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 4, April 2017

Discover User Rare STPs In Browsed 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: Distribution of textual documents on the internet is ever changing in various forms. Instead of mining characteristics of users from published document streams, here we concentrated on browsed document streams. Users are browsing various textual documents from various sources. To understand their personal interests and behaviours are one of the innovative applications of the User-aware Rare Sequential Topic Patterns (URSTPs). STPs can characterise complete browsing behaviours of readers, so compared to statistical methods mining URSTPs better to discover special interests and browsing habits of internet users. Hence, it is capable to give effective and efficient context aware recommendation for them.

KEYWORDS: STPs, web data mining, LDA, hash map, Twitter LDA, PLSI.

I. INTRODUCTION

Document streams are distributed in diverse forms on the internet, such as research paper archives, chatting messages, Web forum discussions and so forth. Contents extracted from these textual documents may be reflecting offline social events and characteristics of users in real life. Most existing works are concentrated to topic modelling and evolution of individual topics to detect and predict their social activities and personalised character. STPs are usually considered under the topic correlation. In STPs topics are entered in to sequentially. Topics extracted from these document streams exposes offline social events and users characteristics in real life. In browsed document streams, there exist some patterns which are globally rare, but occur frequently for specific users. We address them as User 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. Mining URSTPs is a good measure for real-time user behaviour monitoring on the internet [1]. We know that fraudulent activities on internet are increasing day by day. In the case of browsed document streams mining URSTPs can better to prevent these fraudulent activities and especially discover interests and browsing habits of users.

Many technical challenges are raised during this work, first challenge is about the input of the task is textual stream. That is, a pre-processing phase is needed such as topic extraction and session identification. Secondly, real time requirements like accuracy and efficiency of mining algorithms. Thirdly, formal criterion must be well defined. There are four different searching methods are available for the logged users. They are topic based searching, keyword based searching, date based searching and related word searching. A special hash map algorithm is used during keyword and date based searching.

II. RELATED WORK

Topic Detection and tracking aimed to detect and track topics in news streams based on clustering techniques to key words [2]. Semantic association and correlations of words based probabilistic generative models for topic extraction also proposed, such as PLSI [3] and LDA [4]. In many real time applications, document collections generally carry temporal information and thus can be considered as document streams. Various dynamic topic modelling methods proposed to discover topics over time in document streams. These methods are designed to construct the evolution

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 4, April 2017

model of individual topics from document stream. Frequent pattern mining with uncertain data studies the problem of frequent pattern mining on uncertain data [5]. Studies shows broad classes of algorithms can be extended to the uncertain data setting. 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. Probabilistic frequent item set mining in uncertain databases [6] semantically and computationally differs from traditional techniques, 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 [7] have enable 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. Researchers proposes a novel approach for extracting hot topics from disparate sets of textual documents published in a given time period for neglecting above problem. That is, hot topic extraction based on timeline analysis and multidimensional sentence modelling [8] shown in figure 1.

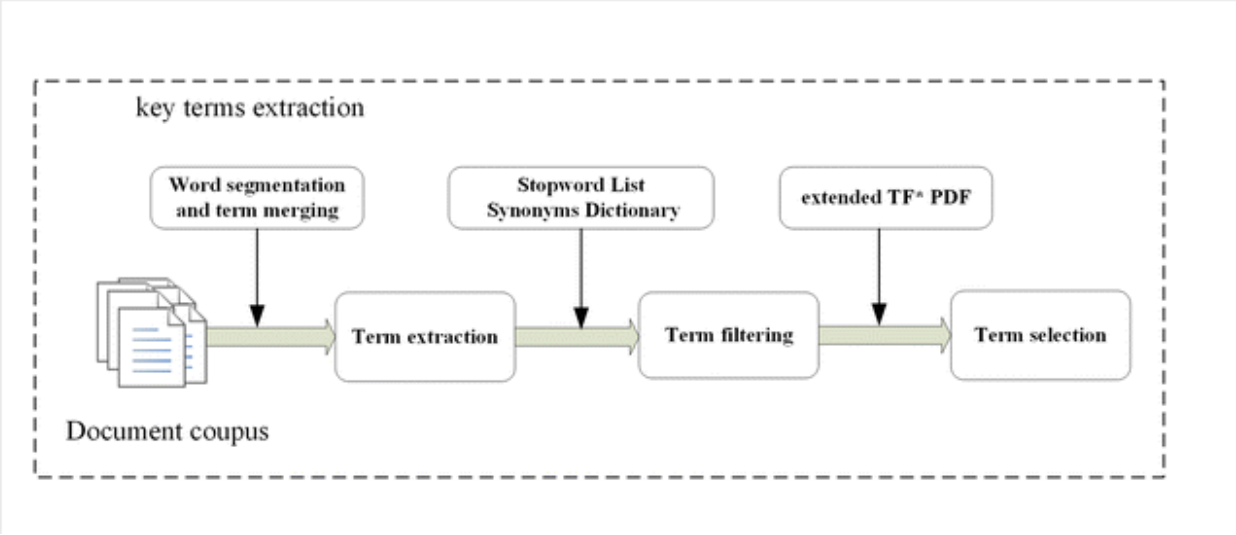

Figure 1: Key term extraction

Muzammal et al. paying attention on sequence-level uncertainty in sequential databases, and proposed methods to evaluate the frequency of sequential pattern based on the expected support, in the frame of candidate generate and test [9] or pattern growth [10]. Expected support is probability of an item is presented in database. However, these works do not consider where the uncertain databases approach from and how the probabilities in the original data are compute, so cannot be directly engaged for our problem which takes document streams as input.

III. PROPOSED SYSTEM

A. *System Architecture:*

Proposed system mainly consists of three phases that are data preparation, data pre-processing and URSTPs mining. In data preparation textual documents are collected from different sources and uploaded in to the system. Three phases are shown in Figure[3]. Then pre-process it, textual streams are transferred in to topic level document stream and divided it into different sessions. We formally called them, session identification and topic extraction. Extraction of topics browsed by the users is done by Twitter LDA [11]. Session identification done by using Time Interval Heuristics or Time Span Heuristics, which divides each document streams in to different sessions for each users. 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. Time interval Heuristics with variable duration is more suitable for our problem. Next we discover STP candidates by pattern

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 4, April 2017

growth. It aims to find all STPs occurring in the document stream associated with specific users, here we use DP based algorithm to derive all the STPs of user and exactly compute the support value of them. Occurrence probability of a session can be correctly computed using dynamic programming. Dynamic Programming matrix method is usually used.

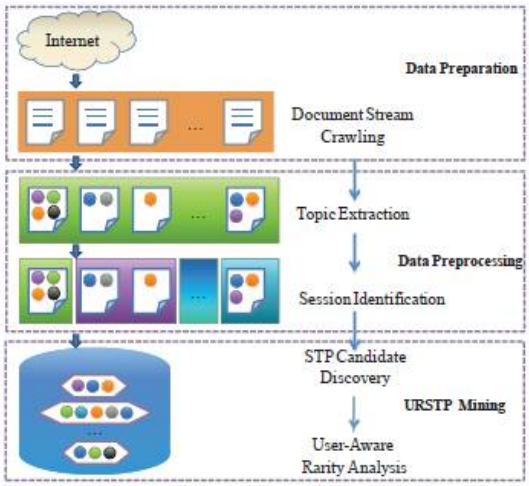

Figure 2: Processing frame work of URSTPs mining

Use DP matrix for traversing all the entries. Finally we use PrefixSpan [12] algorithm to discover STP candidate by pattern growth. If all the STP candidates are discovered, then performs the user aware rarity analysis to find URSTPs. It shows personalised and abnormal behaviours of special user. We use URSTPMiner algorithm. There are two measures are used for the evaluation of this that are 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 rarity threshold hold for same user 'u'. Discovered URSTPs of associated users gives their personalized and abnormal behaviours [1].

B. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to find User Rare STPs. The proposed algorithm is consists of seven main steps.

STEP 1: 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.

$scsupp_\alpha \leftarrow \sqrt[|\alpha|]{supp_\alpha}$; Scaled support is the square root of the support.

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; apply calculated AR in to equation given below

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

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 ofα, checking the conditions.

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 4, April 2017

STEP 7: Return STP support

Discover STP candidates for all users, we will carry user rarity analysis to pick out URSTPs, which imply their personalised, abnormal and significant behaviours. Transform set of user STP-pairs in to a set of user- URSTPs pairs with the set of user session pairs and two thresholds. That is scaled support threshold and relative rarity thresholds. We use them as input parameters. First we get all the derived STPs of all users, for each of them compute the global support as weighted average of its local support and normalize it to a scaled value according to step 3. After that for each user, compute AR_α ,$avrAR$ and RR_α. Next we select locally frequent STPs by the threshold h_{rr}. It generates an STP-RR pair. Finally the set of pairs together with the key value is added to the set of user URSTPs pairs, which will be returned.

We are giving additional facilities to the browsed users; they can search documents four different types. Those are topic based searching, keyword based searching, date based searching and related word searching. During key word based searching hash map is used to find exact match, which uses exact principle of hash function. All objects in java inherit a default implementation of **hashCode()** function defined in Object class. This function produce hash code by typically converting the internal address of the object into an integer, thus producing different hash codes for all different objects. In put() method , same logic is applied in get() method also. The moment HashMap identify exact match for the key object passed as argument, it simply returns the value object stored in current Entry object. If no match is found, get() method returns null. The moment HashMap identify exact match for the key object passed as argument, it simply returns the value object stored in current Entry object.

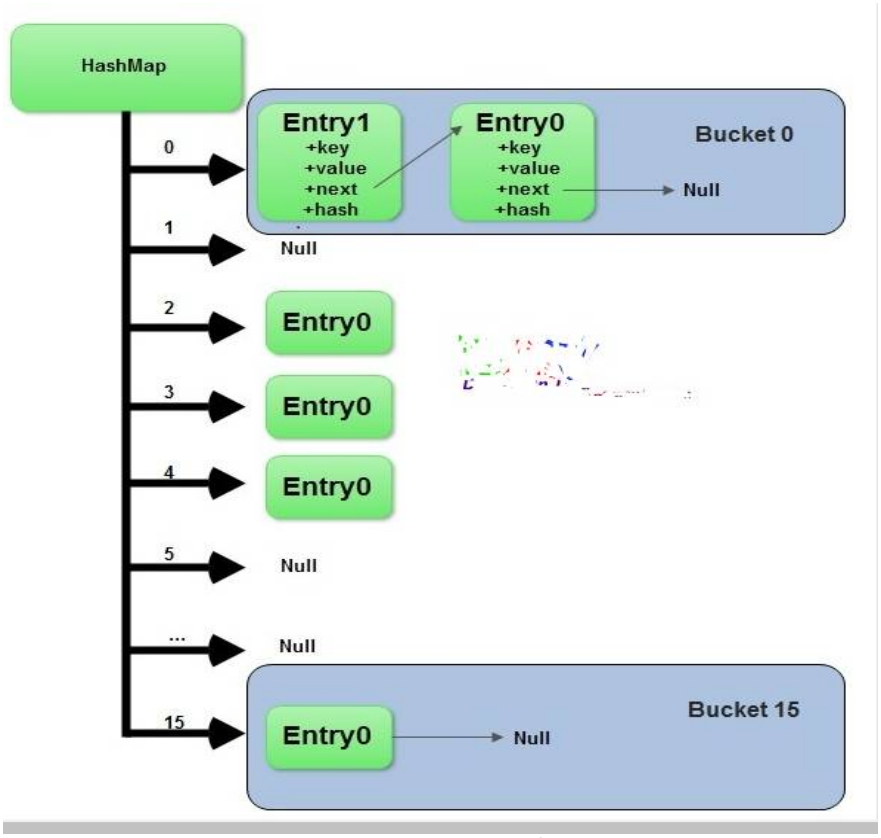

Figure 2: Internal working of hash map

A bucket is used to store key value pairs. A bucket can have multiple key-value pairs. In hash map, bucket used simple linked list to store objects. Hash Map get(Key k) method calls hash Code method on the key object and applies returned hash Value to its own static hash function to find a bucket location(backing array) where keys and values are

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 4, April 2017

stored in form of a map entry . So you have concluded that from the previous line that both key and value is stored in the bucket as a form of Entry object. So thinking that only value is stored in the bucket is not correct and will not give a good impression on the interviewer. After processing hash map we get keywords that we search. Finally admin can view the all histories of the user. Thereby he can discover all the interests, hobbies and especially their abnormal behaviours.

IV. RESULT AND DISCUSSIONS

Problem of mining URSTPs in document streams proposed in this paper is innovative, but the effectiveness of our approach in identifying personalised and abnormal behaviours are need to be validated. We also need to evaluate the efficiency, shown in Figure 4. Collect two Twitter datasets as real document streams, a general dataset and a special sports-related dataset.

At first, we check whether the mined URSTPs are really associated to the users with special or abnormal behaviors. Here, we make a reasonable assumption that "verified" users in Twitter are more likely to have special and abnormal behaviors than normal 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. As discussed above, we mainly concern a small fraction of users with topmost relative rarity values, so recall is insignificant. Moreover, the difference induced by the two topic models for URSTP mining is much smaller than simple topic mining. That indicates compared to individual topics, the sequential patterns can integrate the information in successive and interrelated documents. Examples of mined URSTP are shown in Figure 3.

URSTP	User	Scaled support	Relative rarity
$\langle z^g_7, z^g_{11} \rangle$	u^g_{1299}	0.046	0.441
$\langle z^g_{12}, z^g_1 \rangle$	u^g_{1914}	0.032	0.476
$\langle z^g_8, z^g_{14} \rangle$	u^g_{125}	0.024	0.318
$\langle z^g_{13}, z^g_2, z^g_8 \rangle$	u^g_{207}	0.029	0.340
$\langle z^g_{14}, z^g_1 \rangle$	u^g_{1607}	0.043	0.559

Figure 3: Examples of mined URSTP

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 values as thresholds for analyse precision, recall and F1-measure with different user numbers, shown in Figure 5.

URSTP quality	@5	@10	@15	@20
Self-interpretability-g	4.20	4.00	3.73	3.55
Consistency-g	4.60	4.20	4.20	3.20
Self-interpretability-s	4.00	3.75	3.70	3.25
Consistency-s	4.50	4.50	4.26	4.45

Figure 4: Performance measures on URSTPs mining.

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 4, April 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.

Figure 5: Values of precision, recall and F1

To validate the scalability of the algorithms, we aim at discovers the most time-consuming sub procedure and computes all the user-STP pairs, including the exact version UpsSTP and the approximate version UpsSTPa. The baselines are chosen as the two Apriori based frequent sequential pattern mining algorithms for probabilistic databases, Depth-First Exploration and Breadth-First Exploration. We modify them to accommodate our problem, and compare the execution times of the four algorithms with changing values of the average session number m and the average session size q respectively, as shown in Figure 6.

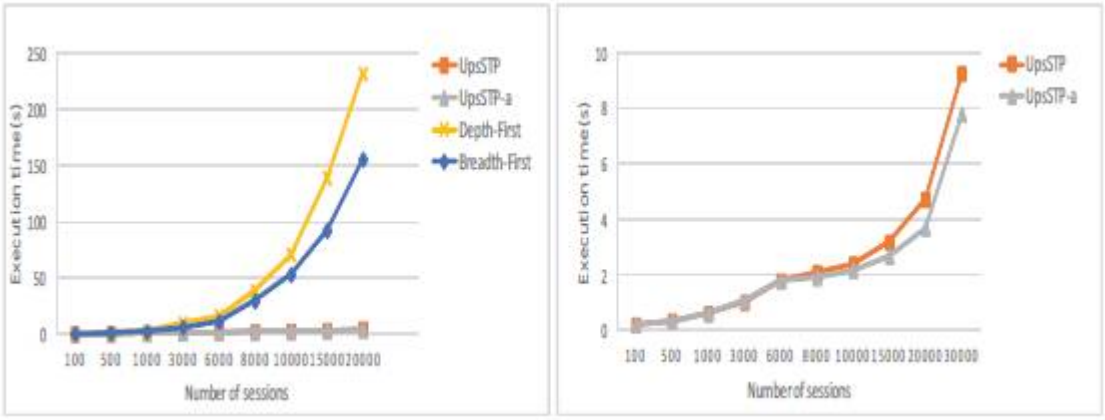

Figure 6: Time Costs of Algorithms with Changing Session Numbers

As no pruning strategies are applied here, the values of thresholds would not affect the time cost. It is observed that the approximation algorithm is indeed a little faster, especially for larger scales. Notice that each execution of this sub procedure is just for one user, so when the user number increases, the time difference for the whole approach will become more and more evident, even with some extent of parallelism.

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 4, April 2017

V. CONCLUSION AND FUTURE WORK

URSTPs mining in browsed document streams are challenging and significant problem on the internet. It gives a new method with wide application scenarios such as real time monitoring and discovering special interests of internet users. We can identify their special habits and interests; hence we can give context aware recommendation for them. 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. J. Allan, R. Papka, and V. Lavrenko, 'On-line new event detection and tracking', in Proc. ACM SIGIR'98, pp. 37-45, 1998
3. T. Hofmann, 'Probabilistic latent semantic indexing', in Proc.ACM SIGIR'99, pp. 50–57, 1999.
4. D. Blei, A. Ng, and M. Jordan, 'Latent Dirichlet allocation', J. Mach. Learn. Res., vol. 3, pp. 993-1022, 2003.
5. R. Agrawal and R. Srikant, 'Mining sequential patterns', in Proc.IEEE ICDE'95, pp. 3–14, 1995
6. D. M. Blei and J. D. Lafferty, 'Dynamic topic models', in Proc.ACM ICML'06, pp. 113–120, 2006.
7. 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.
8. D. Blei and J. Lafferty, 'Correlated topic models', Adv. Neural Inf.Process. Syst., vol. 18, pp. 147–154, 2006.
9. 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.
10. 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.
11. 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.
12. C. K. Chui and B. Kao, 'A decremental approach for mining frequent itemsets from uncertain data', in Proc. PAKDD'08, pp. 64-75, 2008.
13. 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.
14. 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.