PREDICTION OF USER RARE SEQUENTIAL TOPIC PATTERNS OF INTERNET USERS

M.SANGEEGTHA[1], D.SWATHI[2], J.PRIYANKA[3], SHALINI YUVARAJ[4]

*[1]Assistant Professor, Department of Computer Science and Engineering,
Panimalar Engineering College, Tamilnadu, India
[2][3][4]UG Students, Department of Computer Science and Engineering,
Panimalar Engineering College, Tamilnadu, India*

---***---

Abstract-*The advances of technology overtime have enabled the access to textual documents to Internet users all over the world with ease. Sequential patterns have been a focused theme in data mining. Finding the behaviour of a Sequential pattern are helpful in finding many analysing applications like predicting next event has been vital. But there exist a difficulty, since the mining may have to generate or examine a combinatorial abrupt number of intermediate subsequence. In this paper, we scrutinize abnormal behaviours of Internet users in Gmail and Twitter, we propose Sequential topic patterns (STP) and coin the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet. Some contents are frequent for specific users, so it can be used to find the abnormal behaviour of the user in real-time scenario. To achieve this, a set of algorithms is presented. It includes algorithms for pre-processing the user contents, generate all STP support values for efficient pattern growth, and selecting user-aware rare sequential topics by using rare pattern domain analysis.*

Keywords: **document streams, sequential topic pattern, pattern-growth, domain analysis, rare sequential topic.**

INTRODUCTION

Data mining, also known as knowledge discovery in databases has largely been a promising area for database research. Web services like Gmail and twitter provide a rich and freely accessible database for document streams generated and published by the users.

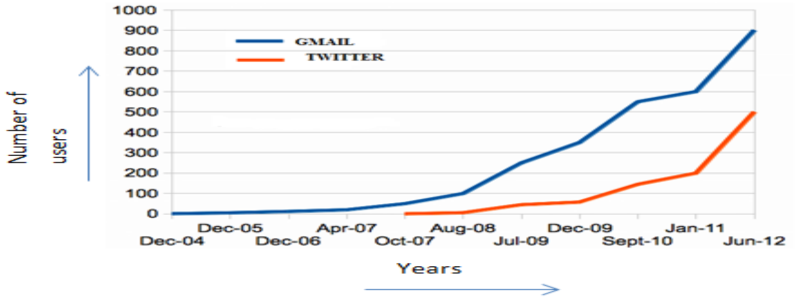

Chart- 1: Depicts Gmail and Twitter users

The number of users of Gmail and twitter are enormously increased over years. So this paper concentrates on the users using these real time applications. The document streams are in the form of micro blogs, tweets, chat messages and emails are extracted to provide a thorough insight on the behavioural analysis of an internet user. There may be some correlations among these obtained topics in successive documents for a specific user, and these correlations could be described by Sequential Topic Patterns. STP not only summarizes on topic modelling, but also investigates the user intrinsic characteristics and psychological statuses. The real intension of publishing these document streams are hard to reveal directly from individual messages, but both content information and temporal relations of messages are required for analysis, especially for abnormal behaviours without prior knowledge. STPs happen to be able to combine a series of inter-correlated messages, and can thus capture such behaviours and associated users. The important clues that trigger out the investigations are global rareness and local frequentness since some illegal behaviours are emerging and their sequential rules are not been explicit yet, but can be exposed through URSTPs. To tackle the problem of mining URSTPS in document streams, many algorithms were being used. First pre-processing phase is carried out in order to get abstract and probabilistic descriptions of documents by topic extraction and then to recognize common and repeated activities of internet users by single sign on capability. In real time run down both the preciseness and potency of mining algorithms are important. Not only frequent terms but an also rare term used by the user is concerned, so that one can characterize user's personalized and abnormal behaviour. Secondly, it performs keyphrase and aspect extraction to determine the topical words and phrases from the documents. The phrases abstracted should be understandable and grammatical. It should provide a nut shell description of the entire document. Aspect extraction, in turn computationally identifies and categorizes opinions expressed as a chunk of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. Finally, rare pattern domain analysis algorithm is applied to perform pattern-matching with the key phrases being extracted and the STP or the dataset being created and the personalized and abnormal behaviour of an individual is being determined.

2. RELATED WORK

Rare pattern behaviour analysis is extracted in document streams has been extensively studied in the literature.[1][3][11]Frequent item sets from uncertain databases that makes use of the probabilistic support concept which considers the aspects of uncertain data completely.[5][6]Presents a visual analytics approach that provides users with scalable and interactive social media data analysis and visualization including the exploration and observation of abnormal topics and events within various social media data sources, such as Twitter, Flickr and YouTube. In order to find and understand abnormal events the analyst can extract major topics from a set of selected messages and rank them probabilistically using Latent Dirichlets Allocation. [14][15][16]This papers deals with uncertainty models in sequential pattern mining. It consider situations where there is uncertainty either about a source or an event, uncertainties could be modelled using probabilistic databases, and give possible-worlds semantics for both. It then describe "interestingness" criteria based on two notions of frequentness namely expected support and probabilistic frequentness.[10] This paper empirically compare the content of Twitter with a traditional news medium, using unsupervised topic modelling. Twitter-LDA model is used to discover topics from a representative sample of the entire Twitter.

[12]This paper presents SPADE, a new algorithm for fast discovery of *Sequential Patterns*. SPADE utilizes combinatorial properties to decompose the original problem into smaller units that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. [7] A novel approach for extracting hot topics from disparate sets of textual documents published in a given time period. Solved by technique consists of two steps. First, hot terms are extracted thereby mapping their distribution over time. Second, from the extracted hot terms, key sentences are identified and then grouped into clusters that represent hot topics by using multidimensional sentence vectors. [4]A novel topic model for multi-part documents, called Multi-Part Topic Model (or MPTM in short), and develop its construction and inference method with the aid of the techniques of collapsed Gibbs sampling and maximum likelihood estimation thus improving the performance in information retrieval and document classification.[8]Mines unseen factors from web logs to personalized web search. This approach is based on probabilistic latent semantic analysis, a model based technique that is used to analyse co-occurrence data.[13] To Evaluate the execution of heuristics a method is utilized to recreate sessions from the server log information. Such heuristics are called to parcel exercises first by client and afterward by visit of the client in the site, where client recognizable proof instruments, for example, treats, might be accessible. [2][9] Propose a projection-based, consecutive example development approach for effective mining of successive examples. In this approach, a succession database is recursively anticipated into an arrangement of littler anticipated databases, and consecutive examples are developed in each anticipated database by investigating just locally visit sections.

3. EXISTING METHODOLOGY

The primary intention of an individual or a group of individuals is to harm the reputation of victims through cybercrimes. According to recent statistics given by the National Crime Records Bureau (NCRB) these figures clearly depicts the cybercrimes or terrorism under various cyber heads.

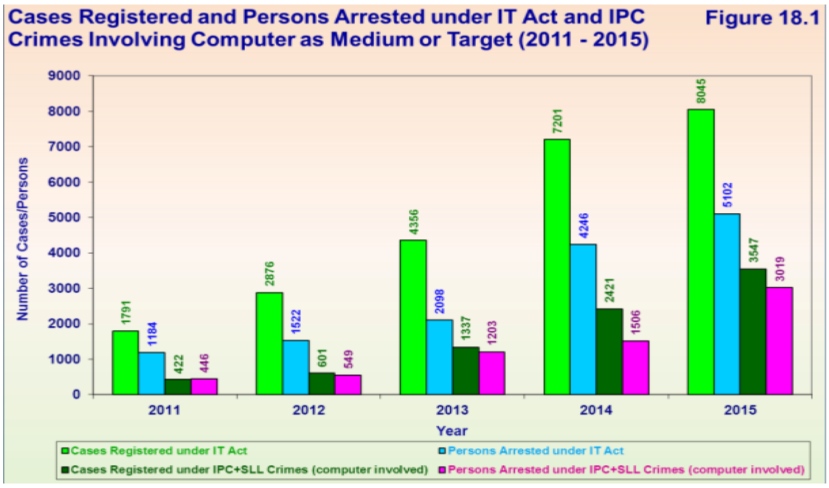

Chart-2: Cyber-crimes in 2011-2015

To overcome this cybercrimes, existing works generally concerns with topic modelling and the evolution of only the individual topics, while sequential relations of topics in successive documents is being ignored. Therefore in the existing method monitoring of users personalized activity is ineffective and in feasible. And due to static content monitoring makes the false alert on the sequential and individual topic extraction. Monitoring and mining individual user's behaviour and activity in single web application doesn't give the effective dataset of topic extraction about the user. So the users' intention and interest are extracted with ambiguous and suspicious manner due to uncertain data set. Hence the user's activity management cannot provide the effective guidance and feasible detection. In existing system, content identification is huge to handle. A session identifier (session ID) or session token is a bit of information that is utilized as a part of system correspondences to recognize a session, a progression of related message trades. A session ID is typically granted to a visitor on his first visit to a site. It is short-lived and may become invalid after a certain goal has been met. The existing methodology uses session identification, which fails to track about sequential pattern of the users and also the personalized and abnormal behaviour of the internet user.

4. PROPOSED METHODOLOGY

In our proposed system, user rare and sequential activities are monitored using a sequence of document streams from multiple web applications. Real time web applications datasets like twitter and Gmail are used with single sign on email id. The documents of inbox and send box mails of Gmail contents and twitter's tweet and individual chats, to extract the topic and mining the user's activity is being used. To extract the topic of document streams an algorithm called Natural Language processing is used to process and monitor dynamic users activities. The mining of URSTP'S in document streams has been a predominant obstacle over time and many new technical challenges came into the scenario which will be tackled in this paper. The input is in the form of textual streams, where probabilistic databases are not relevant to solve the issue. The various phases involved are: i) A preprocessing phase is mandatory to attain the abstract and probabilistic descriptions by topic extraction to acknowledge the completer and frequent activities of internet users through email-id. ii) The accuracy and efficiency of mining algorithms play a vital role, especially in the process of probability computation in terms of time requirements. iii) An alternate notion namely the user aware rare pattern is incorporated to form a well-defined mechanism so as to detect characterization, personalization and abnormality of internet users.

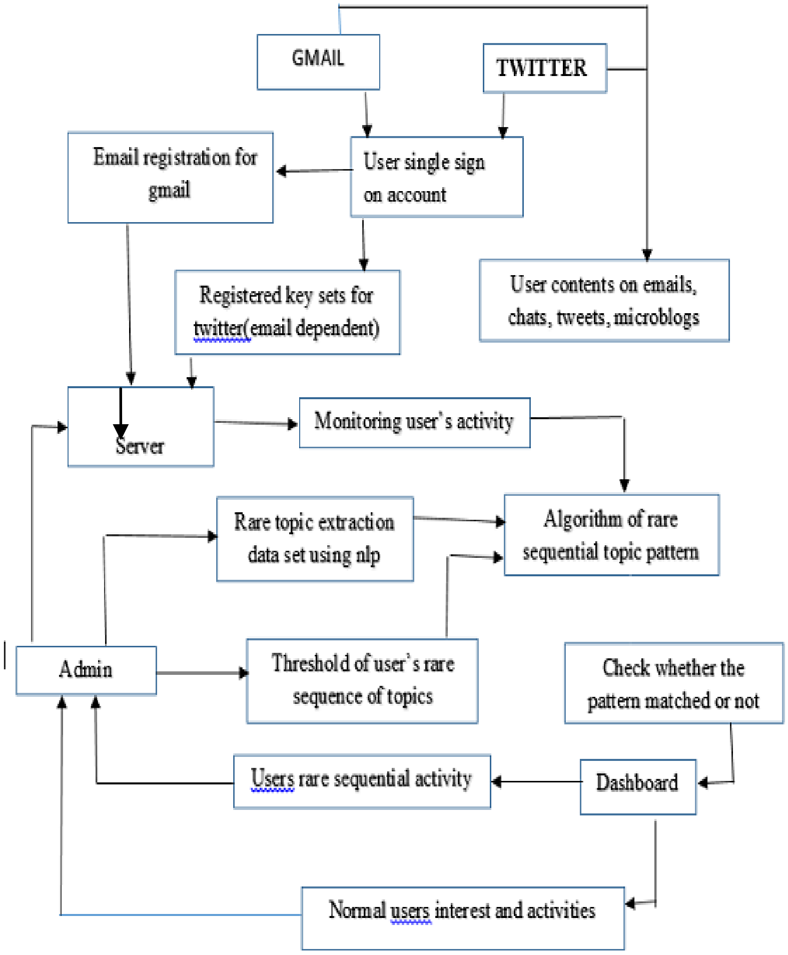
GMAIL
TWITTER
Email registration for gmail
User single sign on account
Registered key sets for twitter(email dependent)
User contents on emails, chats, tweets, microblogs
Server
Monitoring user's activity
Rare topic extraction data set using nlp
Algorithm of rare sequential topic pattern
Admin
Threshold of user's rare sequence of topics
Check whether the pattern matched or not
Users rare sequential activity
Dashboard
Normal users interest and activities

Fig-1: Architecture diagram

The above figure is a theoretical model that characterizes the structure, conduct, and more perspectives of a framework. A design portrayal is a formal depiction and portrayal of a framework, sorted out in a way that backings thinking about the structures and practices of the framework. The architecture explanation is as follows: Gmail and Twitter, being multiple web applications utilize the technique of single-sign on through an email id which is created to access and track the real time datasets of the specific user accounts. These real time applications at first gets the textual documents from the user in the form of mails, chat messages, tweets and micro blogs in order to analyse the user behaviour. For monitoring the user credentials need to be registered. Since Twitter is email dependent few registered key sets such as customer and access token secret keys needs to be specified. These data are saved in the server which in turn used to verify credentials of the user for the further verifications. The administrator maintains the server, the prime task of the administrator is to create and maintain the datasets with the help of NLP algorithm. The administrator has the control over the threshold value. The amalgamation of extracted phrases from the user contents, threshold and the datasets are given as a input to the algorithm of rare pattern domain analysis in order to deduce pattern match. The turnout of the above algorithm is displayed in the dashboard which contains the pattern obtained and also the calculated percentage of the anomalous user's activity.

5. MODULES

5.1. User's registration and creating dataset for user rare topics: -In this module the users have to register their email id and twitter key with our application. The email id and regarded twitter key's id should be a single sign on Gmail and Twitter account. Our application using users details make threshold for every users account by admin process. The data set of user's sequential topic extraction has to provide to the application. . Stanford NLP algorithm to mining the user's activity is built. The data has been maintained and customized in the server. The user's details are stored in server database in the encrypted format because of the security purpose. To implement the effective rare topic extraction on sequential of document stream of the user's activity we used deserved data set of data mining process using Stanford NLP. In this API we implements POS tagging, chunking processing, stemming, spell checking and word net connection. It feasibly extracts the content of the user's rare topics using the above mentioned NLP processing.

5.2. NLP processing on Gmail and twitter content: -The user's details are extracted and monitored from the Gmail and Twitter accounts, to our local server database. Due to the huge amount of data set a threshold is created based on the data retrieved from the Social Medias content. Before proceeding to the content retrieving one must make sure that a single sign on id for Twitter and Gmail is used. Using twitters key and email id the mail content and twitter content can be extracted using Java Mail API and Twitter4j API. The type of data set can be categorized like inbox, sent items, mail chats, user's tweets, twitter chats and micro blogs maintained in our local server database. These social media contents are mined and extracted using Stanford NLP processing. The extracted topics of the user's contents are monitored in the server. POS tagging create the parts of speech of the each content of the user's data set. Stemming process is to group the similar types of words of the content like calling, call, called and callable, etc. Chunking process removes the common words filtering on the content like is, was, the, of, etc.

5.3. Monitoring user's activity using Gmail and Twitter dataset: The Server monitors every user's activity on Gmail and Twitter. Single user activity on the two different web applications can be identified and extracted using single sign on email ids. The sequential topic extraction on sequence of documents are extracted and grouped. The evolution of individual topics was in accordance with the existing methodology while sequential relations of topics in successive documents published by a specific user can be grasped by the proposed application. For a document stream, some STPs may occur frequently and thus reflect common behaviours of involved users. Beyond that, there may still exist some other patterns which are globally rare for the general population, but occur relatively often for some specific user or some specific group of users. They are called as User-aware Rare STPs (URSTPs). Compared to frequent ones, discovering them is especially interesting and significant. Theoretically, it defines a new kind of patterns for rare event mining, which is able to characterized, personalized and abnormal behaviours for special users. Practically, it can be applied in many real-life scenarios of user behaviour analysis.

5.4. Mining rare user sequential activities:-While monitoring and extracting the user's sequential topics, if illegal behaviours are involved, detecting and monitoring them is particularly significant for social security surveillance. It can be uncovered by URSTPs, as long as it satisfies the properties of both global rareness and local frequentness. That can be viewed as essential pieces of information for doubt and will trigger focused on examinations. Therefore, mining URSTPs is a good means for real-time user behaviour monitoring on the Internet. These ideas above are also applicable for another type of document streams, called browsed document streams, where Internet users behave as readers of documents instead of authors. In this case, STPs can characterize complete browsing behaviours of readers, so compared to statistical methods, mining URSTPs can better discover special interests and browsing habits of Internet users, and thus capable to give effective and context-aware recommendation for them. We implement the aware recommendation on admin dashboard. We highlight the rare user's activity and normal user's interest based on their social network.

6. METHODS USED

- NLP Process
- Key phrases and Aspect extraction.
- Rare pattern domain analysis

6.1 NLP(Natural Language Preprocessing):

NLP process includes two phases: i) POS tagging ii) Chunking. The input of pre-processing is user's documents and the output is a list of words and their POS labels. Because of the effectiveness and convenience, word segmentation and POS tagging is being done for the pre-processing stage. The purposes of two modules are as follows:

• Word Segmentation: The main function of this segmentation module is to identify and separate the tokens present in the text in such a way that every individual word, as well as every punctuation mark, will be a different token. The segmentation module considers words, numbers with decimals or dates in numerical

• POS tagging: The output of the segmentation module is taken as input by the POS tagging module. It tells you whether words are nouns, verbs, adjectives, etc., but it doesn't give you any clue about the structure of the sentence or phrases in the sentence. Standard set of tags are used to do POS tagging. One tag is assigned for each part of speech (Eg: N, V, Adj, Adv, Prep).

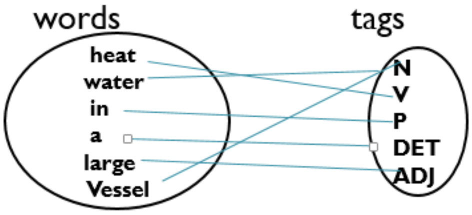

Fig- 2: POS Tagging

Chunking:

It points just to distinguish real constituents. What's more, does not endeavour to distinguish structure, neither (inside the piece), nor outside (between lumps).Chunking will leave some parts of the text unanalysed. Chunks can be represented like tags or like parse trees. Chunking is also called shallow parsing and it's basically the identification of parts of speech and short phrases (like noun phrases).

6.2 Key phrase and Aspect extraction:

Keywords, is a single word term, whereas keyphrase, implies a multi-word lexeme. They describe the content of the single documents and provide a kind of semantic metadata that is suitable for a wide variety of purposes. Before performing keyphrase extraction, key phrase indexing needs to be done. Steps involved:

1) Define a directory that contains all the documents to be indexed. 2) Vocabulary can be defined based on the type of indexing involved. 3) Length of the keyphrase are predefined. Keyphrases do not start or end with stop words defined in the vocabulary. It collects the words that are matched with the thesaurus, which defines the relationship between non-allowed terms and allowed terms. Pseudo phrase matching means removing stop words from the phrase, and then stemming and ordering the remaining words. 4) To collect the exact required keyphrases: [1]TF x IDF – measure describing the specificity of a term for this document under consideration, compared to all other documents. Extracted word, which has high TF x IDF esteem will probably be a Keyphrase. [2]Counting the first

occurrence of phrase in the document. Terms that tend to appear at the start or at the end of a document are more likely to be keyphrases. [3]Length of phrase is also effective factor, works on defining keyphrase. It might be single word, two word or more word combination. [4] Degree of a phrase is the number of phrases in the set that are semantically related to this phrase. This is computed with the help of the thesaurus. Phrases with high degree are more likely to be keyphrases. [6]The final step is extraction of keyphrase, they are started to be extracted from the defined documents. The degrees are counted and according to that measure final list of Keyphrases been retrieved.

6.3 Rare pattern domain analysis:

Rare pattern domain analysis is used to process huge amounts of information in order to extract hidden knowledge to be directly interpreted or exploited to feed other processes. This algorithm is used to discover patterns that can be of interest to a specific domain of application. A pattern is a collection of events/features that occur together in a database. As a matter of fact, different categories of patterns exist, such as sequences, item-sets, etc. To the contents obtained from the key phrase and aspect extraction phase, domain is found, by applying the pattern matching scheme. It is achieved by comparing key phrases with the one which is available in the data set. To filter out the patterns, the support and the confidence information's are used. While the support represents the number of times a pattern occurs in the initial database (frequency), the confidence represents a proportion value that shows how much of the time a piece of the example, called evidence happens among every one of the records containing the entire body. Herein, we classify pattern categories according to the specific use of the support threshold. In fact, by setting fields for the support, one can obtain different categories of patterns. For example, if we set a minimum support threshold that a pattern has to satisfy to be examined as an interesting pattern, we obtain what is called frequent pattern. By setting a maximum support threshold we obtain another category of patterns called rare patterns. Whereas frequent patterns target on mining patterns that appear more frequently in the database, rare patterns aim at discovering patterns that are less frequent. Finally Percentage calculation for rare sequence of that particular user is being computed. This content is made available to the admin via a dashboard (a user interface constantly gets updated eg: pattern either abnormal, normal or little suspicious)

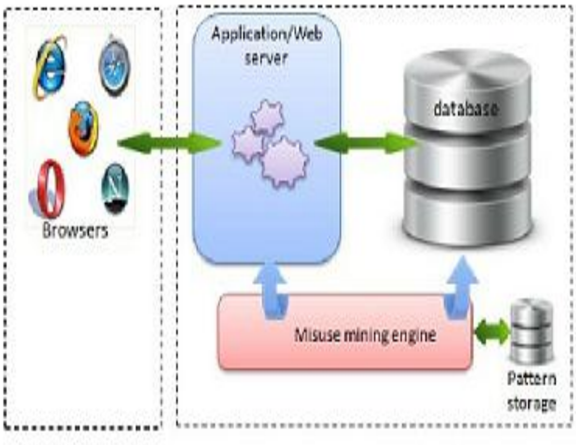

Fig-3: Pattern discovery system based on rare patterns.

7. ALOGRITHM

1. Locate the target word by the key phrase and aspect extraction and start to combine it with two or three surrounding words in order to discover if the target word is a compound one (based on the principle that most compound words are monosemous)

2. Extract all domains from text for each context word and count the frequency and also determine the rarity of target words.

3. Order the domain list obtained on previous step, in a decending order. For each sense of the target word,

- Declare and initialize variable Count with value 0,
- For each domain of this synset:
    - Increse Count with the appearing frequency of this domain on text.
- Assign to an array(Votes),the sense and the received vote(Count).

4. Select from array Votes, maximum value and return the sense associated with this value as the correct sense of the target word and also determine the words rarity by performing pattern matching with words created in the dataset and determine the percentage of rarity in the target words.

8. CONCLUSION AND FUTURE ENHANCEMENT

Mining URSTPs in published document streams on the internet is a remarkable and challenging problem. It constructs new event patterns which are intricate based on document topics, and has wide potential application scenarios, like real-time monitoring on abnormal behaviours of Internet users. This paper deals with various mining problems, and a group of algorithms being designed and combined to systematically solve the problems. The experiments are conducted on real-time applications like Twitter and Gmail to exhibit the proposed idea to prove it efficient and effective to unmask the aberrant behaviour of Internet users.

Since the paper emphasizes on web data mining, one can work to enhance the techniques to take it a notch up. The foremost drawback is it does not assist the detection in encryption of certain code words which is mandatory when the data to be communicated among Internet users are confidential. It indeed accommodates a secret message passing technique which cannot be monitored by the system. The languages into use are of major concern, where it has to be redefined thereby enriching the number of languages present, providing a wider arena for exposing the anomaly of a particular Internet user despite the location. STPs can characterize complete browsing behaviour's of readers, so compared to statistical methods, mining URSTPs can better discover special interests and browsing habits of Internet users, and is thus capable to give effective and context-aware recommendation for them. This application can be left as a recommendation for future work. Elaboration more on the user-aware rarity by obliging a variety of speculation, in order to enhance the mining algorithm to focus on degree of parallelism, and research on-the-fly algorithms targeting at real-time document streams.

REFERENCES:

[1] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in Proc. ACM SIGKDD, 2009, pp. 29–38.

[2] R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc.IEEE Int. Conf. Data Eng., 1995, pp. 3–14.

[3] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle,

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

[5] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlets allocation," J.Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[6] 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 Conf. Vis. Anal. Sci. Technol., 2012, pp. 143–152.

[7] 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,
Aug. 2007.

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

[9] Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in large uncertain databases," IEEE Trans.Knowl. Data Eng., vol. 26, no. 5, pp. 1171–1184, May 2014.

[10] 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 Proc. 33rd Eur. Conf. Adv. Inf. Retrieval, 2011, pp. 338–349.

[11] Q. Zhang, F. Li, and K. Yi, "Finding frequent items in probabilistic data," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008,
pp. 819–832.

[12] M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Mach. Learn., vol. 42, no. 1-2, pp. 31–60, 2001.[32] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng, "Mining uncertain data with probabilistic guarantees," in Proc. 16th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, 2010, pp. 273–282.

[13] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in web-usage analysis," INFORMS J. Comput., vol. 15, no. 2, pp. 171–190, 2003.

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

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

[16] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng, "Mining uncertain data with probabilistic guarantees," in Proc. 16th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, 2010, pp. 273–282.

[17] L. Wang, R. Cheng, S.D. Lee, and D. Cheung, "Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach," Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2010.

[18] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 1994 Int'l Conf. Very Large Data Bases (VLDB '94), pp. 487-499, Sept. 1994.

[19] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, *Sequential Pattern Mining Using a Bitmap Representation*, In Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 429-435, Edmonton, Alberta, Canada, Jul. 2002.

[20]D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp.113–120.