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 3, March 2017

A Research Based On Mining User Aware-Rare STPS in Document Streams

Jaseela Jasmin TK, Ambili K

P.G Scholar, Dept. of Computer Science and Engineering, Cochin College of Engineering, Kerala, India

Assistant Professor, Dept. of Computer Science and Engineering, Cochin College of Engineering, Kerala, India

ABSTRACT: Creation and distribution of document streams on the internet are ever changing in various forms. Existing works are mainly consigned to topic modelling and the expansion of individual topics. They ignore sequential relations of topics in document streams. In this paper, proposes a new kind of patterns for rare event mining, which is able to characterize personalized and abnormal behaviours of internet users. We present a group of algorithms to solve this mining problem through three phases: pre-processing, session identification and mining User-aware Rare Sequential Topic Patterns (URSTPs). Practically, it can be applied to many real life scenarios of user behaviour analysis such as discovering special interests and browsing habits of internet users.

KEYWORDS: Rare event, correlations, jiggles, web data mining, LDA, sequential topic patterns.

I. INTRODUCTION

Textual documents are distributed in different forms on the internet, such as micro-blog articles, chatting messages, and research paper archives, web forum discussions etc. Topics extracted from these document streams mainly reflect offline social events and users characteristics in real life. In order to characterize user's behaviours in document streams, we must study correlations and sequential relations among topics that extracted. Specify them as sequential topic patterns (STPs).

For a document stream, some STPs may occurs frequently and thus reflect offline behaviours of internet users. Beyond that there may exist some patterns which are globally rare, but occur frequently for specific users. These are the User-aware Rare STPs (URSTPs). Compared to frequent ones, discovering them is interesting and significant, so can be applied in many real life scenarios, such as real time monitoring on abnormal user behaviours. For example, micro blog messages are real time, which reports what users are feeling, thinking, and doing. So it reflects users characteristics and statuses, both content information and temporal relations of messages are required for analysis. Fraud behaviours in internets are make award seductiveness, jiggle other user's information, obtaining various payments by cheating, and take illegal coercion if their requests are denied. Hence mining URSTPs is a good measure for real-time user behaviour monitoring on the internet. In the case of browsed document streams mining URSTPs can better discover special interests and browsing habits of internet users. Through this we can give effective and context-aware recommendation for them.

There are many technical challenges are raised and will be solved by this paper. First challenge is about the input of the task is textual stream, so existing techniques cannot be applied directly. That is, a pre-processing phase is necessary such as topic extraction and session identification. Secondly, real time requirements such as accuracy and efficiency of mining algorithms. Thirdly, formal criterion must be well defined. However, determining user interests and abnormal characteristics are somewhat important for worldwide internet 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 3, March 2017

II. RELATED WORK

Frequent pattern mining with uncertain data studies the problem of frequent pattern mining with uncertain data [2]. It shows how broad classes of algorithms can be extended to the uncertain data setting. In particular, we will study candidate generate-and-test algorithms, hyper-structure algorithms and pattern growth based algorithms. One of our insightful observations is that the experimental behaviour of different classes of algorithms is very different in the uncertain case as compared to the deterministic case. In particular, the hyper-structure and the candidate generate-and-test algorithms perform much better than tree-based algorithms. Probabilistic frequent item set mining in uncertain databases [3] semantically and computationally differs from traditional techniques applied to standard "certain" transaction databases. The consideration of existential uncertainty of item, indicating the probability that an item set occurs in a transaction, makes traditional techniques inapplicable. It introduces new probabilistic formulations of frequent item sets based on possible world semantics.

Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition [4] have enabled social media services to support space-time indexed data, and internet users from all over the world have created a large volume of time-stamped, geo-located data. Such spatiotemporal data has immense value for increasing situational awareness of local events, providing insights for investigations and understanding the extent of incidents, their severity, and consequences, as well as their time-evolving nature. Its working flow depicts at figure 1. With the vast amount of digitized textual materials now available on the Internet, it is almost impossible for people to absorb all pertinent information in a timely manner. To alleviate the problem, we present a novel approach for extracting hot topics from disparate sets of textual documents published in a given time period. That is, hot topic extraction based on timeline analysis and multidimensional sentence modelling [5] shown in figure 2.

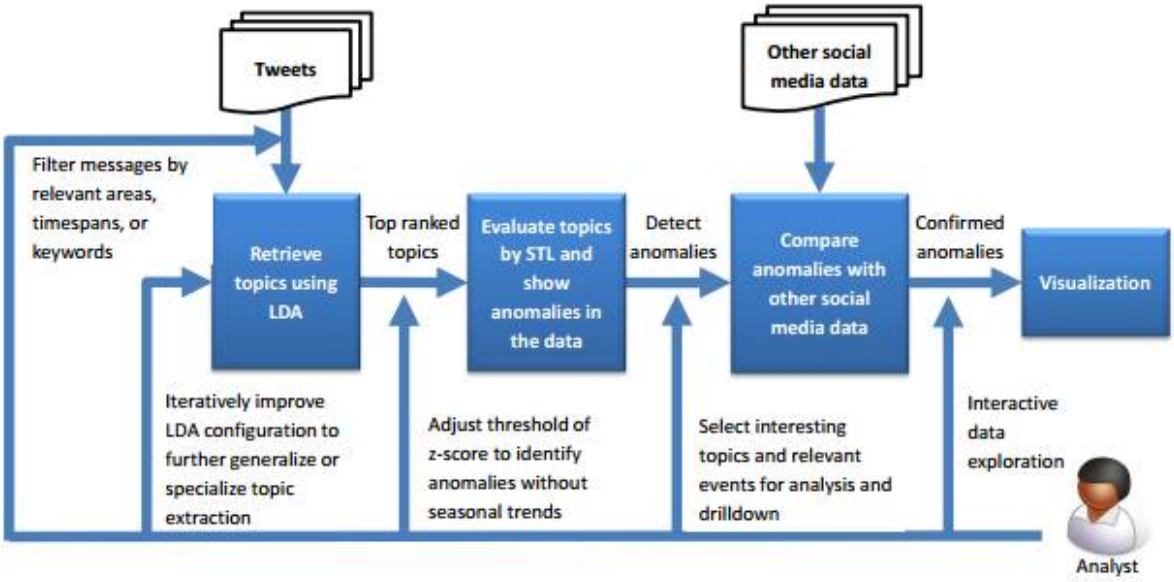

Figure 1: Overview of interactive analysis scheme for spatiotemporal event detection

To extract key terms from a text document, the basic language elements need to be considered. In the process of the key terms extraction, the step of "word segmentation and term merging" is related to languages. So, the first step is to split the documents into a list of separate term via tokenization and part of speech analysis. The terms in the list are sequenced according to their occurrence frequency. Thus, two adjacent terms in the list are merged into a longer term if their occurrence frequencies exceed a predefined threshold. The first term of the pair is accepted as a key term

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 3, March 2017

candidate if its frequency is greater than the threshold and did not merge with its preceding and following term. The process is executed iteratively until no keywords remained for processing in the list.

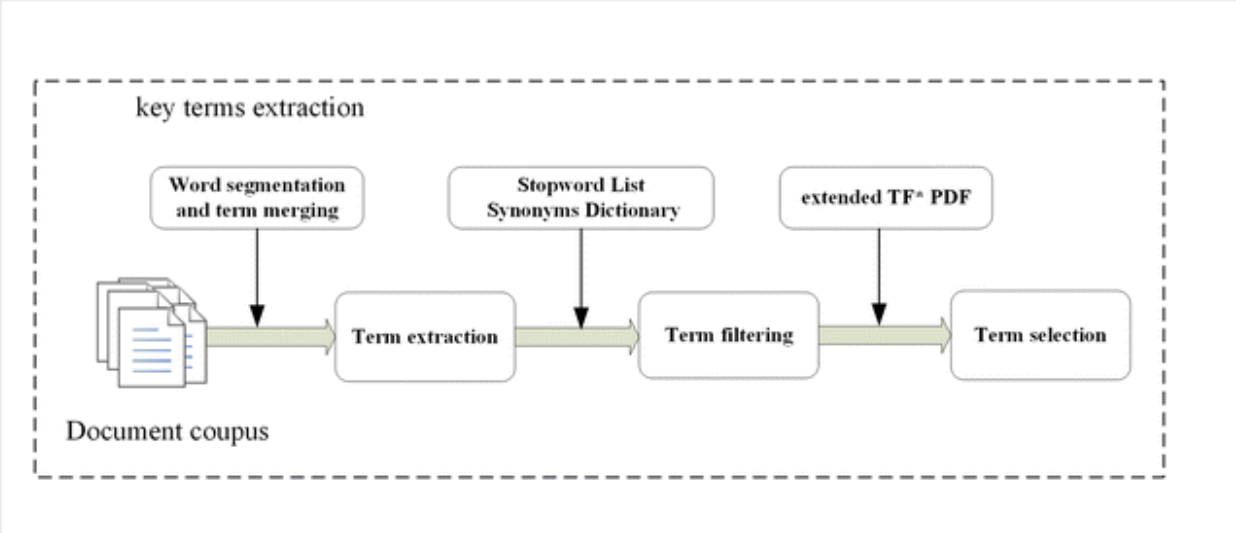

Figure 2: Key term extraction

Hariri et al [7] presented an approach for context-aware music recommendation based on the sequential relations of latent topics. Here the topic set of each song is determined by a threshold on the topic obtained from linear discriminate analysis. Topic set is deterministic, so uncertainty degree of topics is lost due to approximation in threshold based filtering.

However, these works do not consider where the uncertain databases come from and how the probabilities in the original data are computed, so cannot be directly employed for our problem which takes document streams as input. Moreover, they focused on frequent patterns and thus cannot be utilized to discover rare but interesting patterns associated with special users.

III. PROPOSED SYSTEM

Sequential pattern mining concerned with finding statistically relevant patterns between documents where the values are delivered in sequence that is documents are created and distributed in a sequential way. One user can write almost one document at a time. Formally, if $t_p = t_q$ then $u_p \; ! = \; u_q$ ways hold. Here 'up' represents documents that are published by the user 'p' at time t. Correlations that we consider here is the sequential topic patterns. Each session represents the subsequences related to a user during a certain period. For a specific user 'p' there exist disjoint and consecutive multiple sessions in a document stream. From this we need to identify URSTPs. They are globally rare for all sessions relatively frequent for the sessions associated with specific users. Frequency calculation done by using classical concept called support. Support is an indication of how frequently item sets are appears in the database. It is the proportion of transaction't' in the database which contains item sets 'x'. Scaled support is the square root of the support which can be used for the further calculations.

According to these ideas, define two measures absolute rarity and relative rarity. Absolute rarity is the difference between the local support (sessions for particular users) and global rarity (all the sessions). Relative rarity is the difference between the absolute rarity and average of the absolute rarity among all the discovered STPs of particular user. From these we get two thresholds relative rarity threshold and scaled support threshold respectively. Finally we mine URSTPs from two conditions that are scaled support ≤ scaled support threshold and relative rarity ≥ absolute

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 3, March 2017

rarity threshold hold for same user 'u'. Discovered URSTPs of associated users gives their personalized and abnormal behaviours.

1. SYSTEM ARCHETECTURE

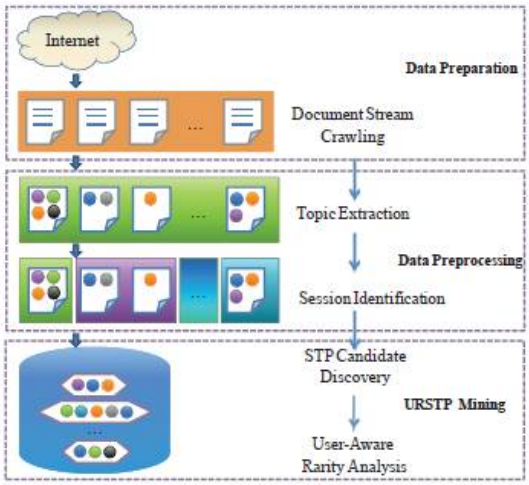

Figure 3: Processing frame work of URSTPs mining

It proposes a novel approach to mining URSTPs in document streams. 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 pre-processing procedures, original stream is transformed to a topic level document stream and then divided into many sessions to identify complete user behaviours. Finally, 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.

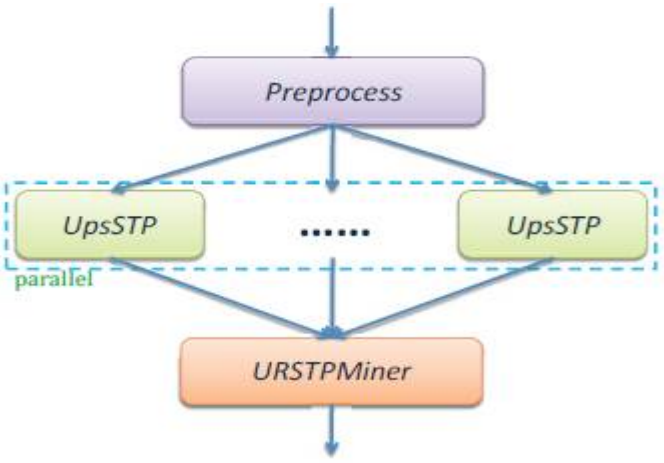

Figure 4: URSTPs work flow

2. PROPOSED ALGORITHM

Purpose: Mining URSTPs
Input: Document stream, scaled support threshold h_{ss} , relative rarity threshold h_{rr}.
Output: Rare sequential patterns

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 3, March 2017

Procedure
STEP 1: Input the textual stream

STEP 2: Pre-processing
2.1: Topic Extraction
2.2: Session Identification
STEP 3: STP candidate discovery
STEP 4: User Aware rarity analysis

3. DATA PREPROCESSING

There are many pre-processing techniques such as linear discriminate analysis to get topic proportion and word distribution. Topic extraction is done by LDA [12] or Twitter LDA [13] to get topic proportion of each document and the word distribution of each learnt topic. Session identification is next step of pre-processing. Each session contain behaviour of individual users. Two classical methods are used for the session identification. That is Time Interval Heuristics and Time Span Heuristics. In time interval heuristics, it examines each document on the input stream orderly for checking new session starting. By checking condition that the time difference between it and previous documents exceeds the given predefined threshold. Time Span Heuristics assumes the duration of each session is less than or equal to a predefined threshold.

4. STP CANDIDATE RECOVERY

It aims to discover all the STPs associated with each user, paired with expected support. We use dynamic programming algorithm to derive all STPs for user and exactly compute the values of them. Use DP matrix for traversing all the entries. After that we use PrefixSpan [14] algorithm to discover STP candidate by pattern growth.

5. USER-AWARE RARITY ANALYSIS

After all the STP candidates are discovered, perform the user aware rarity analysis to find URSTPs. It shows personalised and abnormal behaviours of special user. We use URSTPMiner algorithm. Which transform set of user-STP pairs into a set of user session pairs and two thresholds, the scaled support threshold h_{ss} and the relative rarity threshold h_{rr}. They are used as the input parameter.

URSTPMiner ALGORITHM
STEP 1: Compute set ϕ, containing all the discovered STPs for all the users.
STEP 2: Compute global support as a weighted average of its local support for each user.
STEP 3: Normalize it to a scaled value.

$$scsupp_{\alpha} \leftarrow \sqrt[|\alpha|]{supp_{\alpha}}$$

STEP 4: Calculate absolute rarity AR_{α} for user STP and its average value

$$AR_{\alpha} \leftarrow \sqrt[|\alpha|]{p} - scsupp_{\alpha}$$

$$avrAR \leftarrow avgAR + {AR_{\alpha}}/{|\varphi_u|}$$

STEP 5: Calculate relative rarity

$$RR_{\alpha} \leftarrow AR_{\alpha} - avgAR$$

STEP 6: If $RR_{\alpha} \geq h_{rr}$, it gives relatively high frequency for user.
If $scsupp_{\alpha} \leq h_{ss}$ hold for some user u, indicates global rareness of α
STEP 7: Return STP support

IV. RESULT AND DISCUSSIONS

Here we use IBM data generator [2] to get probabilistic datasets. Take some users and assign sessions for each of them. Each session is a sequence of item sets directly obtained from the generator, where each item sets regarded as document and each item represents a topic. We divide this topic in to two kinds, 80% common topics and 20% rare topics. Rare topics are globally rare and locally frequent, and are assigned to some 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 3, March 2017

Next we assign a probability to each topic occurring in data sets with a uniform distribution and normalises these values. Pre-processing phase is not required due to us already assign session for each user. Initially the user number is set to 50; the number of sessions for each user is selected from a Poisson distribution with mean 100, the size of each session is also picked from the Poisson distribution with mean 5 and the number of topics in each document is randomly chosen. Number of common topics and rare topics are set as $k_c = 20$ and $k_s = 10$ respectively.

URSTP	User	Scaled support	Relative rarity
$\langle z_7^g, z_{11}^g \rangle$	u_{1299}^g	0.046	0.441
$\langle z_{12}^g, z_1^g \rangle$	u_{1914}^g	0.032	0.476
$\langle z_8^g, z_{14}^g \rangle$	u_{125}^g	0.024	0.318
$\langle z_{13}^g, z_2^g, z_8^g \rangle$	u_{207}^g	0.029	0.340
$\langle z_{14}^g, z_1^g \rangle$	u_{1607}^g	0.043	0.559
$\langle z_4^s, z_5^s, z_5^s, z_6^s \rangle$	u_{895}^s	0.025	0.343
$\langle z_2^s, z_9^s \rangle$	u_{426}^s	0.031	0.332
$\langle z_9^s, z_2^s \rangle$	u_{861}^s	0.047	0.502
$\langle z_6^s, z_1^s \rangle$	u_{373}^s	0.033	0.334
$\langle z_1^s, z_3^s, z_7^s \rangle$	u_{875}^s	0.041	0.362

Figure 5: Examples of mined URSTPs

Values of the two thresholds would directly affect the accuracy of mined URSTPs. We find the optimal values by using F1-measure via fixing one and changing other. For the exact mining the optimal values are, $h_{ss} = 0.05$ and $h_{rr} = 0.01$. While for the approximate mining is 0.05, it can be explained by maximum pattern instance probability instead of the exact pattern occurrence probability. Taking these values as thresholds, we analyses precision, recall, and F1-measure with different user numbers.

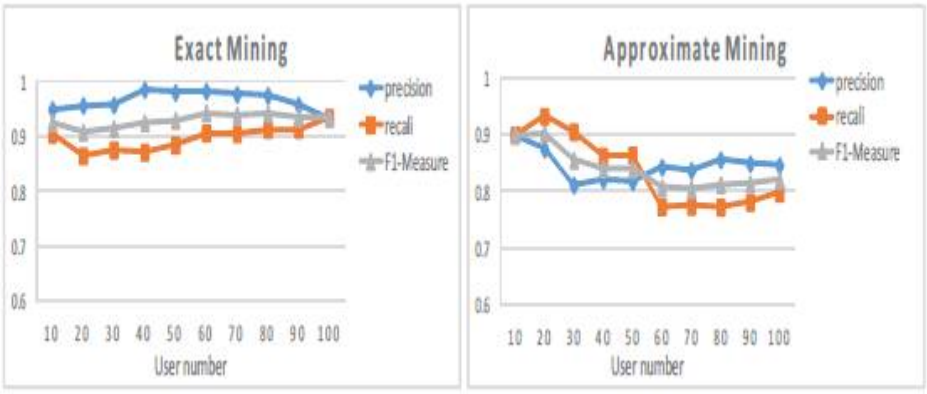

Figure 6: Values of precision, recall and F1

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 3, March 2017

For exact mining, precision varies between 0.90 and 0.98 and recall varies between 0.86 and 0.95, both are high and thus compelling. As the number of users increases from 40, recall shows an upward trend and precision maintain a high value, but decline moderately due to pattern will become sparse. F1-measure is comparatively stable.

V. CONCLUSION AND FUTURE WORK

URSTPs mining in document streams are challenging and significant problem on the internet. It gives a new kind of patterns with wide application scenarios such as real time monitoring and discovering special interests of internet users. We can identify personalised and abnormal characteristics of each user through STPs. The various experiments that conduct on the specially designed databases demonstrate the proposed approach is very effective. This paper, also forwards innovative approach research direction on the web data mining.

REFERENCES

1. Jiaqi Zhu, Member, IEEE, Kaijun Wang, Yunkun Wu, ZhongyiHu, and Hongan Wang, Member, IEEE, 'Mining User-Aware Rare Sequential Topic Patterns in Document Streams', IEEE Transactions on Knowledge and Data Engineering, 2016.
2. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, 'Frequent pattern mining with uncertain data', in Proc. ACM SIGKDD'09, pp. 29-38, 2009.
3. T. Bernecker, H.P. Kriegel, M. Renz, F. Verhein, and A. Zuee, 'Probabilistic frequent itemset mining in uncertain databases', in Proc. ACM SIGKDD'09, pp. 119-128, 2009.
4. 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 V AST'12, pp. 143-152, 2012.
5. 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.
6. C. K. Chui and B. Kao, 'A decremental approach for mining frequent itemsets from uncertain data', in Proc. PAKDD'08, pp. 64-75, 2008.
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. D. Blei, A. Ng, and M. Jordan, 'Latent Dirichlet allocation', J. Mach. Learn. Res., vol. 3, pp. 993-1022, 2003.
9. 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, pp. 93–102, 2012.
10. G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, 'Parameter free bursty events detection in text streams', in Proc. VLDB'05, pp. 181-192, 2005.
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.
12. A. K. McCallum. (2002) MALLET: 'machine learning for language toolkit', [Online]. Available: http://mallet.cs.umass.edu
13. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, 'Comparing Twitter and traditional media using topic models', in Adv. Inform. Retr. LNCS 6611, Springer, pp. 338–349, 2011.
14. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, 'PrefixSpan: Mining sequential patterns by prefix projected growth', in Proc. IEEE ICDE'01, pp. 215–224, 2001.