events are typically evolving over time. Microblogging services such as Twitter, Facebook, and Foursquare have become major sources for information about real-world events. Most approaches that aim at extracting event information from such sources typically use the temporal context of messages. However, exploiting the location information of georeferenced messages, too, is important to detect localized events, such as public events or emergency situations. Users posting messages that are close to the location of an event serve as human sensors to describe an event.

The correlations among topics extracted from the social media posts, 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 seriesparagraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

For a document stream, some STPs may occur frequently and thus reflect common behaviors of involved users. Beyond that, there may still exist some other patterns which are globally rare for the general population, but occur relatively often for some specific user or some specific group of users. We call them ***User-aware Rare STPs (UaRSTPs)***. Compared to frequent ones, discovering them is especially interesting and significant. Theoretically, it defines a new kind of patterns for rare event mining, which is able to characterize personalized and abnormal behaviors for special users.

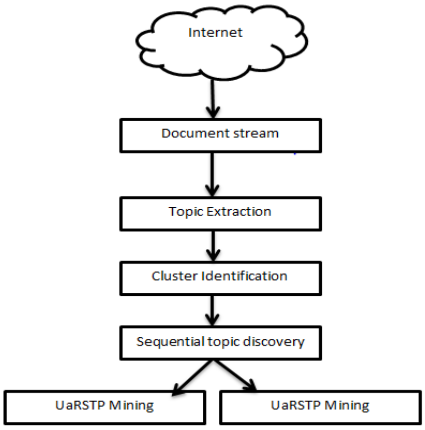
III. PROPOSED WORK

A. User-Aware Rare Sequential Topic Patterns

The term User-Aware Rare Sequential Topic Pattern mining means the topic which is globally rare but frequently used by a specific user. 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. Specifically, when Internet user's publish documents, the personalized behaviors characterized by STPs are generally not globally frequent but even rare, since they expose special and abnormal motivations of individual authors, as well as particular events having occurred to them in real life.

B. Mining UaRSTP and Frequent Events

In this section, we propose an approach to mining UaRSTPs in document streams. The main processing framework for the task is shown in below figure.



It consists of three phases. At first, textual documents are crawled from some social media sites, and these posts 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 clusters to identify complete user behaviors.

Finally and most importantly, we discover all the STPs associated to each user in the document stream for all users, and further pick out significant UaRSTPs associated to specific users by user-aware rarity analysis.it. From these sequential topics we are going to find out the frequent events. In order to fulfill this task, we design a group of algorithms. For event extraction from frequent sequential pattern, we are using database oriented method.

Table below gives the brief idea about the entire algorithm using in these method. The tweets and comments are getting by using APIs of an account. Stemming process means extracting the root words from comments. From these topic streams we are going to create various n-grams.

Procedure (URSTP Mining)

1. Find all messages from Twitter and Facebook using APIs ;
2. Stemming process.
3. N-Gram Creation.
4. Comment vector analysis.
5. Similarity score generation, based on N-gram.
6. Cluster creation.

Figure below shows the architecture of UaRSTP and Event miner. The tweets and comments from Twitter and Facebook respectively are getting by using API and store in our system. Each comments /tweets are extracted topic wise to compute its similarity among topics. The similarity score will be the key to mine frequent and rare sequential topic patterns. If topic similarity is higher, then the STP will be the frequent.

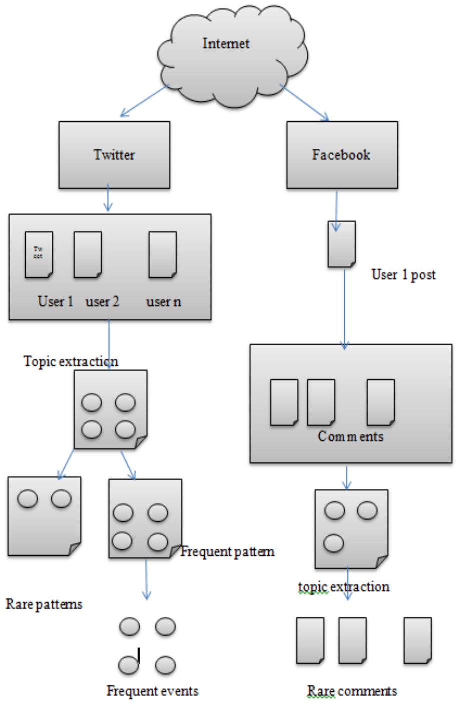


Fig 1 Architecture of proposed system

V. CONCLUSIONS

The Mining URSTPs (User-Aware Rare Sequential Topic Pattern) 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. The different approaches deal with the different algorithm and methods for detecting and mining sequential topic pattern.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: “original version

.

REFERENCES

[1] Rakesh Agarwal and Ramakrishnan Srikant (1995) Mining Sequential Pattern, *Data Mining*, , pp. 3–14

[2] R. Agrawal, T. Imielinski, and A. Swami (1993) Mining association rules between sets of items in large databases, *management of data*, 207-219.

[3] R. Agrawal and Srikant (1994) Fast Algorithm For Mining Association Rules, *Sequential Mining*, .
A. Califano and I. Rigoutsos (1993) A fast look- up algorithm for string homology. *Intelligence system for molecular biology*, (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[4] Jay Ayres, Johannes Gehrke and Tomi Yiu (2002) Sequential Pattern mining using a Bitmap Representation, *knowledge discovery and data mining mining.* 429-435

[5] R. J. Bayardo (1998) Efficiently mining long patterns from databases, 85-93

[6] D. Burdick, M. Calimlim, and J. Gehrke, (2001) Mafia: A maximal frequent itemset algorithm for transactional databases. Heidelberg, Germany, 2001.

[7] Chuan XU and Young CHEN (2009) Sequential Pattern Mining in data streams using the weigted sliding window, *knowledge mining.*3, 265-294

[8] Y. Zhu, and D. Shasha, (2002) , StartStream: Statistical monitoring of thousands of data streams in real time, *VLDB*,358-369.

[9] Dong. Guozhu and Pei. Jian, (2007), Sequence Data Mining, *Database System* , 33,119

[10] H cai, Z Huang and D Srivastava (2015) Indexing Evolving Events From tweet stream, *knowledge and engineering,10, 1041-4347*

[11] . J. Allan, R. Papka, and V. Lavrenko,(1998) , On-line new event detection and tracking, *SIGIR*, 37-45.