International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

Efficient Mining of User-Aware Sequential Topic and Event Patterns in Time Series Data

P.Manvitha[1] G. Sunil Vijaya Kumar[2]

[1] PG scholar G. Pulla Reddy Engineering College, Kurnool, AP, India,

[2] Professor, Dept of Computer Science and Engineering, G.Pulla Reddy Engineering College, Kurnool, India,

Abstract

Recently, micro-blogs such as Twitter are attracting more and more attentions all over the world. Micro blog messages are real-time, spontaneous reports of what the users are feeling, thinking and doing, so reflect users' characteristics and statuses. However, the real intentions of users for publishing these messages are hard to reveal directly from individual messages but both content information and temporal relations of messages are required for analysis, especially for abnormal behaviors without prior knowledge.

If illegal behaviors are involved, detecting and monitoring them is particularly significant for social security surveillance. We can verify the activities of students to analyze the growth and the economy level to understand the behavior of student. For example, a student may opt for part time while studying to manage his economy needs this prevent the students financial problems and helps for better studies. Below are the various activities of student those are used for analysis.

(1).Higher Education (2).Training (3).Joblessness (4).Employment (5).Further education (6).School.

In this paper, we are proposing a system , an efficient mining of User-Aware Sequential Topic and Events Pattern for time series data The proposed system will be used for time series data sets directly connected to source micro-blog applications (like twitter and Facebook) so it fetches the real time streaming time series data It will be optimized for large data sets with parallel processing The proposed system will be used to get the event level information that can be used for identifying the abnormal events those are performed illegally

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

1. INTRODUCTION

There has been a lot of work in the field of data mining about pattern mining. The goal of pattern mining is to discover useful, novel and/or unexpected patterns in databases. Sequences are a very common type of data structures that can be found in many domains such as bioinformatics (DNA sequence), sequences of clicks on websites, the behavior of learners in e-learning, sequences of what customers buy in retail stores, sentences in a text, etc. 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 [1].

In this paper, we are proposing and efficient user aware sequential topic and event pattern in social blogs timelines data This paper will be used for time series data sets directly connected to source micro-blog applications (like twitter and facebook) so it fetches the real time streaming time series data. It will be optimized for large data sets with parallel processing and to get the event level information that can be used for identifying the abnormal events those are performed illegally Taking advantage of these 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 document streams, study on the correlations among topics extracted from these documents, particularly the sequential relations, and specify them as sequential Topic Patterns (STPs). every of them records the complete and recurrent behavior of a user once the user is tweeting a series of documents, and are appropriate for inferring users' intrinsic characteristics and psychological statuses. Firstly, compared to individual topics, STPs capture each mixtures and orders of topics, therefore will nurture as discriminative units of linguistics association among documents in ambiguous things. Secondly, compared to document-based patterns, topic-based patterns contain abstract data of document contents and area unit therefore helpful in cluster similar documents and finding some regularity regarding web users. Thirdly, the probabilistic description of topics helps to take care of and accumulate e uncertainty degree of individual topics, and might thereby reach high confidence level in pattern matching for unsure information.

For a document stream, some STPs might occur oftentimes and therefore mirror common behaviors of concerned users. on the far side that, there should still exist another patterns that area unit globally rare for the overall population, however occur comparatively typically for a few specific user or some specific cluster of users. We have a tendency to decision them User-aware Rare STPs (URSTPs). Compared to frequent ones, discovering them is very attention-grabbing and vital. in theory, it defines a brand new reasonably patterns for rare event mining, that is ready to characterize customized and abnormal behaviors for special users. Much, it are often applied in several real-life eventualities of user behavior analysis, as illustrated within the following example.

Context 1- Joblessness/Depressed (Real-time monitoring on abnormal user behaviors). Recently, micro-blogs like Twitter are attracting additional and additional attentions everywhere the globe. Micro-blog messages are period, spontaneous reports of what the users are feeling, thinking and doing, therefore replicate users' characteristics and statuses. However, the important intentions of users for business enterprise these messages are exhausting to reveal directly from individual messages, however each content data and temporal relations of messages are needed for analysis, particularly for abnormal behaviors while not previous data. What's additional, if unlawful behaviors are concerned, detective work and watching them is especially vital for Social Security police investigation. For example, the joblessness may lead for various disappointments which intern lead to frauds or illegal activities. Tweets like shown below will help to track a user's behavior based on his/her messages: [2]

(1).*I realized Joblessness was greatly impacting my moods and it was making me depressed.*
(2).*I felt worthless and like I was never going to succeed. What was the point? I stopped even looking at jobs.*
(3).*Maybe I can't do anything, maybe I'm generally stupid, this is going to keep happening to me*

The above shown tweets of STPs happen to be able to mix a series of inter-correlated messages, and might therefore capture such behaviors and associated users. moreover, though some black-market behaviors are rising, and their consecutive rules haven't been express however, we are able to still expose them by URSTPs, as long as they satisfy the properties of each international rarity and native frequentness that may be considered necessary clues for suspicion and can trigger targeted investigations. Therefore, mining URSTPs could be a sensible means that for period of time user behavior observation on the web.

It is price noting that the concepts on top of also are applicable for one more variety of document streams, known as browsed document streams, wherever web users behave as readers of documents rather than authors. during this case, STPs will characterize complete browsing behaviors of readers, thus compared to applied math ways, mining URSTPs will higher discover special interests and browsing habits of web users, and is therefore capable to administer effective and context-aware recommendation for them. While, this paper can target revealed document streams and leave the applications for recommendation to future work.

To solve this innovative and important drawback of mining URSTPs in document streams, several new technical challenges are raised and can be tackled during this paper. Firstly, the input of the task could be a matter stream, thus existing techniques of consecutive pattern mining for probabilistic databases can not be directly applied to unravel this drawback. A preprocessing part is important and crucial to urge abstract and probabilistic descriptions of documents by topic extraction, then to acknowledge complete and continual activities of web users by session identification. Secondly, visible of the period of time necessities in several applications, each the accuracy and also the potency of mining algorithms are necessary and will be taken into consideration, particularly for the likelihood computation method. Thirdly, completely different from frequent patterns, the user-aware rare pattern involved here could be a new idea and a proper criterion should be outlined, so it will effectively characterize most of customized and abnormal behaviors of web users, and might adapt to completely different application situations. And correspondingly, unsupervised mining algorithms for this type of rare patterns got to be designed in an exceedingly manner completely different from existing frequent pattern mining algorithms.

To sum up, this paper makes the subsequent contributions:

- To the simplest of our information, this can be the primary work that offers formal definitions of STPs likewise as their rarity measures, and puts forward the matter of mining URSTPs in document streams, so as to characterize and observe personalized and abnormal behaviors of web users;
- We propose a framework to pragmatically solve this drawback, and style corresponding algorithms to support it. At first, we have a tendency to offer preprocessing procedures with heuristic strategies for topic extraction and session identification. Then, borrowing the concepts of pattern-growth in unsure surroundings, 2 various algorithms area unit designed to get all the standard atmosphere candidates with support values for every user. that gives a trade-off between accuracy and potency. At last, we have a tendency to gift a user-aware rarity analysis algorithmic rule in keeping with the formally outlined criterion to select out URSTPs and associated users.
- We validate our approach by conducting experiments on each real and artificial datasets.

The rest of the paper is organized as follows. Section two reviews connected works as well as topic mining and successive pattern mining for settled and unsure databases. In Section three, we have a tendency to offer the key definitions associated with STPs, and formulate the matter of mining URSTPs in document streams. The process framework, preprocessing algorithms and mining algorithms area unit given thoroughly in Section four. Section five shows the experimental results on real Twitter datasets, and leaves the artificial results to Appendix. Section half-dozen concludes the paper and discusses future directions.

2. RELATED WORK

Analytics industry is all concerning getting the “Information” from the data. With the growing quantity of data in recent years, that too largely unstructured, it’s tough to get the relevant and desired info. But, technology has developed some powerful strategies which may be accustomed mine through the information and fetch the knowledge that we have a tendency to square measure yearning for. One such technique within the field of text mining is Topic Modeling because the name suggests, it's a method to mechanically establish topics gift during a text object and to derive hidden patterns exhibited by a text corpus.

Topic Modeling is totally different from rule-based text mining approaches that use regular expressions or wordbook primarily based keyword looking out techniques. it's associate unattended approach used for locating and observant the bunch of words (called “topics”) in massive clusters of texts. Topics will be outlined as “a repeating pattern of co-occurring terms during a corpus”. a good topic model ought to lead to – “health”, “doctor”, “patient”, “hospital” for a topic – healthcare, and “farm”, “crops”, “wheat” for a topic – “Farming”. Topic Models are very helpful for the purpose for document clustering, organizing massive blocks of textual data, info retrieval from unstructured text and have selection. as an example – New York Times are victimization topic models to spice up their user – article recommendation engines. Numerous professionals square measure victimization topic models for recruitment industries wherever they aim to extract latent options of job descriptions and map them to right candidates. they're being employed to prepare massive datasets of emails, client reviews, and user social media profiles. There are many approaches for obtaining topics from a text such as – Term Frequency and Inverse Document Frequency. Non Negative Matrix Factorization techniques. Latent Dirichlet Allocation is the most popular topic modeling technique and in this article, we will discuss the same

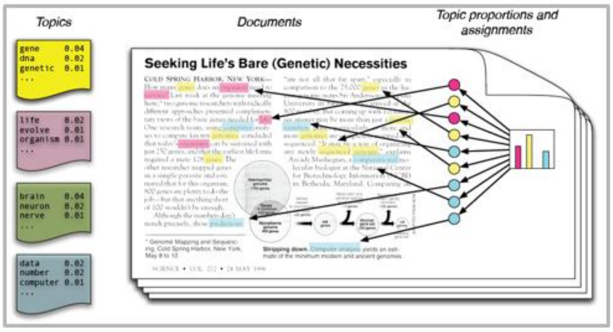


Fig. 1 : Topic Patterns

.

3. PROBLEM DEFINITION

In this section, we give some preliminary notations, define several key concepts related to STPs, and formulate the problem of mining URSTPs to be handled in this paper.[3]

3.1. Definitions

Definition 1(Document): A textual document d in a document collection D consists of a bag of words from a fixed vocabulary V = {w1, w2, • • • , w|V |}. It can be represented as {c(d, w)}w∈V , where c(d, w) denotes the occurrence number of the word w in d.

Given a document collection D and a topic number K, latent topics of these documents can be learnt through probabilistic topic models like LDA [7] and Twitter-LDA [39], and comprise a set T . Each topic is defined as follows.

Definition 2 (Topic): A semantically coherent topic z in the text collection D is represented by a probabilistic distribution of words in the given vocabulary V . It is denoted
Pas {p(w|z)}w∈V , which satisfies w∈V p(w|z) = 1.

In this way, each document can be represented by a probabilistic mixture (proportion) of these K independent topics, which form a structured topic-level document.

Definition 3 (Topic-Level Document): Given an original document d ∈ D and a topic set T , the corresponding topic-level document tdd is defined as a set of topic-probability pairs, in the form of tdd = {(z, p(z|d))}z∈T . Here, the probabilities are obtained through some topic P model and satisfy z∈T p(z|d) = 1. The superscript d can be omitted when the original document is not cared. Actually, we can select some representative topics from T to approximately describe the document, which will be discussed in the preprocessing procedure in the next section.

3.2. Sequential Topic Patterns

On the net, the documents are created and distributed in a very ordered method and so compose varied types of revealed document streams for specific websites. In this paper, we have a tendency to abbreviate them as document streams.

Definition 4 (Document Stream): A document
Stream is defined as sequence DS={(d1, u1, t1), (d2, u2, t2), • • • , (dN , uN , tN)}, where d_i(i = 1, . . . , N) is a document published by user u_i at time t_i on a specific website, and ti ≤ tj for all i ≤ j.
Usually, one user cannot write two documents simultaneously, so we can assume that at any time point, for any specific user, at most one document is published. Formally, if ti = tj , then ui 6=uj always hold.

Obviously, each document stream can be transformed into a topic-level document stream of the form T DS = {(td1, u1, t1), (td2, u2, t2), • • • , (tdN , uN , tN)}, by extracting topics for each document according to Definition 3.

In this paper, we pay attention to the correlations among successive documents published by the same user in a document stream. A kind of fundamental but important correlations is the sequential relation among topics of these documents, which can be defined by sequential topic patterns, and abbreviated as STPs. They are suitable to characterize users' complete and personalized behaviors when publishing documents in a website.

An important data mining problem is to design algorithm for discovering hidden patterns in sequences. There have been a lot of research on this topic in the field of data mining and various algorithms have been proposed.

One should note that the subsequence s′ is not necessarily contiguous in s, but just the order needs to be retained. In addition, when analyzing the pattern instances in a session, the user component and the time component of documents are irrelevant, so we can omit them in the representation of a session. Two examples of sessions in this form are given in Table 1. For the STP hz1, z3i, it has one instance in s1 with probability 0.5 × 0.4 = 0.2, and two instances in s2 with probabilities 0.3 and 0.12, respectively.

Table 1: Examples of Sessions in Topic-level Document Streams

ID	Sequences
Seq1	‹ {a,b},{c},{f}, {g},{e} ›
Seq2	‹ {a,d},{c},{b}, {a,b,e,f} ›
Seq3	‹ {a},{b},{f},{e} ›
Seq4	‹ {b},{f, g, h} ›

We can see that in a specific session, there may be multiple pattern instances for an STP, each with a respective occurrence probability. These instances may be not independent with each other, as some common topics extracted from common documents are involved. Therefore, the probability that an STP α occurs in s, denoted as P r(α ⊑ s), cannot be obtained by simply adding up the probability of each pat-tern instance. For example, the probability P r(hz1, z3i ⊑ s2) is not 0.3 + 0.12 = 0.42, but should be calculated as 0.6 × (0.5 + (1 − 0.5) × 0.2) = 0.36. The idea of dynamic programming can be utilized here, which will be presented in the next section.

3.3. User-Aware Rare Sequential Topic Patterns

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. Therefore, the STPs we would like to mine for user behavior analysis on the Internet should be distinguishing features of involved users, and thus satisfy the following two conditions:

1) They should be globally rare for all sessions involving all users of a document stream;

2) They should be locally and relatively frequent for the sessions associated with a specific user.

Next, we will formally specify this kind of STPs step by step, starting with the classical concept of support to describe the frequency. For deterministic sequential pattern mining, the support of a pattern α is defined as the number or pro-portion of the sequences containing α in the target database, but inapplicable for uncertain sequences like topic-level document streams. Instead, the expected support is appropriate to measure the frequency on uncertain sequences, and can be computed by summing up the occurrence probabilities of α in all sequences . In other words, it expresses the expected number of sequences containing α. To measure the frequency of STPs, we modify it a little to record the proportion of sessions where α occurs also in terms of expectation, via dividing the summation by the number of sessions. That is necessary because the session number here is no longer a constant when we consider both the global frequency and the local frequency of α for different users. For simplicity, this measure is still denoted as support instead of expected support in this paper.

Definition 6 (Support of STP): Given a session set S = {s1, s2, • • • , s|S|} as the database and an n-STP α = hz1, z2, • • • , zni, the support of α in S is defined as

$$supp(\alpha, S) \mathrel{\hat{=}} \frac{\sum_{i=1}^{|S|} Pr(\alpha \sqsubseteq s_i)}{|S|} \qquad (1)$$

If the document stream is fixed with no ambiguity, the second parameter T DS can be omitted from the notations.

Since the lengths of significant STPs are unknown, we have to consider all the possible STPs with different lengths and compute their support values. Notice that for an n-STP, the probability of each instance is the product of n probability values (decimals), and its support can be approximately thought as a combination (summation) of these products, so longer STPs tend to have lower support values. In order to fairly compare among the supports of STPs with different lengths, we should normalize their values to the same order of magnitude, i.e., a single probability value, by correspondingly taking the n-th root. That will reduce the impact of pattern lengths in the later comparison and unified computation. To this end, we further define scaled support as follows.

Definition 7 (Scaled Support of STP): Given the support supp(α, S) for an n-STP α, the scaled support scsupp(α, S) is defined as the n-th root of the support value. Formally,

scsupp(α, S) =b supp(α, S) N (2)

This formula is applicable not only for the general definition on a session set, but also for the global and local variants on a (possibly defaulted) document stream, such as

scsupp(α)|∗ and scsupp(α)|u. (3)

Based on the scaled support, we can evaluate these STPs in terms of their abilities in characterizing personalized and abnormal behaviors of Internet users, and pick out some significant and representative ones. Firstly, each selected STP should be linked to a specific user, so can be called a user-aware STP. Secondly, it should reflect the particularity on frequency, not only at user level but also at pattern level. According to these ideas, we define two new measures,

$$\overset{\beta\in\Phi}{RR(\alpha)|u = b\ AR(\alpha)|u - \quad U} \qquad (4)$$

where Φu is the set containing all the discovered STPs for the user u, i.e., for all β ∈Φu, supp(β)|u > 0 holds. Compared to where the average value is taken among all STPs for any user, this measure is fully user-specific and thus more accurate to reflect the particularity of the STP with respect to the user.

Now, we can define the core concept of this paper, user-aware rare STPs, abbreviated as URSTPs.

Definition 8 (User-Aware Rare STP). Given a topic-level document stream T DS, a scaled support threshold hss, and a relative rarity threshold hrr, an STP α is called a User-aware Rare STP (URSTP) if and only if both scsupp(α) ≤ hss and RR(α)|u ≥ hrr hold for some user u.

Here, the first condition indicates the global rareness of α, and the second one assures its relatively high frequency for the specific user u. That is consistent with the requirement of our mining task proposed at the beginning of this subsection. Based on these definitions, the problem of mining URSTPs in this paper can be formulated as follows:

Given a (published) document stream on the Internet, discover all URSTPs and associated users, which character-ize users' personalized and abnormal behaviors.

4. MINING URSTP

In this section, we propose a novel approach to mining URSTPs in document streams. The main processing frame-work for the task is shown in Fig. 2. 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.[4]

In order to fulfill this task, we design a group of algorithms. To unify the notations, many variables are denoted and stored in the key-value form. For example, User_Session represents the set of user-session pairs, and each of its elements is denoted as hu : Sui, in which the user u is the key of the map and its value Su is a set containing all the sessions associated with u. All the structures of such sets and simple description of topics are summarized in Table 2.

The workflow of our approach is presented in Fig. 2, and Algorithm 1 gives the pseudo-code of the main procedure. The input includes an original document stream DS = h(d1, u1, t1), (d2, u2, t2), • • • , (dN , uN , tN)i, a scaled support threshold hss and a relative rarity threshold hrr. As discussed later, there are still some thresholds used in preprocessing procedures, but since preprocessing strategies will be chosen with some common rules according to the characteristics of the input stream, we think preprocessing as a separate and independent module, and thus do not regard the thresholds defined there as the input parameters of the whole mining problem.

Algorithm 1. Main(DS, hss, hrr)

```
User_Session ← Preprocess(DS);
U ser ST P ← ∅;
for all (u : Su)∈User_Session do
Start a new thread;
ST P Suppu ← UpsSTP(∅, Su, ∅, Su);

User_ST P ← User_ST P ∪ {(u : ST P_Suppu)};
User_U  RST  P    ←EfficientURSTPMiner(User  ST P, User_Session,hss, hrr );
return  U ser U RST P ;
```

After preprocessing, we obtain a set of user-session pairs. For each of them with a specific user u, a new thread is started and a pattern-growth based subprocedureUpsSTP is recursively invoked to find all the STP candidates for u, paired with their support values, and add the combined user-STP pair to the set User STP . These threads can be executed in parallel relying on the hardware environment. When all of them finish, another subprocedure Efficient URSTP Miner will be called to make user-aware rarity analysis for these STPs together and get the output set User URSTP , which contains all the pairs of users and their corresponding URSTPs with values of relative rarity. Next, we will present the preprocessing procedures and the mining algorithms in detail.

4.1. Data Preprocessing

4.1.1. Topic Extraction

Twitter topics are extracted using Twitter R packages as explained below.

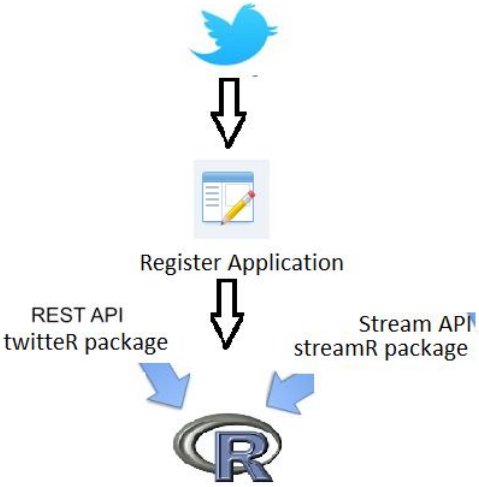


Figure 2: Data Preprocessing

The Twitter Streaming API, one of 3 such APIs (search, streaming, “firehose“), gives developers (and data scientists!) access to multiple types of streams (public, user, site), with the difference that the streaming API collects data in real-time (as opposed to the search API, which retrieves past tweets).

Algorithm 2. Partition(T DSu, hts)

```
j, k ← 1;
for i = 1 to N do
if ti − ti−1 >hti then
sj ← h(tdk, u, tk), • • • , (tdi−1, u, ti−1)i;
j ← j + 1;
k ← i;
sj ← h(tdk, u, tk), • • • , (tdN , u, tN )i;
m ← j;
return Su = {s1, s2, • • • , sm};
```

Intuitively, the primary heuristic with unfixed durations is a lot of appropriate for our drawback than the other. Additionally, these two heuristics don't need further data from the texts of documents, and presume that the behaviors of all users manifested in an exceedingly document stream adjust to a unified time-oriented rule, therefore neglect completely different business characteristics of users. on the far side that, some web-sites enable users to make hyperlinks among revealed documents, therefore during this case, it's attainable to seek out a lot of correct and user-specific partitions if users very manufacture these links to point complete behaviors, however this policy is restricted by the perform of the web site and therefore the business habits of users. Therefore, for a selected document stream, we want to decide on associate degree acceptable partition algorithmic rule to form a possible and relatively correct session identification.[5]

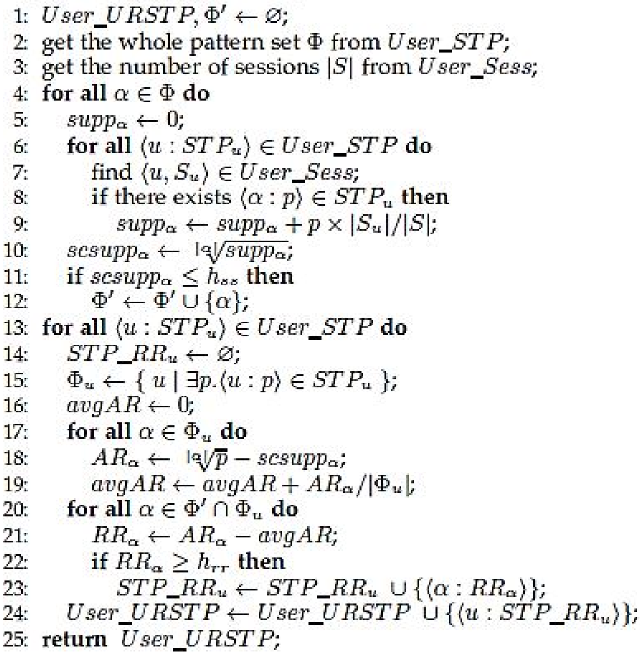
4.2. STP Candidate Discovery by Pattern-Growth

The sub procedure of STP candidate discovery is executed in parallel for each user. It aims to find all STPs occurring in the document stream associated with a specific user, paired with the (expected) support values of these STPs. According to Equation 1, the key step is to compute P r(α ⊑si), which is the probability that an STP α occurs in a session si belonging to the user. In this subsection, we at first present a DP-based algorithm to derive all STPs for the user and exactly compute the support values of them. Then, in order to improve the efficiency of our approach, we also give an approximation algorithm to estimate the support values for all STPs. Both algorithms are designed in the manner of pattern-growth.

4.3. 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. That is implemented by the sub procedure Efficient URSTP Miner shown in Algorithm 2. 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. At first, we get the set _ containing all the derived STPs for all users (line 2), and for each of them (denoted as _), compute the global support supp_ as a weighted average of its local support for each user (lines 5-9), and normalize it to a scaled value scsupp_ according to Equation 2 (line 10).

Algorithm 3
EfficientURSTPMiner(User_STP,User_Session, h_{ss}, h_{rr})



```
1: $User\_URSTP, \Phi' \leftarrow \varnothing$;
2: get the whole pattern set $\Phi$ from $User\_STP$;
3: get the number of sessions $|S|$ from $User\_Sess$;
4: for all $\alpha \in \Phi$ do
5:     $supp_\alpha \leftarrow 0$;
6:     for all $\langle u : STP_u \rangle \in User\_STP$ do
7:         find $\langle u, S_u \rangle \in User\_Sess$;
8:         if there exists $\langle \alpha : p \rangle \in STP_u$ then
9:             $supp_\alpha \leftarrow supp_\alpha + p \times |S_u|/|S|$;
10:    $scsupp_\alpha \leftarrow \sqrt[|q|]{supp_\alpha}$;
11:    if $scsupp_\alpha \leq h_{ss}$ then
12:        $\Phi' \leftarrow \Phi' \cup \{\alpha\}$;
13: for all $\langle u : STP_u \rangle \in User\_STP$ do
14:    $STP\_RR_u \leftarrow \varnothing$;
15:    $\Phi_u \leftarrow \{ u \mid \exists p.\langle u : p \rangle \in STP_u \}$;
16:    $avgAR \leftarrow 0$;
17:    for all $\alpha \in \Phi_u$ do
18:        $AR_\alpha \leftarrow \sqrt[|q|]{p} - scsupp_\alpha$;
19:        $avgAR \leftarrow avgAR + AR_\alpha/|\Phi_u|$;
20:    for all $\alpha \in \Phi' \cap \Phi_u$ do
21:        $RR_\alpha \leftarrow AR_\alpha - avgAR$;
22:        if $RR_\alpha \geq h_{rr}$ then
23:            $STP\_RR_u \leftarrow STP\_RR_u \cup \{\langle \alpha : RR_\alpha \rangle\}$;
24:    $User\_URSTP \leftarrow User\_URSTP \cup \{\langle u : STP\_RR_u \rangle\}$;
25: return $User\_URSTP$;
```

If the STP is globally rare checked by the threshold hss, it will berecorded in a set _′⊆ _ (lines 11-12). After that, for eachuser u, we calculate firstly the absolute rarity AR_ for all of her STPs according to Equation 3 as well as the average value for the user (lines 15-19), and then the relative rarity RR_ for those STPs globally rare and found for u accordingto Equation 4 (lines 20-21). Next, the locally frequent STPs are selected by the threshold hrr, each of which forms an STP-RR pair for u (lines 22-23). At last, the set of these pairs together with the key value u is added to the set of user-URSTP pairs (line 24), which will be returned when all the users have been handled (line 25).

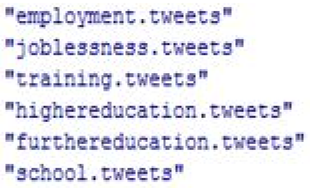
5. EXPERIMENTS

Since the problem of mining URSTPs in document streams proposed during this paper is innovative, there are not any different complete and comparable approaches for this task because the baseline, however the effectiveness of our approach in discovering customized and abnormal behaviors, particularly the reasonability of the URSTP definition, has to be much valid. During this section, we have a tendency to conduct attention-grabbing and informative experiments on message streams in Twitter datasets, to indicate that the majority of users discovered by our approach are literally special in reality, and also the mined URSTPs will so capture customized and abnormal behaviors of web users in a visible means.

In addition, we have a tendency to additionally valuate the potency of the approach on artificial datasets, and compare the two various subprocedures of STP candidate discovery to demonstrate the trade-off between accuracy and potency. As these results don't serve directly for the real-world mining task, they're left to the Appendix.

5.1. Experimental Setup

We got the random data from 15,000 users and 9,20,121 tweets which contains various sequence patterns as shown below



```
"employment.tweets"
"joblessness.tweets"
"training.tweets"
"highereducation.tweets"
"furthereducation.tweets"
"school.tweets"
```

In the preprocessing phase, we use a public package of the Twitter-LDA model in Github developed by the SMU Text Mining Group, with the topic number K = 16 and K = 13, respectively for the two datasets. In this model, each tweet is assumed to talk about only one topic, so each derived topic-level document just contains a single topic with probability 1. Although later computations are still feasible, the uncertainty degree is totally lost. We re-cover it by recording the topic values of each tweet at 10 iteration points after the burn-in period (900 iterations), and computing the proportion of them to get probabilistic topics. We find 45% of tweets involve a unique topic, and others follow biased distributions. That implies convergence and coincides with the characteristics of short tweets. Then, the Topic Probability Threshold with value 0.25 is adopted to select representative topics, and sessions are identified through the Time Interval Heuristics with the threshold set to 5 hours.

Afterwards, STP candidates for all users are discovered by calling the sub procedure Ups STP with five parallel threads. Here, we restrict the STP length in between 2 and 4, as longer STPs are generally insignificant and hard to interpret. Then, we apply Efficient URSTP Miner on these STPs to mine user-aware rare ones. Notice that our target is to find special and abnormal behaviors of Internet users, which are intuitively in minority for the general population, so the effectiveness of our approach should be reflected by the quality of those URSTPs with topmost values of the relative rarity, as well as their associated users. To this end, we set hss = 0.03 and hrr = 0.29 to get relatively strong conditions, and evaluate on a small but representative result set.

In addition, we also use the approximation algorithm UpsSTP- to replace UpsSTP, and carry out the two steps of mining for comparison. Very similar results of URSTPs are obtained and omitted here due to the page limit.

All the experiments were conducted on a desktop with Intel(R) 3GHz i5 CPU and 8GB RAM. The algorithms are implemented in R scripting, and run in command line R studio in Windows7.

5.2. Quality of Mined URSTPs

Besides the users associated to mined 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 their behaviors, and determine whether it is consistent with the abstract understanding above.[6]

Table 2: Top Words and Simple Description of Topics.

Topic	Top Words	Descriptio n
TP1	New class sessions start this week at @Training Buddies in Livonia	Study
TP2	Saturday Football and support the children who never get picked @training today #dropped and follow us	Joblessne ss

TP3	@seahawksPR @training camp this close to @DangeRussWilson bout to get an autograph when he sees my phone cover oops	School
TP4	What's \"secret sauce\" of this small college's success? @highereducation	higheredu cation
TP5	If you not talking bout work @school don't even talk to me !! don't got time to be playing games with irrelevant people	School
TP6	Well done @school of environment @uniofbrighton. We did a good job of making our curriculum future proof	school
TP7	Great things happening this morning Palm Lane Elem. @CartwrightSD W/The Mission Continues volunteers @school staff…	school
TP8	@SCHOOL @TEACHERS @SUBJECTS PLEASE BE GOOD TO ME TOMORROW, IDK WHAT I'LL DO IF SOMETHING BAD OR ANNOYING HAPPEN.	school
TP9	@zenyonces @school hopefully it does, i'm starting to become less scared of high school but still, i'm scared af.	school

Intuitive examples for this process from the two datasets are demonstrated here. Table 2 shows a part of top-10 URSTPs in the respective datasets, their associated users, scaled supports and relative rarities, and lists some top words as well as simple descriptions of involved topics. From the results of the general dataset, we can infer some personalized characteristics and abnormal missions of Twitter users. For the first two users who are verified, HumpDestiny is studying school but he badly needs to go out to watch movie; while kiera_battle is happy with school studies and willing to attend classes without fail.. As to ordinary users, SE0KJ1NN13 is likely to say bye because he is tired of schooling decided not to go to school again.

To make quantitative analysis, 10 volunteer students are divided into two groups, rating for the 30 topmost URSTPs respectively on their self-interpretability and consistency, with the values between 1 (worst) and 5 (best). The average scores of URSTPs for different top positions from the two datasets ("−g" and "−s" respectively) are recorded. The results indicate our approach can indeed capture personalized behaviors of Internet users and express them in an understandable way, especially for those topmost results. In addition, the patterns from the general dataset are easier to understand, but once interpreted, the results from the sports-related dataset are more consistent with tweet contents. The reason may be that domain knowledge is required to understand patterns in specialized datasets, whereas meaningful patterns there tend to coincide specialized but not diversified behaviors of Internet users.

Synthetic Experimental Results:

As the commencement, we have a tendency to imagine some users and make some sessions for each of them. every session may be a sequence of item sets directly obtained from the generator, wherever every itemset will be thought to be a document and every item represents a subject. These topics area unit divided into 2 varieties, common topics and special topics. For eightieth of users, we have a tendency to solely assign common topics, whereas for the opposite 2 hundredth, each styles of topics area unit concerned, and every special topic is simply for one user. These special topics lead to globally rare, however domestically frequent STPs, which may be observed manually and function the bottom truth of URSTPs to be strip-mined by our approach. Then, we have a tendency to assign a chance to every topic occurring within the dataset following a standardized distribution over (0, 1), and normalize these values to ensure that the summation for every document is a smaller amount than one, controlled by a dummy topic in every document. The obtained sets of element-level probabilistic sequences change to the topic-level document stream outlined during this paper. Notice that the stream has already been divided into sessions, therefore the preprocessing section isn't needed here, and also the take a look at can so target the mining algorithms.

Initially, the user variety is about to fifty, the quantity of sessions for every user is picked from a statistical distribution with the mean m¯ = a hundred, the scale of every session (length of sequences) is additionally drawn from a statistical distribution with the mean q¯ = five. All the reportable results area unit averaged on 5 runs for every such as configuration explained by the usage of the most pattern instance chance rather than the precise pattern incidence chance, that ends up in larger variations among STPs' support values, at each user level and pattern level.

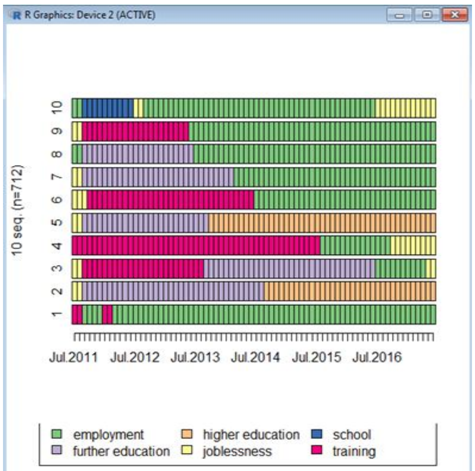


Figure 3: Sequence index plot: Visualization Of Individual Sequences of employment, higher education, school, further education, joblessness, training.

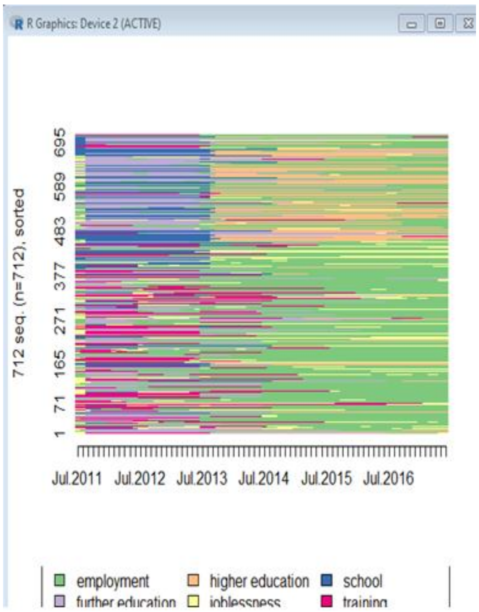


Figure 4: Sequence index plot: Visualization Of Individual Sequences change with timestamp of employment, higher education, school, further education, joblessness, training.

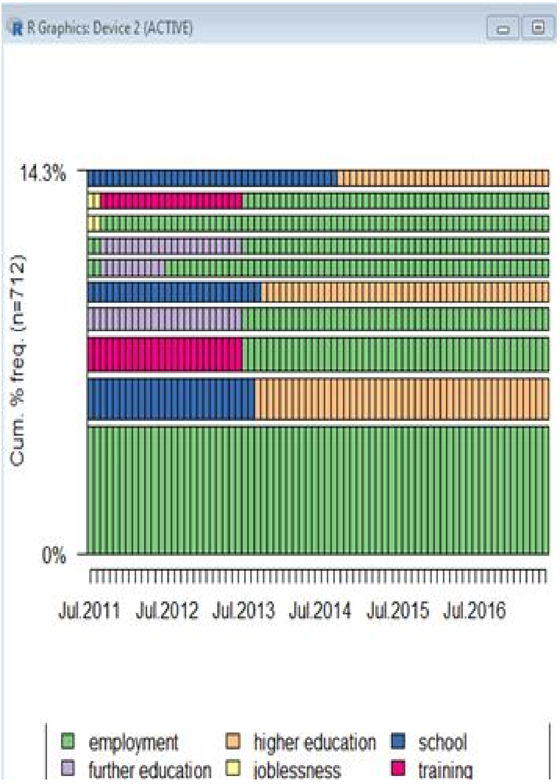


Figure 5: The Frequency Of Sequences of employment, higher education, school, further education, joblessness, training.

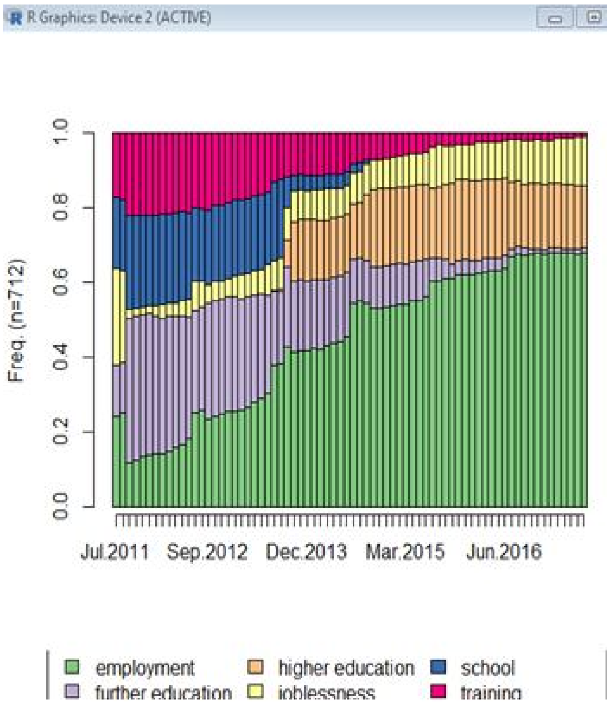


Figure 6: The Sequence Of State Distributions of employment, higher education, school, further education, joblessness, training.

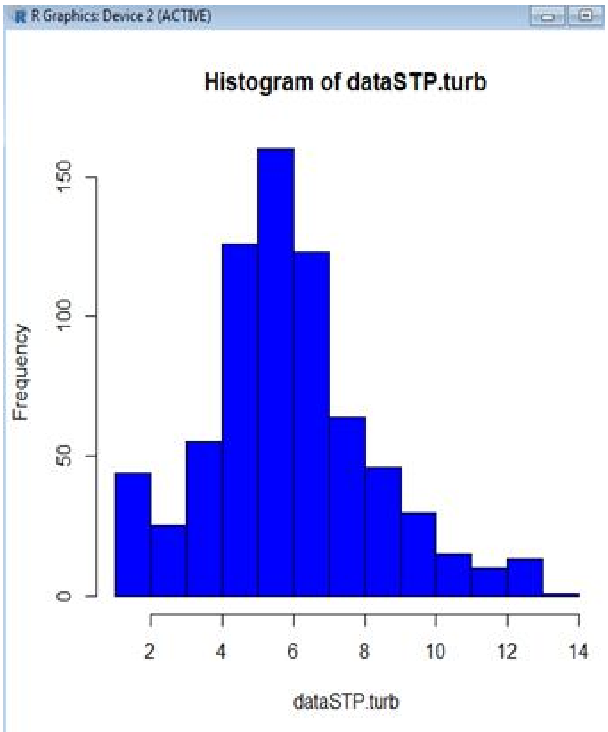


Figure 7: Computation of sequence turbulence for employment, higher education, school, further education, joblessness, training.

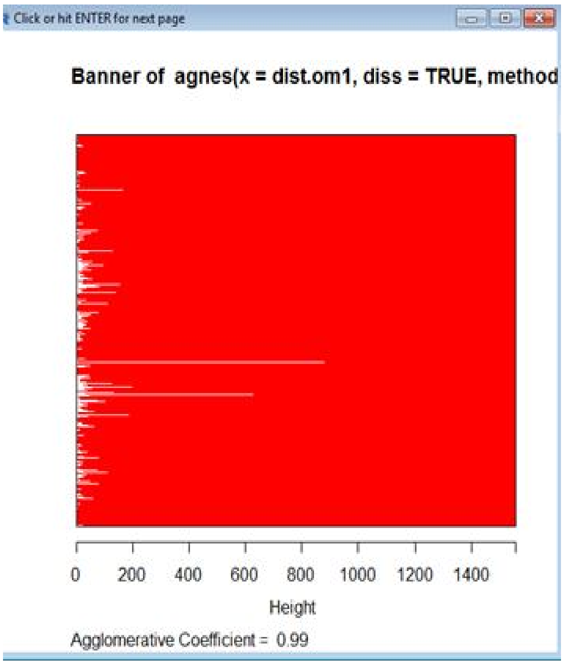


Figure 8: Hierarchical clustering of a sequences for employment, higher education, school, further education, joblessness, training.

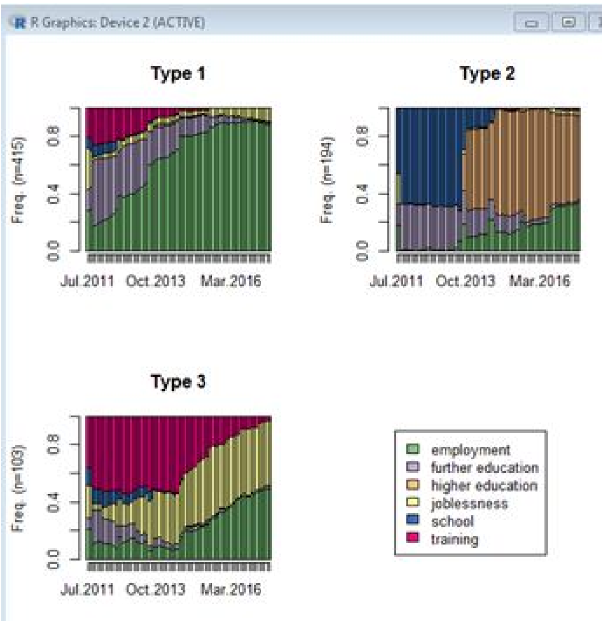


Figure 9: Calculation of categorical distribution for employment, higher education, school, further education, joblessness, training.

Taking these values as thresholds, we analyze on precision, recall and F1-measure with different user numbers, and compare the two algorithms. For the exact mining, the precision varies between 0.90 and 0.98 and the recall varies between 0.86 and 0.95. They are both high and thus compelling. In addition, as the number of users increases from 40, recall shows an upward trend, while precision maintains a high value but declines moderately. The reason is that the patterns will become sparser with more users, and the discrepancy among users will get more obvious, so some insignificant but relatively rare STPs will also be found. As a trade-off metric, F1-measure is comparatively stable. The results for the approximation mining is a little worse, especially for precision and larger user scales, but still acceptable. The F1-measure is around 0.8 and tends toward stability.

Taking these values as thresholds, we have a tendency to analyze on preciseness, recall and F1-measure with totally different user numbers, and compare the 2 algorithms. For the precise mining, the preciseness varies between zero.90 and 0.98 and also the recall varies between zero.86 and 0.95. they're each high and so compelling. additionally, because the variety of users will increase from forty, recall shows associate upward trend, whereas preciseness maintains a high price however declines moderately. the rationale is that the patterns can become sparser with a lot of users, and also the discrepancy among users can get a lot of obvious, thus some insignificant however comparatively rare STPs will be found. As a trade-off metric, F1-measure is relatively stable. The results for the approximation mining may be a very little worse, particularly for preciseness and bigger user scales, however still acceptable. The F1-measure is around zero.8 and tends toward stability.

For the primary case, we have a tendency to fix q¯ = five, however modification m¯ for every user from a hundred to 30000. For the second case, we have a tendency to fix m¯ = a hundred, however modification q¯ from four to eighteen. With the info size increasing, the performances of each UpsSTP and UpsSTP-a area unit virtually stable, whereas for the Apriori-based algorithms, they refuse sharply at m¯ = 6000 and q¯ = ten, and area unit severally terribly sensitive to at least one parameter.

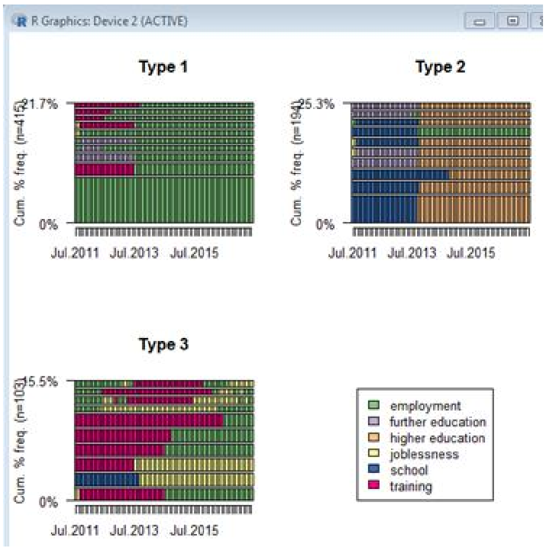


Figure 10: Calculation of categorical frequency for employment, higher education, school, further education, joblessness, training.

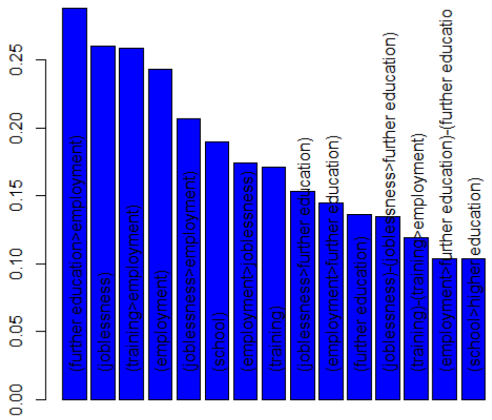


Figure 11: Frequent subsequences of employment, higher education, school, further education, joblessness, training with minimal support =0.05

To make a finer comparison between our 2 algorithms, we have a tendency to amplify the lower elements of the 2 charts and show them on the proper of several figures. It is discovered that the approximation algorithmic rule is so a trifle quicker, particularly for larger scales. Notice that every execution of this subprocedure is simply for one user, thus once the user variety will increase, the time distinction for the total approach can become a lot of and a lot of evident, even with some extent of correspondence. Therefore, we are able to conclude that the 2 algorithms have their several benefits that one is acceptable for the $64000 task reflects a trade-off between mining accuracy and swiftness, and will rely upon the particular needs of application eventualities.

6. CONCLUSION

Mining URSTPs in published document streams on the web could be a important and difficult problem. It formulates a replacement quite complicated event patterns supported document topics, and has wide potential application eventualities, like period observation on abnormal behaviors of net users. During this paper, many new ideas and also the mining drawback area unit formally outlined, and a bunch of algorithms area unit designed and combined to consistently solve this drawback. The experiments conducted on each real (Twitter) and artificial datasets demonstrate that the projected approach is extremely effective and economical in discovering special users further as attention-grabbing and explicable URSTPs from net document streams, which may well capture users' personalized and abnormal behaviors and characteristics.

As this paper puts forward associate degree innovative analysis direction on net data processing, a lot of work is engineered thereon within the future. At first, the matter and also the approach also can be applied in different fields and eventualities, particularly for browsed document streams we are able to regard readers of documents as personalized users and create context-aware recommendation for them. Also, we are going to refine the measures of user-aware rarity to accommodate totally different necessities, improve the mining algorithms primarily on the degree of correspondence, and study on-the-fly algorithms aiming at period document streams. Moreover, supported STPs, we are going to attempt to outline additional complicated event patterns, like imposing temporal order constraints on consecutive topics, and style corresponding economical mining algorithms. we tend to also are inquisitive about the twin drawback, i.e., discovering STPs occurring often on the total, however comparatively rare for specific users. What's additional, we are going to develop some sensible tools for real-life tasks of user behavior analysis on the web.

References

[1]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.

[2]R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc. IEEE ICDE'95, 1995, pp. 3–14.

[3]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.

[4]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, 2014, pp. 533–541.

[5]N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," ACM Comput. Surv., vol. 43, no. 1,pp. 3:1–3:41, 2010.

[6]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.

Author

P. Manvitha received B.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University, Anantapur.
Currently an M.Tech student in Computer Science and Engineering in G.PullaReddy Engineering College, Kurnool.

Dr. G. Sunil Vijaya Kumar received B.Tech. from SVUCE Tirupathi. And M.Tech in Computer Science from SC&SS JNU, Delhi and Ph.D in Computer Science from SOCIS,IGNOU Delhi .Currently working as a professor in Dept of Computer Science in GPREC, Kurnool.