Certain Investigations In Text Mining With Respect To Abnormal User Behavior Using Sequential Topic Patterns

Ms.Christy Abraham[1], Ms. Pradheeba P[2], Ms. Hyba H[3]
[1, 2, 3] CMS College of Engineering and Technology

Abstract- *The executions conducted under the title of Mining User Aware Rare Sequential Topic Patterns in Document Streams had to take care of datasets of Tweets which was collected in large quantities by using Twitter API. It carry out the extraction of probabilistic themes from data set to dig up abstract and probabilistic depiction which forms the initial step in data mining method proposed through the study. These topics are utilized to make out various sessions of information transfer for publishing and insinuate absolute and repetitive tricks. There are possibilities of monotonous mode of activities from an assortment of users thereby displaying a pattern which is though not confined to the entire stream but to several explicit users termed as Sequential Topic Patterns (STPs).*

Keywords- Data Mining, Twitter Dataset, STP, URSTP, User Aware Rarity Analysis, Real-Time Monitoring On Abnormal User Behaviours, Data Preprocessing

I. INTRODUCTION

Document streams are fashioned and dispersed in a variety of outlines on the Internet, for example news pour out, emails, micro-blog critique, chatting communication, research doc-ument annals, web forum deliberations, and huge pile dif-ferent versions. The stuffing of these credentials in the main ponders on a quantity of explicit subject matter, which is a sign of offline social procedures and distinctiveness of user in factual being. In order to mine such available stream of data, innumerable explorations were performed giving core importance to text mining concentrated on digging out mat-ter under discussion from manuscript anthology and doc-ument streams in the course of diverse probabilistic theme sculpts.

Intriguing benefit of these hauled out topics in document rivulets, most of obtainable works scrutinized the progres-sion of entity topics to become aware of and envisage social proceedings as well as nature of client activities. On the oth-er hand, few researches remunerated interest to the relation-ships in the midst of diverse foci coming into view in con-secutive documents made into public by an explicit end-user, so a little concealed although momentous information to expose adapted behaviours has been abandoned.

For typifying user activities in available document streams, a revision is performed on the correlations between topics taken out from these papers, particularly the chrono-logical dealings, and stipulates them as Sequential Topic Patterns (STPs). Everyone among the available group, ac-count on the absolute and repetitive performance of a user whilst a series of documents are uploaded via networked communication schema. The correlated yield from the dig-ging, while contributes towards the concluding note of the fundamental features and psychosomatic statuses of the person. First, contrasted to isolated matter of discussions, STPs incarcerate equally mixture and assortment of topics, so know how to give out usefully as discriminative entities of semantic alliance along with documents in vague state of affairs. Second, while considering normal document de-pendent models, the approach with handling of topics will offer nonfigurative information of document enclosure and are consequently advantageous in crowding together of analogous documents plus ruling several reliabilities on users. Third, the probabilistic portrayal of topics facilitates in upholding and amassing the indecision degree of sepa-rate themes of discussion, and can in that way arrive at high buoyancy echelon in pattern harmonizing for indecisive data. A strategic style is adopted to create an emotional segmentation which will output Sentimental Analysis on the available dataset.

II. PROPOSED SYSTEM

The paper presents a scheme where the following activities are realised and explained with validated outcomes. Through this paper, I have put forward prescribed descrip-tion on STPs and scarcity methods linked with them. Lay presumptuous the quandary of mining URSTPs in docu-ment streams, with the purpose of setting apart and becom-ing responsive of tailored and anomalous activities of Inter-net users. A scaffold is kept brazen to rationally unravel this dilemma, and devise analogous modus operandi to prop up it. At the outset, I have furnished pre-processing trial with heuristic modes for focus drawing out and gathering discovery. After that, having access to the facts of escalation of prototype in indecisive milieu, two unconventional algorithms are premeditated to determine the entire STP aspir-ants with prop up principles for every user. That affords a substitution linking precision and effectiveness. At length, a user responsive uncommonness breakdown algorithm is illustrated in relation to the officially defined standard to make a choice of URSTPs and allied users. The advance-ment in learning performed is corroborated by carrying out experimentations on bona fide datasets from twitter.

With the intention of typifying and becoming aware of modified and uncharacteristic actions of Internet users, Sequential Topic Patterns (STPs) were anticipated and devise the predicament of mining User-aware Rare Sequential Top-ic Patterns (URSTPs) in article rivulets on the Internet. So as to exemplify user manners in available document torrents, the correspondences among subject matters hauled out from these documents were premeditated; more than ever the se-quential dealings and spell out them as Sequential Topic Patterns (STPs). Apiece of them accounts the absolute and repetitive activities of a user when bringing out a progres-sion.

Topic mining in document anthology has been lengthily premeditated in the narrative under related works. Topic Detection and Tracking (TDT) assignment intended to be-come aware of and trail topics (proceedings) in hearsay tor-rents with huddling-founded practices on keywords. The trials carried out on realistic datasets of (Twitter) as well as imitation datasets make obvious that the anticipated meth-od is exceptionally effectual and proficient in discerning exceptional users and remarkable and analysable URSTPs from Internet article arrays, which can finely incarcerate tai-lored and nonstandard conduct and features of user.

The scheme has taken care of to provide a validating ap-proach to the recognition of STPs and connected rarity deal-ings so that hitches submitted by mining of URSTPs in doc-ument streams are tackled by portrayal and perception of bespoke and activities of Internet punters. An agenda was projected to rationally decipher this predicament and design analogous algorithm to prop up with it; at the outset pre-processing steps are executed which include drawing out topics and session credentials. The particulars on pattern intensification in uncertain setting are hired to model proce-dures owing to the breakthrough of entire array of STP con-tenders by use of magnitude of everyone. Finally, a User-Aware Rarity Analysis algorithm was on hand to single out URSTPs and allied users.

III. SYSTEM DESIGN AND IMPLEMENTATION

1. System Requirements

1) Hardware Requirements

Processor	:	Intel Core I3 Processor
Ram	:	512 MB
Hard Disk	:	80 GB

2) Software Requirements

Operating System:	Windows 8	
Language	:	Java
Tool	:	Netbeans 8.0.2
Database	:	MySQL

2. Main Processing Outline.

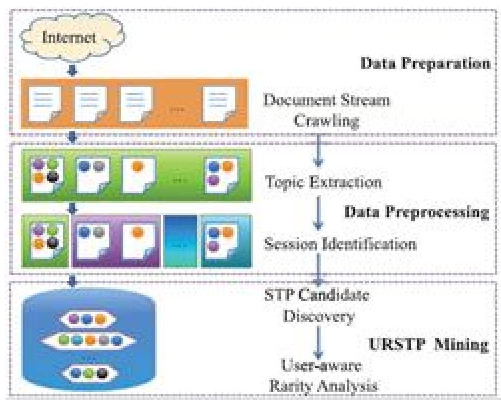

Figure 1. Processing Framework of URSTP Mining

The diagrammatical representations above in this section depicts the entire processing schema and the supportive workflow implemented through my system of data mining to dig out the USRTP and user duos.

The hierarchies shown here are of three-tier initial of which is the keying in stage where the text streams pub-lished in the twitter are gathered in bulk through the process of crawling. These streams are then fed to Data Prepro-cessing subdivision where these are converted into topics and sessions through two different activities with the name as Topic Extraction and Session Identification. Later it is fed to URSTP Mining department where STP Candidate Discov-ery and User-Aware Rarity Analysis are performed to finally get the twosomes of user-USRTP.

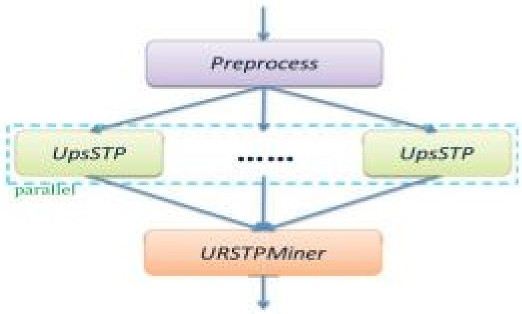

Figure 2. Workflow of URSTP Mining

3. Module Description

Data Preparation

Document Stream Crawling

This is the preliminary stage prior to the actual implementa-tion of the scheme where the data to be processed is gathered by crawling through the website, here twitter, to garb the textual torrents for my analysis; this is the exact input of the system developed.

Data Pre-processing

This is at this level of schema where the input documents are processed so that it may be undergone actual execution to dig out User-Aware Rarity; the original flow is altered to a topic echelon article stream and after that separated keen on numerous gathering to make out absolute user activities.

Topic Extraction

With the purpose of getting hold of a subject-level document flow, I have foremost made use of the traditional probabilis-tic topic models like LDA [4] and Twitter-LDA [39] to ac-quire the focus percentage of every document and the ex-pression allocation of all studied topic, with an already characterized topic number K.

For apiece file, the engendered topic share may enclose some areas with stumpy probability. They are incapable of replicating the substance of the document with lofty buoy-ancy, as a result can be expelled from the topic-level demon-stration to diminish the intricacy of anon calculations. For getting these achieved, I have selected a few delegate topics to dig up a fairly accurate topic-level text; there are two chief miscellany tactics used in my implementation:

Topic Probability Threshold:

It decides on all the likelihoods in excess of or identical to a pre-defined thresh-old htp.

Probability Summation Threshold:

subsequent to categorization of the probability principles of the K subjects in the diminishing direction, it goes for them in keeping with the order up to a achievable limit in a way that their abridgment is not more than or the same as a pre-defined threshold hps.

Session Identification

Seeing as every session ought to be full of an absolute bring-ing out activities of a personage user, I had to in the begin-ning segregate the document torrent consistent with dissimi-lar users, which is a trouble-free work as the creator of every document is unequivocally prearranged in the contribution stream. The upshot for each client u is a follower of the topic-level file flow controlled by that user. Subsequently, I was made to implement the partition of the subsequence to cate-gorize inclusive and continual performance as uninterrupt-ed and non-shared sessions.
In the subsequent, I have provided the consequent algo-rithms deliberated for my mining mission.

Time Interval Heuristics

It presupposes that the time intermission of every two ad-joining documents in the similar session is beneath or alike a pre-defined threshold hti. The algorithm scrutinizes every text on the keying in rivulet methodical to distinguish whether it should be the preparatory indication of an inno-vative session, by inspection on the circumstance that the occasion discrepancy connecting it and its preceding file goes beyond the threshold

Time span heuristics

It takes for granted that the extent of every session is either at lower lever or at the same level as a predefined threshold hts. Exclusively, the instance tip of the prior document ti 1 is re-instated by the preliminary moment of the in progress ses-sion tk, and hts is brought into play as threshold.

Instinctively, the opening heuristic with open to discus-sion extents is further apposite for my quandary than the subsequent one. On top, both the heuristics do not have need of added knowledge as of the wording of articles, and sup-pose that the manners of the entire set users apparent in a text torrent be conventional to a amalgamated time-oriented decree, so overlook changed circulating qualities of users.

URSTP Mining

Ultimately and largely decisively, I determine every bit of the STP candidates in the text stream for every single user, and additionally make a choice momentous URSTPs allied to explicit users by User-Aware Rarity Analysis.

STP Candidate Discovery by Pattern-Growth

The sub-system of STP candidate finding is carried out in analogous for apiece user which targets to come across all STPs taking place in the file stream connected with a partic-ular user, harmonizing with the accepted bearing standards of these STPs. At the start, I keep forward a DP-based algo-rithm to draw from all STPs for the user and unerringly fig-ure the prop up values of those. After that, with the aim of improving the effectiveness of my method, I had also prear-ranged an estimate algorithm to guesstimate the bearing values for all STPs; together algorithms are deliberate in the style of pattern-intensification.

DP-Based Algorithm

The incidence likelihood of an STP in a session, a succession of subject-level files, can be worked out by dynamic encod-ing. I could see that in the course of go across and working out all the ingress in the DP-matrix separately can be got hold of ultimately as it is presently the assessment of A[n,q]. Still, this technique will fetch elevated involvedness a propos mutually time along with space, i.e., O[n,q]. In reality, a few openings in the template are not obligatory to be up-held and figured. Every implementation of UpsSTP executes single stride of pattern-growth starting with the put in STP α to an unmitigated one $\beta = \alpha z$, by adding on a latest theme z.

Approximation Algorithm

In the algorithm explained in the preceding section, the working out of the literal likelihood is a solution stage, which engages inter-reliant occurrences of α in the session s and therefore convoluted. It is an innate thought to stumble on and employ an approximate numeral in its place to shorten the totalling. With the intention of creating a high-quality substitution connecting the aptitude of the STP can-didate detection algorithm and the exactness of the held up prices each step was executed. To conclude with, I had al-leged that every STP in point of fact transpires in a session paramount one time. Explicitly sensible to a few degrees because low-probability themes have been done away with the indecisive topic-level progression, and in addition, the occurrence of an STP should be replicated by the manifold happenings in altered sessions, however not in the un-changed session. Therefore, although an STP has in excess of one illustration in a session, I could pick the one with the principal probability as the delegate episode of the STP in the session.

User-Aware Rarity Analysis

In spite of everything the STP contenders for the entire users are revealed, I could craft the User-Aware Rarity Analysis to make a choice on URSTPs, which entail uncommon, atypi-cal, and accordingly momentous manners. It renovates the deposit of User-STP couples obsessed to a faction of User-URSTP duos, by way of the mix of user-session braces and two verges, the Scaled Support Threshold hss and the Rela-tive Rarity Threshold hrr, as participation constraints. Initial-ly, I got the set Φ restraining all the derived STPs for every user and apiece of them denoted as α, calculate the global support suppα as a prejudiced middling of its restricted sup-port for all users and standardize it to a resized value scsuppα. If the STP is internationally infrequent chequered by the threshold hss, it was proofed in a set $\Phi 0$ is a subset of Φ. Subsequently, for every user u, I worked out foremost the Absolute Rarity ARα for the entire of STPs over and above the standard value for the client and after that the Relative Rari-ty RRα for those STPs globally rare and set up for u. Then, STPs that are recurrent in neighbourhood are curtained by the threshold hrr, every one of which shapes an STP-RR pair for u. after a long wait, the group of these twosomes in com-mon with the key value u is made together with the cluster of User-URSTP pairs, which was revisited whilst the entire the users was used for execution.

4. Experiments

Given that the setback of mining URSTPs in document rivu-lets anticipated in this paper is ground-breaking, there are no former inclusive and akin advancements for this under-taking as the underpinning apart from the base paper cho-sen [41]. However the efficacy of my loom in discerning per-sonalized and atypical behaviours, more than ever the rea-sonability of the URSTP characterization, requests to be sen-sibly corroborated. Sentimental Analysis is an additional attraction amalgamated with the exact work execution. Ap-pealing and edifying tests were carried out on message flows in Twitter datasets, to give an idea about that largely of users revealed by my method are essentially exceptional in existent life, and the mined URSTPs can without a doubt incarcerate bespoke and anomalous activities of Internet users in a comprehensible line of attack.

In totting up, I have also appraised the competence of the system on synthetic datasets, and measure up to the two substitute sub dealings of STP contender breakthrough to make obvious on the swapping connecting accurateness and effectiveness.

IV. CONCLUSION AND FUTURE WORK

Mining URSTPs in the document streams made accessible for global view has turned out to be under keen observation and decisive magnitude; it is evidently challenging quandary. It put together a fresh category of intricate episode prototypes derived from document subject, and has extensive latent application circumstances, for instance real-time screening on anomalous behaviours of Internet accesses. In this tabloid, quite a lot of new-fangled notions and the mining setbacks are methodically delineated, and a collection of algorithms are premeditated and pooled to analytically unravel this hitch. The trials carried out on realistic streams of dataset from Twitter to demonstrate that the projected method is very effectual and competent in ascertaining exceptional customers as well as appealing and assessable URSTPs from Internet article brooks, which can finely incarcer-ate bespoke of user and nonstandard deeds and sorts; Senti-mental Analysis features the emotional departmentalization of the user tweets.

The effectiveness of the schema can be utilized for further expansions in the utilities; taking out probabilistic matters from dataset can help out in getting hold of nonfigurative as well as probabilistic metaphors. It will be easier to categorize the sessions for diverse users to be familiar with absolute and recurring manners. Sequential Topic Patterns can be acknowledged in order to accomplish the portrayal and re-vealing of the adapted actions and conduct of Internet users; the mining procedure can be carried out to spot out User-Aware-Rare-Sequential-Patterns.

V. ACKNOWLEDGMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCES

[1] Aggarwal, C. Charu, Li, Yan, Wang, Jianyong, and Wang, Jing (2009) “Frequent Pattern Mining with Uncertain Data”, Proc. ACM SIGKDD, Paris, France, pp. 29 - 38

[2] Bernecker, Thomas, Kriegel, Hans-Peter, Renz, Matthias, Verhein, Florian and Zuefle, Andreas (2009) “Probabilistic frequent itemset mining in uncertain databases,” Proc. 15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, Paris, France, pp. 119–128

[3] Blei, M David and Lafferty, D John (2006) “Dynamic topic models,” Proceedings of the 23rd international Conference on Machine Learning, pp. 113-120

[4] Blei, M David, Ng, Y Andrew and Jordan, I Michael (2003) “Latent Dirichlet Allocation” Journal of Machine Learning Research, Vol. 3, pp. 993–1022

[5] C. H. Mooney and J. F. Roddick (2013) “Sequential pattern mining - approaches and algorithms,” ACM Comput. Surv., Vol. 45, No. 2, pp. 19:1–19:39

[6] C. K. Chui and B. Kao (2008) “A decremental approach for mining frequent itemsets from uncertain data,” Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, pp. 64–75

[7] Chae, Junghoon, Thom, Dennis, Bosch, Harald, Jang, Yun, Maciejewski, Ross, Ebert, S David and Ertl, Thomas (2012) “Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition,” Proc. IEEE Conf. Vis. Anal. Sci. Technol., pp. 143–152

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

[9] D. Mimno, W. Li, and A. McCallum (2007)“Mixtures of hierarchical topics with Pachinko allocation,” Proc. ACM Int. Conf. Mach. Learn., pp. 633–640

[10] Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu, Yu, S Philip and Lu, Hongjun (2005) “Parameter free bursty events detection in text streams,” Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp. 181–192

[11] Hariri, Negar, Mobasher, Bamshad, and Burke, Robin (2012) “Context-Aware Music Recommendation Based on Latent Topic Sequential Patterns”, RecSys’12, Dublin, Ireland, pp. 131 - 138

[12] J. Allan, R. Papka, and V. Lavrenko (1998) “On-line new event detection and tracking,” Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 37–45

[13] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu (2000) "FreeSpan: frequent pattern-projected sequential pattern mining," Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 355–359

[14] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu (2001) "PrefixSpan: Mining sequential patterns by prefix-projected growth," Proc. IEEE Int. Conf. Data Eng., pp. 215–224

[15] K. Chen, L. Luesukprasert, and S. T. Chou (2007) "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," IEEE Trans. Knowl. Data Eng., Vol. 19, No. 8, pp. 1016–1025

[16] Krause, J. Leskovec, and C. Guestrin (2006) "Data association for topic intensity tracking," Proc. ACM Int. Conf. Mach. Learn., pp. 497–504

[17] L. Hong and B. D. Davison (2010) "Empirical study of topic modeling in Twitter," Proc. 1st Workshop Soc. Media Anal., pp. 80–88

[18] M. J. Zaki (2001) "SPADE: An efficient algorithm for mining frequent sequences," Mach. Learn., Vol. 42, No. 1-2, pp. 31–60

[19] M. Muzammal (2011) "Mining sequential patterns from probabilistic databases," Proc. 5th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, pp. 210–221

[20] M. Muzammal (2011) "Mining sequential patterns from probabilistic databases by pattern-growth," Proc. BNCOD, pp. 118–127

[21] M. Muzammal and R. Raman (2010) "On probabilistic models for uncertain sequential pattern mining," Proc. 6th Int. Conf. Adv. Data Mining Appl., pp. 60–72

[22] M. Seno and G. Karypis (2002) "SLPMiner: An algorithm for finding frequent sequential patterns using length-decreasing support constraint," Proc. IEEE 11th Int. Conf. Data Mining, pp. 418–425

[23] Mabroukeh, R NIZAR and Ezeife, C. I. (2010) "A taxonomy of sequential pattern mining algorithms," ACM Comput. Surv., Vol. 43, No. 1, pp. 3:1 – 3:41

[24] Mei, Qiaozhu, Liu, Chao, Su, Hang and Zhai, ChengXiang (2006) "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," Proc. 15th Int. Conf. World Wide Web, pp. 533–542

[25] Pei, Jian, Han, Jiawei, Mortazavi-Asl, Behzad, Pinto, Helen, Chen, Qiming, Dayal, Umeshwar and Hsu, Mei-Chun (2001) "PrefixSpan: Mining sequential patterns by prefix-projected growth," Proc. IEEE Int. Conf. Data Eng., pp. 215–224

[26] Seno, Masakazu and Karypis, George (2002) "SLPMiner: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint", Proc. IEEE 11th Int. Conf. Data Mining, pp. 418 – 425

[27] Srikant, Ramakrishnan, and Agrawal, Rakesh (1995) "Mining Sequential Patterns: Generalizations and Performance Improvements", Proc. IEEE Int. Conf. Data Eng., pp. 3 - 14

[28] Sun, Liwen, Cheng, Reynold, Cheung, W. David and Cheng, Jiefeng (2010) "Mining uncertain data with probabilistic guarantees," Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 273–282

[29] T. Hofmann, "Probabilistic latent semantic indexing," (1999) Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 50–57

[30] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou (2012) "LeadLine: Interactive visual analysis of text data through event identification and exploration," Proc. IEEE Conf. Vis. Anal. Sci. Technol., pp. 93–102

[31] W. Li and A. McCallum (2006) "Pachinko allocation: DAG-structured mixture models of topic correlations," Proc. ACM Int. Conf. Mach. Learn., Vol. 148, pp. 577–584

[32] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li (2011) "Comparing Twitter and traditional media using topic models," Proc. 33rd Eur. Conf. Adv. Inf. Retrieval, pp. 338–349

[33] X. Wang and A. McCallum (2006) "Topics over time: A non-Markov continuous-time model of topical trends," Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 424–433

[34] X. Yan, J. Guo, Y. Lan, and X. Cheng (2013) "A biterm topic model for short texts," Proc. 22nd Int. Conf. World Wide Web, pp. 1445–1456

[35] Y. Li, J. Bailey, L. Kulik, and J. Pei (2013) "Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases," Proc.

IEEE 11th Int. Conf. Data Mining, pp. 448–457

[36] Yang, Yiming, Pierce, Tom and Carbonell, Jaime (1998) "A Study on Retrospective and On-Line Event Detection", Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 28 – 36

[37] Zhang, Qin, Li, Feifei and Yi, Ke (2008) Finding Frequent Items in Probabilistic Data, Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 819–832

[38] Zhang, Zhongfeng, Li, Qiudan, and Zeng, Daniel (2011) "Mining evolutionary topic patterns in community question answering systems," IEEE Transactions On Systems, Man, And Cybernetics - Part A: SYSTEMS AND HUMANS, Vol. 41, No. 5, pp. 828–833

[39] Zhao, Wayne Xin, Jiang, Jing, Weng, Jianshu, He, Jing, Lim, Ee-Peng, Yan, Hongfei and Li, Xiaoming (2011) "Comparing Twitter and Traditional Media using Topic Models", Proc. 33rd Eur. Conf. Adv. Inf. Retrieval, pp. 338–349

[40] Zhao, Zhou, Yan, Da and Ng, Wilfred (2014) "Mining probabilistically frequent sequential patterns in large uncertain databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 5, pp. 1171–1184

[41] Zhu, Jiaqi, Wang, Kaijun, Wu, Yunkun, Hu, Zhongyi, and Wang, Hongan (2016) "Mining User-Aware Rare Sequential Topic Patterns In Document Streams", Proceedings of IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 7, pp. 1790 - 1804