International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 09
August 2017

Mining User Aware Rare Sequential Topic Patterns in Document Streams

Ch. Dayakar Reddy
MCA, M.Tech, M.Phil, (Ph.D)
Professor and HOD MCA
CMR College of Engineering & Technology, Hyderabad

Mr. Siva Rama Krishna Reddy
MCA 3rd Year, II Sem
CMR College of Engineering & Technology, Hyderabad

ABSTRACT: *The advances display is empowered the entrance to printed records to Internet clients everywhere throughout the world easily. Consecutive examples are engaged subject in information mining. Finding the conduct of successive example are useful in finding many breaking down applications like foreseeing next occasion has been fundamental. Records made and circulated on the Internet are regularly changing in different structures. The greater part of existing works is given to point displaying and the development of individual subjects. Reports made and disseminated on the Internet changing many structures. The vast majority of backend works is given to subject displaying and the advancement of individual points. The advances of innovation additional time have empowered the entrance to printed reports to web clients. Successive examples have been a centered in information mining. Consecutive example are useful in finding many dissecting applications like foreseeing next occasion has been key. A great deal of explores of content mining concentrated on removing subjects from record accumulations and archive streams numerous probabilistic theme models. To accomplish this, an arrangement of calculations is introduced for pre-preparing the client substance, create all STP bolster esteems for proficient example development, and choosing client mindful uncommon successive subjects by utilizing uncommon example area investigation.*

Index Terms: Data Mining, Information Retrieval, Document Streams, Dynamic Programming, Pattern-Growth,

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 [1].

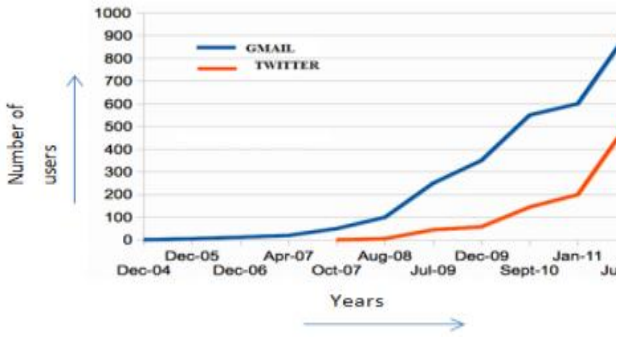

Fig no 1: Depicts Gmail and Twitter users

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 [2]. Traditional Information Retrieval (IR) has same objective of automatically retrieving as many relevant documents as possible, whilst filtering out irrelevant documents at the same time. However, IR-based systems do not provide users with what they really need. Many text mining methods have been developed for retrieving useful information for users [3]. Most text mining methods use keyword based approaches, whereas others choose the phrase method to construct a text representation for a set of documents. The phrase-based approaches perform better than the keyword-based as it is considered that more information is carried by a phrase than by a single term. New studies have been focusing on finding better text representatives from a textual data collection [4].

1. RELATED WORK

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. [5] 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 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 [6]. These strategies were intended to find visit successive examples whose backings are at least a client characterized edge minsupp. Notwithstanding, the acquired examples are not continually fascinating, on the grounds that those uncommon but rather noteworthy examples are pruned for their low backings. Moreover, the incessant successive example mining from deterministic databases is totally not the same as the STP mining that handles vulnerability of points [7]. Subsequent to extricating topics from archives by LDA and sorting the record stream into sessions for various clients amid various eras, the proposed calculations find uncommon STPs digging STP possibility for every client through a proficient calculation in view of example development, and creating client related uncommon STPs by example irregularity examination [8].

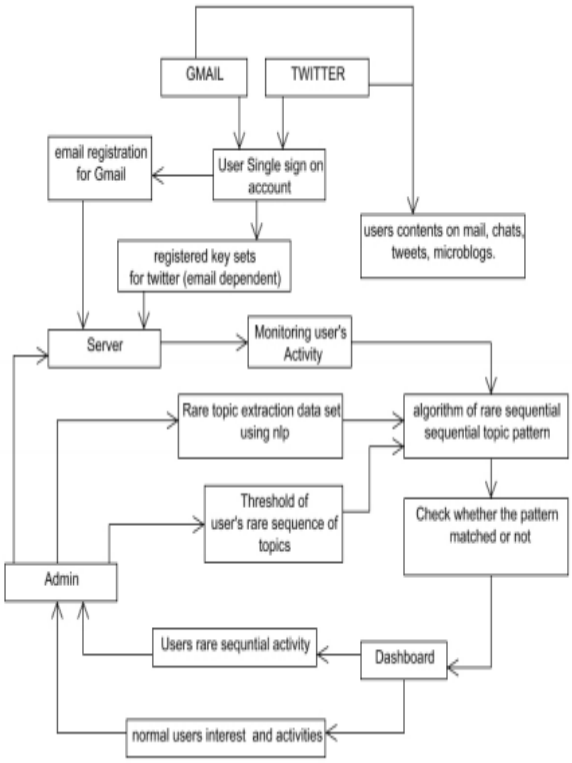

Fig No 2. Multi-Part Topic Model

2. SYSTEM ARCHITECTURE

Sequential Topic Patterns (STPs). Each of them records the complete and repeated behavior of a user when she is publishing a series. Topic mining in document collections has been extensively studied in the literature. Topic Detection and Tracking (TDT) task aimed to detect and track topics (events) in news streams with clustering-based techniques on keywords. The experiments conducted on both real (Twitter) and synthetic datasets demonstrate that the proposed approach is very effective

and efficient in discovering special users as well as interesting and interpretable URSTPs from Internet document streams, which can well capture users' personalized and abnormal behaviors and characteristics [9]. 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 [10]

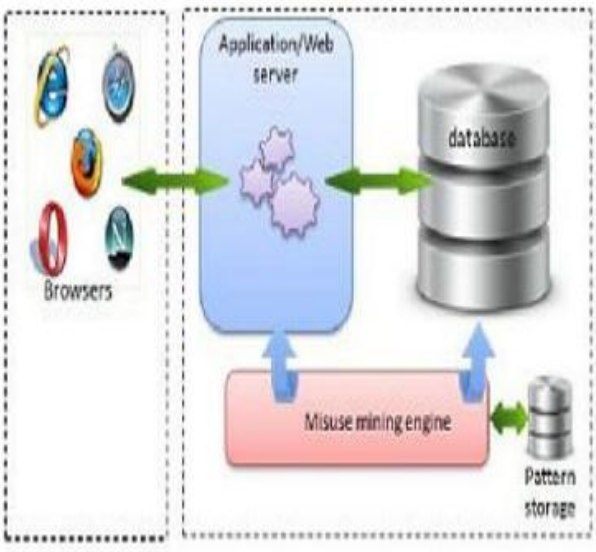

Fig-4:Pattern Discovery System

3. PROPOSED METHODOLOGY

In our proposed system, user rare and sequential activities are monitored using a sequence of document streams from multiple web applications. 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 [11]. We take preprocessing strategies with heuristic techniques for subject extraction and session ID. The exchange off amongst precision and proficiency. We exhibit a client mindful irregularity examination calculation as indicated by the formally characterized model to select URSTPs and related clients [12].

A. MATHEMATICAL MODEL

The Mathematical model this Query I1 is submitted to state q1 where the Data preparation is done then it is passed to state q2 where the Data is pre-processed then in state q3 the URSTP Mining is done and the output is generated in final state O [13].

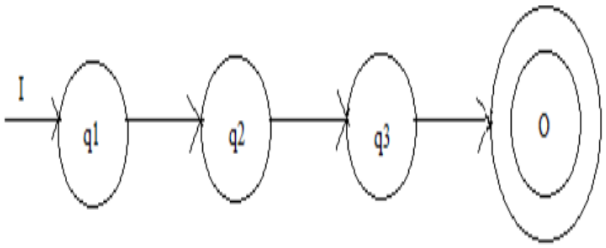

Figure-5: Mathematical Model of the Proposed System

Input Parameter (I)

I = I1 where I is set of Input. I1= It is the textual stream which is submitted to state q1.

Functional Parameter (Q)

Q=q1,q2,q3,q4

where Q is functions/process done in the URSTP mining.

q1 = Data preparation stage in which the document stream crawling is done.

q2 = Data pre-processing stage in this topic extraction is done and based on that sessions are being identified.

q3 = URSTP mining stage in this STP candidate discovery is done and user-aware rarity analysis is done [14]

Output Parameter (O)

O = O1

Where O is an Output parameter

O1 = Result generated [15].

B. 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 word segmentation and POS tagging is being done for the pre-processing stage.

• 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 [16].

• 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. Standard set of tags are used to do POS tagging. One tag is assigned for each part of speech [17].

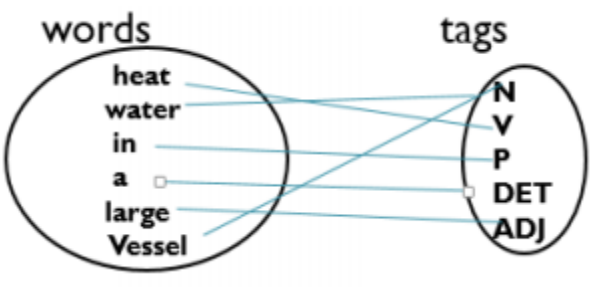

Fig- 2: POS Tagging

C. 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

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 descending order. For each sense of the target word,

4. Select from array Votes, maximum value and return the sense associated with this value as the correct sense of the target word

5. CONCLUSION

It indeed accommodates a secret message passing technique which cannot be monitored by the system The proposed display Maximum coordinated Pattern-based Topic Model comprises of subject disseminations depicting point inclinations of every record or the archive accumulation and example based theme representations speaking to the semantic significance of every theme. Here suggested that an organized example based point representation. 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. As this paper puts forward an innovative research direction on Web data mining, much work can be built on it in the future.

6. FEATURE ENHANCEMENT

We open up the lower parts of the two outlines. And estimation calculation is in reality somewhat speedier particularly for bigger scales. Notice that every execution of the sub strategy is only for one client, so when the client number expands the time distinction for the entire methodology will turn out to be increasingly obvious, even with some degree of parallelism. Subsequently, together with the outcomes,

we can reason that the two calculations have their particular points of interest proper for the genuine undertaking mirrors an exchange off between mining precision and execution speed, and ought to rely on upon the particular prerequisites of utilization situations.

7. REFERENCES

[1] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in Proc. ACM SIGKDD'09, 2009, pp. 29–38.

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

[3] Y. Li, J. Bailey, L. Kulik, and J. Pei, "Mining probabilistic frequent spatio - temporal sequential patterns with gap constraints from uncertain databases," in Proc. IEEE ICDM'13, 2013, pp. 448 – 457.

[4] 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.

[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, Aug. 2007.

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

[7] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD, 2009, pp. 119–128.

[8] 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.

[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] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in Proc. IEEE VAST'12, 2012, pp. 93–102.

[12] 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, 2005, pp. 181–192.

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

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

[15] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD, 2009, pp. 119–128.

[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 ModelBased 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.