ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 3, Issue 12, December 2016

Survey on Mining Sequential Topic Patterns Methods

[1] Snehal R. Tanksale [2] Dr. A. B. Bagwan
[1][2] Dept. of Computer Engineering
Rajashri Shahu College of Engineering, Tathawade,
Pune, Savitribai Phule Pune University,
Pune, India.

Abstract :-- **Sequential Pattern mining belongs to data mining techniques. Such techniques can applied to a sequential database to find out the correlation among list of ordered text documents. It is also used to identify time related user behavior in database. This can be useful in various areas such as, medical records, marketing etc. This is very challenging task to find out the sequential patterns from list of documents created by online users on social networking sites. The text mining is a popular solution to retrieve the information from text documents. Till now number of useful sequential patterns mining techniques from text documents has been developed and improvements are still going to improve the efficiency. Previous approaches do not consider the time stamp attribute of dataset. But it should be consider to improve the accuracy and efficiency of extracted information. So this paper makes survey of some recent approaches that focus on the sequential topic pattern mining for user abnormal behavior detection, also compare them on the basis of technique used and their advantages and disadvantages. It will be further used to improve the user behavior detection systems.**

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

I. INTRODUCTION

Document streams are created in different forms on the Internet, for example, news streams, emails, microblog articles, instant messages, research paper archives, web forum discussion threads, et cetera. These document streams for the most part focus on particular topics. For instance, people in the same social group may discuss some common topics or talk about a public or private events on the web. So far, most of text mining research focused on finding topics in document streams. Topics can be extracted from the stream including both semantic and temporal data by different topic modeling techniques. Clearly, some correlations among these obtained topics in successive documents for a particular client, and these correlations could be described by Sequential Topic Patterns(STPs). Since catching both topic combinations and their requests, STPs serve well as discriminative units of semantic association in ambiguous circumstances. In addition, the abstract and probabilistic depiction of topics can help to solve the cold start issue and achieve high confidence level in pattern matching.

Some STPs happen frequently in a document stream also, in this manner reflect common behaviors of clients. Moreover, there are still some others which are uncommon for the general populace, however happen moderately regularly for some particular client or some particular group of clients. Contrasted with frequent ones, mining these user-related rare STPs is all the more fascinating. Hypothetically, it characterizes another kind of patterns for event mining, which can describe those individual and personalized behaviors in a certain context.

Exploiting advantage of these extracted topics in document streams, a large portion of existing works analyzed the development of individual topics to recognize and predict social events as well as client practices. However, few explores paid attention on the relationships among various topics appearing in progressive documents published by a particular client, so some hidden but significant data to reveal personalized behaviors has been neglected.

In order to describe client behaviors in published document streams, we examine on the relationships among topics extracted from these documents, particularly the sequential relations, and indicate them as Sequential Topic Patterns (STPs).

II. LITERATURE REVIEW

In this paper [1], in order to characterize and identify personalized and abnormal behaviors of Internet clients, authors propose Sequential Topic Patterns (STPs) and detail the issue of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet. They are uncommon on the whole but relatively frequent for particular clients, so can be applied in many real-life situations, for example, real-time monitoring on abnormal user behaviors. We exhibit a group of algorithms to solve this innovative mining issue through three stages: preprocessing to extract probabilistic topics and distinguish sessions for various clients, generating all the STP candidates with (expected) support values for every client by example development, and selecting URSTPs by making user-aware rarity examination on determined STPs.

In this paper [2], a personalized minimum support (P_minsup) threshold with client indicated minimum items or min_i is presented. The P_minsup is produced for every k-sequence by analyzing the overall support pattern distribution of the click stream information; while the min_i esteem gives the client the adaptability to gain control on the number of patterns to be created on the next k-sequence by utilizing the top min_i items. This methodology is then applied in the SPADE Algorithm utilizing vector cluster as an expansion from the previous strategy for utilizing relational database and pre-defined threshold.

As the web log [3] information is considered as complex and temporal, applying Sequential Pattern Mining strategy becomes a challenging task. The min sup threshold issue is highlighted - as a pattern is considered as frequent if it meets the predetermined min sup. If the min sup is high, few patterns are found else the mining procedure will be longer if excessively numerous patterns produced utilizing low min sup. The organization of web log information that makes consecutive occurring pages has made it hard to create frequent sequences. Additionally, as every client' behaviour is unique; utilizing one min sup esteem for all clients may influence the pattern generation. This research presented a personalized minimum support threshold for every web clients utilizing their Median item access (support) value to curb this issue. The pSPADE execution was the highest on the discovery of user’s origin and also interesting pattern discovery attribute.

The fundamental point [4] of this study is to investigate the patterns of client fragments' structural changes. Up to now, there has been no examination on this specific point. This is the main study that researches the effect of client dynamics on fragments' structural changes. This paper tries to build up a technique to depict and clarify this issue. Another strategy is proposed based on the clustering and sequential rule mining methods. Moreover, another definition and structure for finding recognizing sequential rules is developed. The proposed strategy is implemented on the client information of a telecommunication service provider.

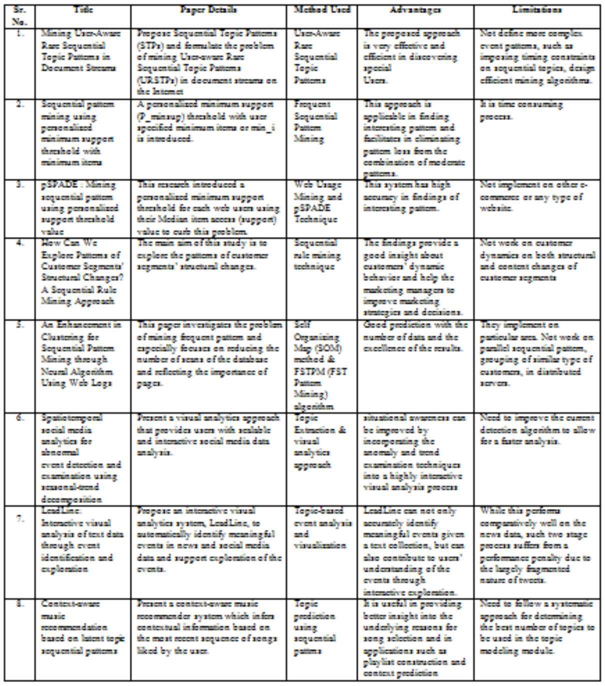
This paper [5] examines the issue of mining frequent pattern and particularly concentrates on decreasing the number of scans of the database and reflecting the significance of pages. In the present work a novel technique for pattern mining is introduced to take care of the issue through FSTSOM. In this Paper, the proposed technique is a change to the web log mining strategy and to the online navigational pattern forecasting. Here, Neural based approach ie. Self Organizing Map (SOM) is utilized for clustering of sessions as a pattern investigation. SOM relies on upon the clustering execution with the number of requests. In the proposed technique, utilizing the SOM algorithm for Frequent Sequential Traversal Pattern Mining called FSTSOM. In this technique, first utilizing SOM algorithm and getting some cluster of web-logs. At that point loading that web-log cluster, which is almost identified with frequent pattern. After that applying Min-Max Weight of Page in Sequential Traversal Pattern.

In this paper [6], authors display a visual analytics approach that gives clients scalable and interactive social media data analysis and visualization including the exploration and, examination of abnormal topics and events within different social media data sources, for example, Twitter, Flickr and YouTube. In order to find and understand abnormal events, the analyst can first extract major topics from a set of selected messages and rank them probabilistically utilizing Latent Dirichlet Allocation. He can then apply seasonal trend decomposition together with traditional control graph techniques to discover uncommon peaks and outliers within topic time series.

This paper [7] proposes an interactive visual analytics framework, LeadLine, to automatically distinguish important events in news and social media information and support investigation of the events. To describe events, LeadLine integrates topic modeling, event identification, and named entity recognition methods to automatically extract information regarding the investigative 4 Ws: who, what, when, and where for each event.

This paper [8] presents a context-aware music recommender framework which infers contextual data based on the most recent sequence of songs liked by the client. This methodology mines the top frequent tags for songs from social tagging Web sites and uses topic modeling to determine a set of latent topics for each song, speaking to various connections. Utilizing a database of human-compiled playlists, every playlist is mapped into a sequence of topics and frequent sequential patterns are found among these topics. These patterns represent frequent sequences of transitions between the latent topics representing contexts.

Table 1. Survey Table



Sr. No.	Title	Paper Details	Method Used	Advantages	Limitations
1.	Mining User-Aware Rare Sequential Topic Patterns in Document Streams	Propose Sequential Topic Patterns (STPs) and formulate the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet	User-Aware Rare Sequential Topic Patterns	The proposed approach is very effective and efficient in discovering special Users.	Not define more complex event patterns, such as imposing timing constraints on sequential topics, design efficient mining algorithms.
2.	Sequential pattern mining using personalized minimum support threshold with minimum items	A personalized minimum support (P_minsup) threshold with user specified minimum items or min_i is introduced.	Frequent Sequential Pattern Mining	This approach is applicable in finding interesting pattern and facilitates in eliminating pattern loss from the combination of moderate patterns.	It is time consuming process.
3.	pSPADE : Mining sequential pattern using personalized support threshold value	This research introduced a personalized minimum support threshold for each web users using their Median item access (support) value to curb this problem.	Web Usage Mining and pSPADE Technique	This system has high accuracy in findings of interesting pattern.	Not implement on other e-commerce or any type of website.
4.	How Can We Explore Patterns of Customer Segments' Structural Changes? A Sequential Rule Mining Approach	The main aim of this study is to explore the patterns of customer segments' structural changes.	Sequential rule mining technique	The findings provide a good insight about customers' dynamic behavior and help the marketing managers to improve marketing strategies and decisions.	Not work on customer dynamics on both structural and content changes of customer segments
5.	An Enhancement in Clustering for Sequential Pattern Mining through Neural Algorithm Using Web Logs	This paper investigates the problem of mining frequent pattern and especially focuses on reducing the number of scans of the database and reflecting the importance of pages.	Self Organizing Map (SOM) method & FSTPM (FST Pattern Mining) algorithm	Good prediction with the number of data and the excellence of the results.	They implement on particular area. Not work on parallel sequential pattern, grouping of similar type of customers, in distributed servers.
6.	Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition	Present a visual analytics approach that provides users with scalable and interactive social media data analysis.	Topic Extraction & visual analytics approach	situational awareness can be improved by incorporating the anomaly and trend examination techniques into a highly interactive visual analysis process	Need to improve the current detection algorithm to allow for a faster analysis.
7.	LeadLine: Interactive visual analysis of text data through event identification and exploration	Propose an interactive visual analytics system, LeadLine, to automatically identify meaningful events in news and social media data and support exploration of the events.	Topic-based event analysis and visualization	LeadLine can not only accurately identify meaningful events given a text collection, but can also contribute to users' understanding of the events through interactive exploration.	While this performs comparatively well on the news data, such two stage process suffers from a performance penalty due to the largely fragmented nature of tweets.
8.	Context-aware music recommendation based on latent topic sequential patterns	Present a context-aware music recommender system which infers contextual information based on the most recent sequence of songs liked by the user.	Topic prediction using sequential patterns	It is useful in providing better insight into the underlying reasons for song selection and in applications such as playlist construction and context prediction	Need to follow a systematic approach for determining the best number of topics to be used in the topic modeling module.

III. PROPOSE SYSTEM

A proposed system is implemented to solve the problem of mining URSTPs in document stream, to detect the abnormal behavior of users and design corresponding algorithms to support it. At first, preprocessing is used with heuristic methods for topic extraction and session identification. Then, borrowing the ideas of pattern-growth in uncertain environment, two alternative algorithms are designed to discover all the STP candidates with support values for each user. 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. To improve the user behavior identification, we will also use the user web search history and combine with twitter tweet stream document data. This system enhanced with topic aware recommendation to users, by using collaborative filtering approach.

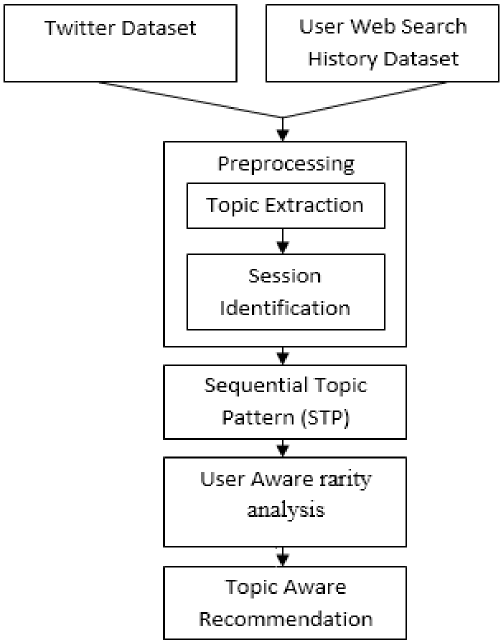

Twitter Dataset
User Web Search History Dataset
Preprocessing
Topic Extraction
Session Identification
Sequential Topic Pattern (STP)
User Aware rarity analysis
Topic Aware Recommendation

Fig 1. Propose System

IV. CONCLUSION

This paper presents the several recent approaches of recent sequential pattern mining, which is applied to some real time applications. From this survey we conclude that the timestamp management is the major issue in knowledge discovery process from real time datasets. This paper mainly focuses on the abnormal user behavior detection by using sequential topic pattern mining technique. The main aim of all these technique is to improve the accuracy of results.

REFERENCES

1. J. Zhu, K. Wang, Y. Wu, Z. Hu and H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1790-1804, July 1 2016.

2. S. Alias, M. N. Razali, Tan Soo Fun and M. S. Sainin, "Sequential pattern mining using personalized minimum support threshold with minimum items," 2011 International Conference on Research and Innovation in Information Systems, Kuala Lumpur, 2011, pp. 1-6.

3. S. Alias and N. M. Norwawi, "pSPADE: Mining sequential pattern using personalized support threshold value," 2008 International Symposium on Information Technology, Kuala Lumpur, Malaysia, 2008, pp. 1-8.

4. E. A. Z. Noughabi, A. Albadvi and B. H. Far, "How Can We Explore Patterns of Customer Segments' Structural Changes? A Sequential Rule Mining Approach," Information Reuse and Integration (IRI), 2015 IEEE International Conference on, San Francisco, CA, 2015, pp. 273-280.

5. S. Sahu, P. Saurabh and S. Rai, "An Enhancement in Clustering for Sequential Pattern Mining through Neural Algorithm Using Web Logs," Computational Intelligence and Communication Networks (CICN), 2014 International Conference on, Bhopal, 2014, pp. 758-764.

6. 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 VAST’12, 2012, pp. 143–152.

7. 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 VAST’12, 2012, pp. 93–102.

8. N. Hariri, B. Mobasher, and R. Burke, “Context-aware music recommendation based on latent topic sequential patterns,” in Proc. ACM RecSys’12, 2012, pp. 131–138.