ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Mining User-Aware Rare STPs in Document Streams

Mr. Somesh D. Kalaskar, Dr. Archana Lomte

M.E. Student, Dept. of Computer Engg., JSPM's BSIOTR, Wagholi, Pune, Maharashtra, India

Asst. Professor, Dept. of Computer Engg., JSPM's BSIOTR, Wagholi, Pune, Maharashtra, India

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

KEYWORDS: Data Mining, Document Streams, Rare Users, RSTPs, STPs, URSTPs.

I. 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 analysing 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 behaviour 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 foranother type of document streams, called browsed documentstreams, where Internet users behave as readers of documentsinstead of authors. In this case, STPs can characterize completebrowsing behaviours of readers, so compared to statisticalmethods, mining URSTPs can better discover special interestsand browsing habits of Internet users, and is thus capable togive effective and context-aware recommendation for them.While, this paper will concentrate on published documentstreams.

In order to find rare users from their published documentstreams, we study the correlations among topics extracted fromtheir document streams, especially the sequential relations, andspecify them as Sequential Topic Patterns (STPs). Some ofthese STPs are frequently common for all the users but thereare some patterns which are rare and infrequent. These RareSTPs (RSTPs) over the user-aware document streams constitutethe URSTPs which are used to find the rare users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

II. RELATED WORK

Topic mining in document collections has been extensivelystudied in the literature. The authors in [2] proposed RareSequential Topic Patterns over document stream. In this plaintext documents are produced and circulated over the Internetin dynamically changing form. It focuses on point displayingand disregarded the successive examples of topics in archivestream. Also, conventional consecutive example mining calculations basically centredon successive examples fordeterministic information sets and henceforth not appropriatefor document streams with topic uncertainty and uncommonexamples. Our work can be compared with this work as weare proposing the system which discovers the rare patterns socalled STPs in document streams.

Z. Zhao, D. Yan, W. Ng in their work has focused on probabilistic sequential pattern mining in large uncertain databases.Wireless sensors, GPS are large and uncertain databases wheredata is changing dynamically in huge contexts. The numbers of users are great in numbers across the world using suchGPS facility and all. The data in this database application ishenceforth very uncertain as it changes user to user instantaneously. The authors in[3],proposed probabilistic sequentialpattern mining in large uncertain database. The author uses*PrefixSpan* algorithm; the author derived two new forms as*U-PrefixSpan* for *p-SFS* mining and *UPrefixSpan* to avoidthe problem of possible world explosion. Algorithms can beverified by experiments on real and synthetic datasets.

X. Yan, J. Guo, Y. Lan, and X. Cheng have proposedmodel for short messages as bit term topic model (BTM)in their work [4]. This model is used to reveal topics insidethe short messages such as tweets and texts etc. instead ofcustomary topic models viz.LDA, PLSA. The authors foundthat BTM beats LDA in short text and ordinary writing.BTMunequivocally models the word co-event examples to improvethe theme learning. BTM utilizes the accumulated examplesas a part of the entire corpus for learning topics to take careof the issue of inadequate word co-event designs at documentlevel. We do broad examinations on real-world short contentaccumulations. The outcomes exhibit that our approach canfind more unmistakable and lucid topics, and fundamentally outflank standard techniques on a few assessment measurements. Moreover, we find that BTM can beat LDA even onordinary writings, demonstrating the potential consensus andmore extensive utilization of the new point show.

The sequential patterns for topics in context aware musicrecommendation system are proposed by the authors N. Hariri,B.Mobasher [5].In this topic set of each song is at first determined by a threshold on the topic probabilities obtained fromLDA. Then frequent topic-based sequential patterns occurringamong playlists are discovered to play next song. The songsare played in system according to user's context.The data generated with respect to some real time applications such as wireless sensors; moving object tracking etc. isdynamic. The authors in [6], author concentrates on exampledigging for dubious groupings and present incessant spatialpatterns with consecutive example with gap constraints. Suchexamples are essential for disclosure of learning given undetermined direction information. Author propose a dynamic programming approach for processing the recurrence likelihoodof these examples, which has direct time intricacy, and Authorinvestigate its inserting into example specification calculationsutilizing both broadness first pursuit and profundity first huntprocedures. Our broad experimental study demonstrates theproficiency and viability of our techniques for engineered andreal - world datasets.

C.H. Mooney, J.F. Roddick the authors [7] have proposedpattern mining for interval based events. They proposed*CTPrefixScan* algorithm for it. The interesting patterns aremined by applying multiple constraints on the events to getInteresting patterns and thereby topics. Sequence of events,things, or tokens happening in a requested metric space showup regularly in information and the necessity to identifyand dissect visit sub-sequences is a typical issue. ConsecutivePattern Mining emerged as a subfield of information miningto concentrate on this field.

III. SYSTEM ARCHITECTURE AND WORKFLOW

A. *System Architecture:*

In the proposed system, the users can sign up or sign in by entering their details. The system admin can manage the users' entries with their details and credentials in repository. In this context we are using textual files as document streams which the user can update and publish from its side. System admin can upload these to the server or database in encrypted form once the user has submitted this. The user of system can perform various search operations by using different search key attributes. The results are retrieved and displayed to the user accordingly. Among the search results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

the topics are extracted from the document streams published by specific users. These topics are used to determine the behaviour of users and if certain infrequent patterns observed it will then designate the rare users.

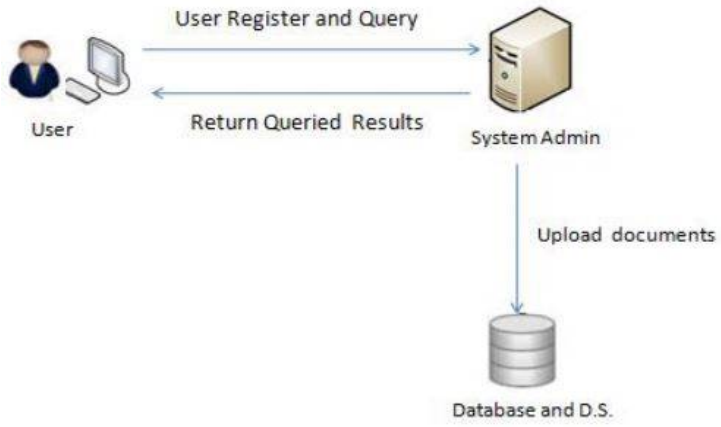

Fig. 1 System Architecture

B. *Mining Workflow :*

In our proposed work we have to mine RSTPs over user-aware document stream called URSTPs[1]. It involves mainlythree phases as document streams crawling, pre-processing totransform into topic level document stream and mining RSTPsover user-aware document streams.

The operations are stated below.

• **Document Stream Crawling:** It crawls the textual documents and act as input stream for topic extraction.

• **Topic Extraction:** In this we are pre-processing thecrawled document stream by certain algorithms to formthe topic level document stream.

•**Session Identification:** In this topic level documentstreams are mapped to different sessions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

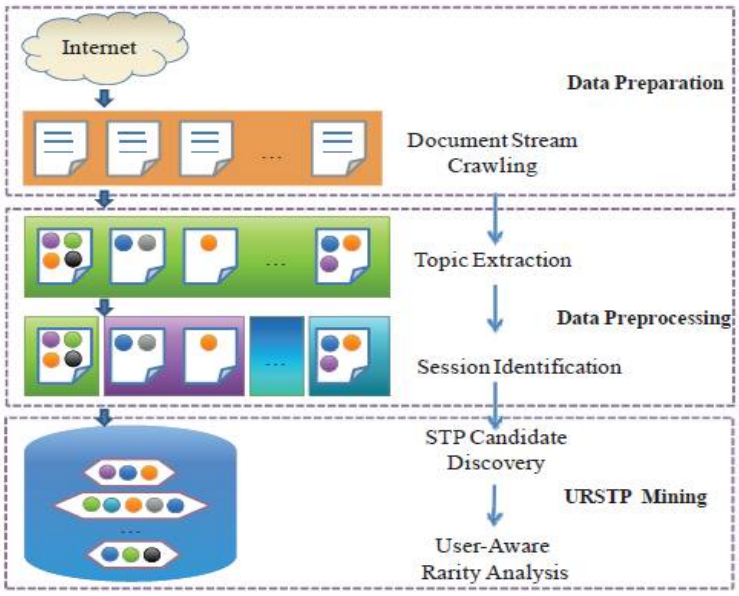

Fig 2. Mining Workflow

• **STP candidate discovery:** The sessions contain the topiclevel document streams for different users. In this step the STPs are identified for particular user.
• **RSTPs mining:** This step deals with RSTPs mining. These are very rare and infrequent patterns which are used to detect rare users.

IV. PSEUDO CODE

A. *System Operations:*

Step 1: Sign up or Sign in done by user by entering its details or credentials respectively.
Step 2: System admin module will update the entries fornew users.
Step 3: Users publishes document streams.
Step 4: System admin uploads document streams published by users.
Step 5: Users can view results by various search key attributes such as date, name etc.
Step 6: Users can find out rare users.

These steps shows the operations that system can perform from sign in or sign up to retrieving results according to users query. The results can be based on various attributes passed by the users and finally can return rare users.

B. *URSTP Mining:*

Step 1: Pre-process document streams.
Step 2: Extract Sequential Topic Patterns (STPs).
Step 3: Rarity analysis of STPs from derived sessions.
Step 4: Discover Rare Sequential Topic Patterns (RSTPs) from STPs.
Step 5: Identify rare users from RSTPs.

The above steps briefly show the mining workflow.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

V. RESULTS

For obtaining the results over published document streams the user has to log on to his personal accounts. Users can view results with various search key attributes e.g. name, date etc. and in addition can publish document streams. To determine the behaviour of user we need to keep track on its published document streams.

Table 1. Shows the results of displaying the users.

Table 1. Topics Extraction from Users Document Streams

Users	Top words extracted from STPs	Description
User1	game run team play football win lost toss	Sports
User2	hockey ground match hockey-stick	Sports
User3	movie cinema-hall ticket theatre	Entertainment
User4	YouTube download movie songs music	Entertainment
User5	attack kill gun fire	Terrorism
User6	online-fraud spoof hacking	Cybercrime

In table 1, we can see the result of operation, displaying users of the system; performed by both modules of our system i.e. user and system module. User name and top words extracted and the description of its document streams published is retrieved and returned to be displayed.

Here from the description column we can say that terrorism, cybercrime are rare patterns. User can perform various operations of search by using multiple search key attributes for searching the different topics and retrieving the users' details accordingly. We have skipped those result-sets as we are interested in finding the rare users.

Table 2. Rare Users

Users	RSTPs	Description
User5	attack kill gun fire	Terrorism
User6	online-fraud spoof hacking	Cybercrime

Table 2, shows the result of operation, finding the rare users which can be performed by both system admin module and user module.

A. *Advantages of Proposed System:*

1) Proposed system discovers the Rare Sequential Topic Patterns (RSTPs) from document streams of users.
2) System can identify rare users by URSTPs analysis.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

3) Proposed system is more time efficient than that of existing ones.

B. *Applications of Proposed System:*

1) 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.
2) This can be used to minimize cybercrimes by keeping track on abnormal and rare users.

C. *Analysis of Results:*

The figure below shows the time complexities of existing and proposed system. The graph is plotted time in second(s) vs. average number of sessions processed for discovering RSTPs and thereby detecting rare users.

The existing and proposed systems are analyzed using this time parameter and the average no. of sessions containing equal number of topics for this purpose. The number of sessions here are the average no. of sessions containing particular number of topics.

Initially we see that if averagenumbers of sessions are less the existing system take some more time than that of proposed system. The time difference to process average number of sessions remains somewhat constant with increase in size of average number of sessions.

Initially the time difference for same number of session is more. Then with increase in average number of sessions the time difference somewhat reduces; with respect to existing system and the proposed system. After some threshold value the time difference required to process average numbers of sessions for existing system and proposed system remains constant.

The experiment is performed on the session size with average numbers of sessions viz. 2, 4, 6, 8 ,10 etc. Time required to process them by existing system and proposed system is calculated. The graph is plotted below.

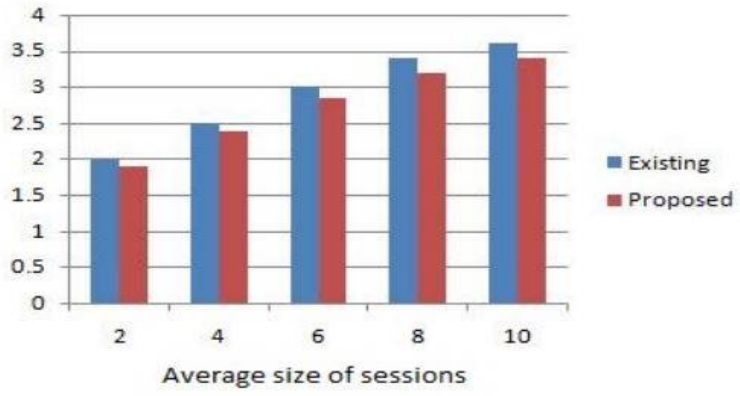

Fig 3. Time Complexities of Existing and Proposed System

VI. CONCLUSION AND FUTURE WORK

Knowledge discovery by various data mining techniques in documents streams is crucial. Topics are extracted in document streams and by topic modelling 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

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 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'sbehaviour by mining RSTPs over its browsed/surfed document streams and designing recommendation system.

REFERENCES

1. Jiaqi Zhu, Kaijun Wang, Yunkun Wu, Zongyi Hu, "Mining User Aware Rare Sequential Toppic Patterns in Document Streams", IEEE Transactions on Knowledge and Data Engineering,vol.28, no. 2, pp.1790-1804,2016.
2. Z. Hu, H.Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, " Discovery of rare sequential topic patterns in document stream", in Proc.SIAM SDM'14, pp. 533-541,2014.
3. Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in large uncertain databases", IEEE Trans. Knowledge Data Eng., vol. 26, no. 5, pp. 1171-1184, 2014.
4. X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts", in Proc. ACM WWW'13, pp. 1445-1456,2013.
5. N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latent topic sequential patterns", in Proc. ACM RecSys'12, pp. 131-138,2012.
6. Y. Li, J. Bailey, " L. Kulik, and J. Pei, Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases", in Proc. IEEE ICDM'13, pp. 448-457,2013.
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. T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Prob-abilistic frequent itemset mining in uncertain databases", in Proc. ACM SIGKDD'09, pp. 119-128,2009.
9. Somesh D. Kalaskar, Dr.Archana Lomte, "A survey on User-Aware STPs in Document Streams" ,International Journal of Innovative Re-search In Computer And Communication Engineering(IJIRCCE),vol. 4,no. 12,pp:21016-21021,2016.
10. 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.
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 SIGKDD'00, pp. 355-359,2000.

BIOGRAPHY

Mr. Somesh D.Kalaskar is pursuing M.E.at Computer Dept. JSPM's Bhivarabai Sawant Institute of Technology and Research,Wagholi, Pune Maharashtra, India and has completed B.Tech from Walchand College of Engineering Sangli. His area of interest is data mining He is working as an Assistant Professor at SSGMCE, Shegaon, Maharashatra,India.

Dr. Archana Lomte is an Assistant Professor at, Department of Computer Engineering, JSPM'sBhivarabaiSawant Institute of Technology and Research, Pune, Maharashtra, India.