Rare Sequential Topic Patterns in Document StreamS

Priya Surana[1], Khusboo Singh[2], Aishwarya Pandharkar[3], Ekta Wason[4], Sneha Vadali[5]
[1, 2, 3, 4, 5] Department of Compute Engineering
[1, 2, 3, 4 5] Pimpri Chinchwad College of Engineering

Abstract- *Facebook, LinkedIn and Twitter have been a crucial source of information for a wide spectrum of users. The network propagates popular information that is deemed important by the community. It is important to study the characteristics of the messages for a number of tasks, such as breaking news detection, personalized message recommendation, friends' recommendation, sentiment analysis and others. While many researchers wish to use standard text mining tools to understand messages on Twitter, the restricted length of those messages prevents them from being employed to their full potential. By studying how the models can be trained on the dataset we address the problem of using standard topic models in microblogging environments. We propose several schemes to train a standard topic model and compare their quality and effectiveness through a set of carefully designed experiments from both qualitative and quantitative perspectives. By training a topic model on aggregated messages we can obtain a higher quality of learned model which results in significantly better performance in to real world classification problems.*

Keywords- Web Mining, sequential Pattern, document streams, rare events, pattern growth.

I. INTRODUCTION

Document streams, such as emails, micro-blog articles, news streams, research paper archives, chatting messages, web forum discussions is contained by the Internet. The contents of these documents generally concentrate on some specific topics, which reflect offline social events and users' characteristics in real life. To mine these pieces of information, a lot of researches of text mining focused on extracting topics from document collections and document streams through various probabilistic topic models. Taking advantage of extracted topics in document streams, most of existing works analyzed the evolution of individual topics to detect and predict social events as well as user behaviors. However, few researches paid attention to the correlations among different topics appearing in successive documents published by a specific user, so some hidden but significant information to reveal personalized behaviors has been neglected. In order to characterize user behaviors in published document streams, we study on the correlations among topics extracted from these documents, especially the sequential relations, and specify them as Sequential Topic Patterns (STPs).

II. SEQUENTIAL TOPIC PATTERN

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. It is a special case of structured data mining.

There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, recovering missing sequence members, comparing sequences for similarity, and extracting the frequently occurring patterns. In general, sequence mining problems can be classified as string mining which is typically based on string processing algorithms and item set mining which is typically based on association rule learning

Each of them records the complete and repeated behavior of a user when she is publishing a series of documents, and are suitable for inferring users' intrinsic characteristics and psychological statuses. Firstly, compared to individual topics, STPs capture both combinations and orders of topics, so can serve well as discriminative units of semantic association among documents in ambiguous situations. Secondly, compared to document-based patterns, topic-based patterns contain abstract information of document contents and are thus beneficial in clustering similar documents and finding some regularities about Internet users. Thirdly, the probabilistic description of topics helps to maintain and accumulate the uncertainty degree of individual topics, and can thereby reach high confidence level in pattern matching for uncertain data. For a document stream, some STPs may occur frequently and thus reflect common behaviours 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 (URSTPs). 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 behaviours for special users. Practically, it can be applied in many real-life scenarios of user behaviour analysis.

Knowledge discovery in databases, has been recognized as a promising new area for database research also known as, data mining. This area can be defined as efficiently discovering interesting rules from large databases. A new data mining problem, discovering sequential patterns, was introduced. The input data is a set of sequences, called data-sequences. Each data sequence is a list of transactions, where each transaction is a sets of literals, called items. Typically there is a transaction-time associated with each transaction. A sequential pattern also consists of a list of sets of items. The problem is to find all sequential patterns with a user-specified minimum support, where the support of a sequential pattern is the percentage of data-sequences that contain the pattern.

1. RARE EVENTS

The events that occur with low frequency, or the term is often used in particular reference to hypothetical or infrequent events that have potentially widespread impact and which might sabotage society. Rare events encompass anthropogenic hazards (acts of terrorism, industrial accidents, financial and commodity market crashes, warfare and related forms of violent conflict, etc.), natural phenomena (tsunamis, solar flares, hurricanes, floods, asteroid impacts, earthquakes, etc.), , as well as phenomena for which natural and anthropogenic factors interact in complex ways (global warming-related changes in climate and weather, epidemic disease spread, etc.).

2. DOCUMENT STREAMS

A document stream is a sequence of data published by a user at time on a specific website. These document streams are available in various forms on the internet such as email blogs, internet forums etc. The co relation among the successive document streams by a single user can be made .These document stream can be represented as a topic-level document stream by extracting the topics from it. The co-relative topics can be defined by sequential topic patterns (STP). STPs reflect user's behavior which probably show repeated behaviors, their instance involve some subsequences related to a specific user during a certain time period.

Each of such proportion is called a session of the document stream, consists of a series of possibly correlated messages posted by a user during a time period on some micro-blog sites or Internet forums. Hence, in order to find significant STPs, a document stream should be divided into independent sessions in advance with the definition below.

For a specific user, the sessions should be disjoint and consecutive if they have multiple sessions. Many heuristic techniques can be applied to identify a session.

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 users' 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.

III. MINING URSTP

It consists of three phases. Firstly, textual documents are crawled from some micro-blog sites or forums, and constitute a document stream as the input of our approach. Then, preprocessing procedures, the original stream is transformed to a topic level document stream and then divided into many sessions to identify complete user behaviors. After preprocessing, we obtain a set of user-session pairs. 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. Next we will represent mining algorithms.

IV. ALGORITHMS

1. LDA

A generative statistical model that allows set of observations to be explained by unobserved groups that explain why some part of data are similar is Latent Dirichlet allocation (LDA). For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics that each words creation is attributable to one of the documents topic. There are two methods used in LDA such as preprocessing which is a data mining technique that involves transforming raw data into an understandable format. Stemming and stopping are main in preprocessing stage. Clustering is also used in LDA which is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters.

Given below is the figure showing preprocessing.

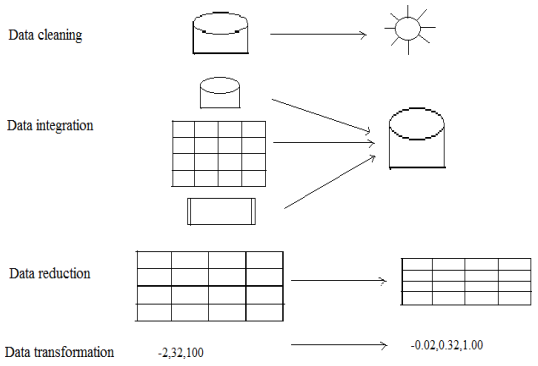

Figure 1. Preprocessing

2. SESSION IDENTIFICATION

Session definition is a Group of activities performed by a user from the moment he entered the website to the moment he left it. A set of user clicks usually referred to as a click stream, across web servers is defined as user session. Session identification is grouping the different activities of a single user. The process of segmenting the access log of each user into individual access session. Session identification uses time interval and time span heuristics to calculate session of the users. In the following, we give the corresponding algorithms designed for our mining task.

i) Time Interval Heuristics.
ii) Time Span Heuristics.

3. APPROXIMATION ALGORITHM

It is an idea to find and use an approximate value instead of solving the computation. That will make a good trade-off between the efficiency of the STP candidate discovery algorithm and the accuracy of the support values. We assume that each STP occurs in a session at most once. Which is reasonable to some extent since low-probability topics have been discarded from the uncertain topic-level sequence, and furthermore, the frequency of STP should be reflected by the multiple occurrences in different session.

V. USER AWARE RARITY ANALYSIS

After all the STP candidates for all users are discovered, we will make the user-aware rarity analysis to pick out URSTPs, which imply personalized, abnormal, and thus significant behaviors. It transforms the set of user-STP pairs into a set of user-URSTP pairs, with the set of user-session pairs and two thresholds, the scaled support threshold hss and the relative rarity threshold hrr, as input parameters. hss globally searches for users with same content. On the other hand hrr locally searches for user with same content.

VI. EXPERIMENTAL SETUP

We make use of a general dataset and a sports-related dataset To create the general dataset, we start from a user which is famous like “Steve Nash”, we check 150 latest tweets and select 50 random active friends and put them in a waiting queue. The activeness is determined by the total number of tweets (not less than 150) and friend number (not less than 50). This process is repeated for all the users in the queue. This helps us in removing the users with too high or too low tweets as well as very short and non-English tweets.

The special dataset is also obtained in the same manner, except that the seed user becomes a sports journalist. His friends are mostly connected to sports, such as journalists, players and commentators. To control the contents of tweet, we remove the users who are irrelevant by seeing their description in their profiles. Hence, the topic of tweets in dataset will focus on sports.

In pre-processing, we use a public package of Twitter-LDA model in Github, with topic number K=15and K=10, for the two datasets. In this model, each tweet talks about only one topic. It is found that 60% of tweets involve a unique topic, and others follow biased distribution. Afterwards, STP candidates for all users are discovered by calling sub procedure UpsSTP. Later, we apply URSTP Miner on these STPs to mine user-aware rare ones. Our target is to find special and abnormal behavior of users. In addition, we also use approximation algorithm.

1. QUALITY OF ASSOCIATED USERS

At first, we check if the mined URSTPs are really associated to the users with special or abnormal behaviors. It is very difficult to find the exact ground truth of these users for randomly crawled datasets. Here, we make an assumption that “verified” users in Twitter are more likely to have special and repeated behaviors than ordinary users, so they can be regarded as approximate ground truth of special users. But for the sports-related dataset, most of users are verified, and the user particularity is not obvious in a specific field, so the test here is only conducted on the general dataset. We mainly concern a small fraction of users with topmost relative rarity values, so recall is insignificant, especially when the ground truth is approximate. . Hence, we take precision @K as the evaluation metrics. For comparison, we also consider some alternative methods. The first one named URSTP-L is almost same as our approach except that LDA in the toolkit MALLET is used to directly extract probabilistic topics. In addition, as baseline methods to solve this innovative mining

problem, special users can be found by computing the relative rarity of a single topic z for a user u, instead of an STP. The topics are still extracted by Twitter-LDA or LDA, and the methods are named Topic-L and Topic respectively.

Table 1. Precision of the top most associated users for the general dataset

Precision	@5	@10	@15	@20	@30
URSTP	0.80	0.70	0.73	0.65	0.60
URSTP-L	0.80	0.60	0.67	0.60	0.53
TOPIC	0.60	0.50	0.53	0.55	0.47
TOPIC-L	0.20	0.30	0.33	0.25	0.27

Table 2. Examples of mined URSTPs

URSTP	USER	SCALED SUPPORT	RELATIVE RARITY
(Z^g_7,Z^g_{11})	U^g_{1299}	0.046	0.441
(Z^g_{12},Z^g_1)	U^g_{1914}	0.032	0.476
(Z^g_8,Z^g_{14})	U^g_{125}	0.024	0.318
(Z^g_{13},Z^g_2,Z^g_8)	U^g_{207}	0.029	0.340
(Z^g_{14},Z^g_1)	U^g_{1607}	0.043	0.559
$(Z^s_4,Z^s_5,Z^s_5,Z^s_6)$	U^s_{895}	0.025	0.343
(Z^s_2,Z^s_9)	U^s_{426}	0.031	0.332
(Z^s_9,Z^s_2)	U^s_{861}	0.047	0.502
(Z^s_6,Z^s_1)	U^s_{373}	0.033	0.334
(Z^s_1,Z^s_3,Z^s_7)	U^s_{875}	0.041	0.362

2. QUALITY OF MINED URSTPs

Besides the users associated to miner URSTPs, the quality of these patterns themselves also needs to be validated. They can reveal those special and abnormal behaviors on the Internet concretely, and should be self-interpretable and consistent with tweet contents. Similar as above, we mainly pay attention to the topmost URSTPs in terms of relative rarity. Specifically, for each mined URSTP, we firstly examine top words of involved topics from the topic model, summarize a rough description of each topic, and see whether an abstract and reasonable understanding can be obtained by integrating the meanings of ordered topics in the pattern. Afterwards, we check the profile of the user associated to the URSTP and the original tweets published by the user, to form a concrete understanding of her behaviors, and determine whether it is consistent with the abstract understanding above. Intuitive examples for this process from the two datasets are demonstrated here. The superscripts “g” and “s” on topics or users represent the results come from the general dataset or the special sports related dataset respectively.

Table 3. Top words and simple description of topics in table 2

TOPIC	TOP WORDS	DESCRIPTION
Z^g_1	People life black women support children love kids	family
Z^g_2	World hours production women goods things oil skin photo	product
Z^g_7	Game run team exercise play football strong week league win	sports
Z^g_{11}	Photo halloween fun happy space night water beautiful sun video	entertainment
Z^g_{12}	Watch tonight episode TV season movie series star excited favorite	TV show
Z^g_{13}	Love health body cool news life pretty skin enjoy great	health
Z^g_{14}	Day game weekend play happy team class win things amazing	(sports) game

VII. CONCLUSION

Mining URSTPs in document streams on the Internet is a challenging and significant 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. This paper contains several new concepts and the mining problem are ceremoniously defined, and a group of algorithms are designed and combined to analytically solve this problem. The experiments conducted on synthetic datasets demonstrate that the proposed approach is very efficient and effective in discovering special users as well as interesting and interpretable URSTPs from Internet document streams, which can well capture users’ abnormal and personalized characteristics and behaviors.

REFERENCES

[1] Jiaqi Zhu, Member, IEEE, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang, Member, IEEE, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams",2016

[2] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent pattern mining with uncertain data,” in Proc. ACM SIGKDD’09, 2009, pp. 29–38.

[3] R. Agrawal and R. Srikant, “Mining sequential patterns,” in Proc. IEEE ICDE’95, 1995, pp. 3–14.

[4] J. Allan, R. Papka, and V. Lavrenko, “On-line new event detection and tracking,” in Proc. ACM SIGIR’98, 1998, pp. 37–45.

[5] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, “Probabilistic frequent itemset mining in uncertain databases,” in Proc. ACM SIGKDD’09, 2009, pp. 119–

128.

[6] D. Blei and J. Lafferty, “Correlated topic models,” Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.

[7] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in Proc. ACM ICML’06, 2006, pp. 113–120.

[8] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.