AN ENHANCED PROBABILITY BASED CONTENT SEARCH METHOD USING DATAMINING

[1] S. Niranjana Devi, [2] K. Krishnaveni, [3] S. Chidambaram.

1. UG Student, Dept .of IT, National Engineering College, Kovilpatti.

2. UG Student, Dept. of IT, National Engineering College, Kovilpatti.

3. Asst.Professor (Senior Grade), Dept.of IT, National Engineering College, Kovilpatti.

ABSTRACT

Among various multi keyword and sequential topic semantics, is selected for implementation the efficient similarity measure of "coordinate matching", i.e., as many matches to capture the relevance of data documents to the search query. In this paper we have used inner product similarity, number of query keywords appearing in a document, to quantitatively evaluate such similarity measure of that document to the search query. By investigating privacy and efficiency it guarantees the schemes of the proposed system of the real-world data set and further show the proposed schemes indeed introduce low overhead on computation and communication. We have implemented STP (Sequential Topic Pattern) which captures both combinations and orders of topics when compared to other document-based patterns, such as topic-based patterns which contains abstract information of document contents. We proposed a probability model that can capture the normal mentioning of a user, which comprises of both the number of users per post and the frequency of users occurring in viewing the post. The proposed technique can detect changes in the user patterns even in a realistic setting when only some part of the users react to the emerging topic.

I. INTRODUCTION

Document streams are created and distributed in various forms on the Internet, such as news streams, emails, micro-blog articles, chatting messages, research paper archives, web forum discussions, and so forth. The contents of these documents generally concentrate on some specific topics, which reflect offline social events and users characteristics in real life. To mine these pieces of information, a lot of researches of text mining focused on extracting topics from document collections and document streams through various probabilistic topic models. During the index construction, each document is associated with a binary vector as a sub-index where each bit represents whether corresponding keyword is contained in the document. Thorough analysis investigating privacy and efficiency guarantees of the proposed schemes is given, and experiments on the real-world dataset further show the proposed schemes indeed introduce low overhead on computation and communication. We implemented STP (Sequential Topic Pattern) which captures both combinations and orders of topics and compared to document-based patterns, topic-based patterns contain abstract information of document contents and are thus beneficial in clustering similar documents and finding some regularities about Internet users. STPs happen to be able to combine a series of inter-correlated messages, and can thus capture such behaviors and associated users. We propose a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users

occurring in the mentions. Then this model is used to measure the anomaly of future user behavior. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. We aggregate the anomaly scores obtained in this way over hundreds of users and apply a recently proposed change point detection technique based on the sequentially discounting normalized maximum-likelihood coding.

- Sequential Topic Patterns (STPs) and formulate the problem of mining
- User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet.

They are rare on the whole but relatively frequent for specific users, so can be applied in many real-life scenarios, such as real-time monitoring on abnormal user behaviors. We present a group of algorithms to solve this innovative mining problem through three phases: preprocessing to extract probabilistic topics and identify sessions for different users, generating all the STP candidates with (expected) support values for each user by pattern-growth, and selecting URSTPs by making user-aware rarity analysis on derived STPs. Experiments on both real (Twitter) and synthetic datasets show that our approach can indeed discover special users and interpretable URSTPs effectively and efficiently, which significantly reflect users' characteristics.

Taking advantage of these extracted topics in document streams, most of existing works characterize user behaviors in published document streams, we studied on the correlations among topics extracted from these documents, especially the sequential relations, and specify them as Sequential Topic Patterns (STPs). Each of them records the complete and repeated behavior of a user when he/she is publishing a series of documents, and are suitable for inferring users' intrinsic characteristics and psychological statuses. First, compared to individual topics, STPs capture both combinations and orders of topics, so can serve well as discriminative units of semantic association among documents in ambiguous situations. Second, compared to document-based patterns, topic-based patterns contain abstract information of document contents and are thus beneficial in clustering similar documents and finding some regularities about Internet users. Third, the probabilistic description of topics helps to maintain and accumulate the uncertainty degree of individual topics, and can thereby reach high confidence level in pattern matching for uncertain data. For a document stream, some STPs may occur frequently and thus reflect common behaviors 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. We call them 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 characterize personalized and abnormal behaviors for special users. Practically, it can be applied in many real-life scenarios of user behavior analysis, individual topics to detect and predict social events as well as user behaviors. However, few researches paid attention to the correlations among different topics appearing in successive documents published by a specific user, so some hidden but significant information to reveal personalized behaviors has been neglected.

The idea of paper [2] 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 behaviors 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. While, this paper will concentrate on published document streams and leave the applications for recommendation to future work. To solve this innovative and significant problem of mining URSTPs in document streams, many new technical challenges are raised and will be tackled in this paper. First, the input of the task is a textual stream, so existing techniques of sequential pattern mining for probabilistic databases cannot be directly applied to solve this problem. A preprocessing phase is necessary and crucial to get abstract and probabilistic descriptions of documents by topic extraction, and then to recognize complete and repeated activities of Internet users by session identification. Second, in view of the real-time requirements in many applications, both the accuracy and the efficiency of mining algorithms are important and should be taken into account, especially for the probability computation process. Third, different from frequent patterns, the user aware rare pattern concerned here is a new concept and a formal criterion must be well defined, so that it can effectively characterize most of personalized and abnormal behaviors of Internet users, and can adapt to different application scenarios. And correspondingly, unsupervised mining algorithms for this kind of rare patterns need to be designed in a manner different from existing frequent pattern mining algorithms.

In the rest of this paper, Section 2 reviews related works including topic mining and sequential pattern mining for deterministic and uncertain databases using some references. In Section 3,describes the key definitions related to STPs, and formulate the problem of mining URSTPs in document streams using the probability model through problem definition. Section 4 shows the comparison of existing and proposed methods. Section5 shows the experimental results on real datasets, and leaves the synthetic results. Section 6 concludes the paper and discusses future directions.

II. RELATED WORKS:

Y. Li, J. Bailey in [1] demonstrated that the problem of mining probabilistic frequent spatio-temporal sequential patterns in uncertain databases. This proposed a dynamic programming approach for computing the frequentness probability with linear time complexity. This is a somewhat surprising result by using Apriori-based algorithms and Clustering algorithm. However it has the limitations of Uncertainty is common in real-world applications, for example, in sensor networks and moving object tracking, resulting in much interest in item set mining for uncertain transaction databases. So we planned extend current approach to be able to handle the trajectory data where the identity of objects is uncertain.

C. H. Mooney in [2] demonstrated that an algorithm to quickly find all frequent sequences in a list of transactions. The algorithm utilizes a depth-first traversal of the search space combined with a vertical

bitmap representation to store each sequence using Sequential pattern mining algorithm. Building on a state-of-the-art sequential pattern mining algorithm Prefix Span for mining transaction databases.

Z. Hug, H. Wang in [3] demonstrated that a group of new algorithms proposed to solve the mining problem. The experiments based on both synthetic and real data sets show that the proposed approach is very effective and efficient for discovering interesting and interpretable rare STPs from document streams on the Internet using recommendation algorithm, pattern-growth based algorithm and Sequential Topic Patterns algorithm. However it has the limitations of Plain text documents created and distributed on the Internet are ever changing in various forms. Mining topics of these documents has significant applications in many domains. Most of the literature is devoted to topic modeling, while sequential patterns of topics in document streams are ignored. Moreover, traditional sequential pattern mining algorithms mainly focused on frequent patterns for deterministic data sets, and thus not suitable for document streams with topic uncertainty and rare patterns.

X. Yan in [4] demonstrated that a novel probabilistic topic model for short texts, namely biterm topic model (BTM). BTM can well capture the topics within short texts as it explicitly models the word co-occurrence patterns and uses the aggregated patterns in the whole corpus using Gibbs sampling algorithm and greedy algorithm. However it has the limitations of Short texts are popular on today's web, especially with the emergence of social media. Inferring topics from large scale short texts becomes a critical but challenging task for many content analysis tasks. Conventional topic models such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA) learn topics from document-level word co-occurrences by modeling each document as a mixture of topics, whose inference suffers from the sparsity of word co-occurrence patterns in short texts.

Z. Zhao in [5] demonstrated that the problem of mining probabilistically frequent sequential patterns (or p-FSPs) in uncertain databases.Our proposed there are two new U-PrefixSpan algorithms to mine p-FSPs from data that conform to our sequence level and element-level uncertain sequence models using PrefixSpan algorithm and pattern-growth algorithm. However it has the limitations of Data uncertainty is inherent in many real-world applications such as environmental surveillance and mobile tracking. Mining sequential patterns from inaccurate data, such as those data arising from sensor readings and GPS trajectories, is important for discovering hidden knowledge in such applications.

N. Hariri in [6] demonstrated our approach discovers patterns at a more abstract level rather than songs. This generalization makes it easier to track and detect any changes in the users' preferences. Also, it is useful in handling the cold start problem where a new song hasn't occurred in the training data using recommendation algorithm and content-based algorithm. However it has the limitations of Contextual

factors can greatly influence the users' preferences in listening to music.

III. PROBLEM DEFINITION

Most of existing works analyzed the evolution of individual topics in topic based as well as document based to detect and predict social events as well as user behaviors. Many mining algorithms have been proposed based on support, such as Prefix Span, Free Span and SPADE. They discovered frequent sequential patterns whose support values are not less than a user-defined threshold, and were extended by SLPMiner to deal with length decreasing support constraints.Muzammal et al. focused on sequence-level uncertainty in sequential databases, and proposed methods to evaluate the frequency of a sequential pattern based on expected support, in the frame of candidate generate-and- test or pattern-growth.

IV. EXISTING SEQUENTIAL PATTERN MINING ALGORITHM:

Textual documents created and distributed on the Internet are ever changing in various forms. Most of existing works are devoted to topic modeling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. Most of existing works on sequential pattern mining focused on frequent patterns, but for STPs, many infrequent ones are also interesting and should be discovered. Specifically, when Internet users' publish documents, the personalized behaviors characterized by STPs are generally not globally frequent but even rare, since they expose special and abnormal motivations of individual authors, as well as particular events having occurred to them in real life. The existing methodology is based on textual documents are crawled from some micro-blog sites or forums, and constitute a document stream as the input of our approach. Then, as preprocessing procedures, the original stream is transformed to a topic level document stream and then divided into many sessions to identify complete user behaviors. Finally and most importantly, we discover all the STP candidates in the document stream for all users, and further pick out significant URSTPs associated to specific users by user-aware rarity analysis. In order to fulfill this task they use a group of algorithms

4.1 DP-Based Algorithm

Similar to, the occurrence probability of an STP in a session (a sequence of topic-level documents) can be computed by dynamic programming. During the search process two statistical knowledge sources are combined: a translation model and a bigram language model. This search algorithm expands hypotheses along the positions of the target string while guaranteeing progressive coverage of the words in the source string.

V. PROPOSED METHODOLOGY:

We proposed a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. This model is used to measure the anomaly of future user behavior. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. We aggregate the anomaly scores obtained in this way over hundreds of users and apply a recently proposed change point detection technique based on the sequentially

discounting normalized maximum-likelihood coding.

A taxonomy of existing sequential pattern-mining algorithms, which lists the algorithms, showing a comparative analysis of their different important features. This proposed taxonomy is composed of three main categories. of sequential pattern-mining algorithms, namely, Apriori-based, pattern-growth and early-pruning algorithms.. Frequent sequential pattern discovery can essentially be thought of as association rule discovery over a temporal database. While association rule discovery [Agrawal et al. 1993] covers only intratransaction patterns (itemsets), sequential pattern mining also discovers inter transaction patterns (sequences), where ordering of items and itemsets is very important, such that the presence of a set of items is followed by another item in a time-ordered set of sessions or transactions. The set of all frequent sequences is a superset of the set of frequent itemsets. Due to this similarity, the earlier sequential pattern-mining algorithms were derived from association rule mining techniques. The first of such sequential pattern-mining algorithms is the **AprioriAll algorithm** [Agrawal and Srikant 1995], derived from the **Apriori algorithm** [Agrawalet al. 1993; Agrawal and Srikant 1994]. An algorithm can fall into one or more (hybrid algorithms) of the categories in the proposed taxonomy. Algorithms mainly differ in two ways:

(1) **Breadth-first search**: AP-pattern growth algorithms are described as breath-first (level-wise) search algorithms because they construct all k-sequences together in each k^{th} iteration of the algorithm as they traverse the search space.

(2) **Generate-and-test**: This feature is introduced by the **AP-pattern growth** algorithm [Agrawalet al. 1993] and is used by the very early algorithms in sequential pattern mining. In BIDE (an algorithm for mining frequent closed sequences [Wang and Han 2004]), it is referred to as "maintenance-and-test". It entails using exhaustive join operators(e.g., Apriori-generate, GSP-join), in which the pattern is simply grown one item at a time and tested against the minimum support.

(3) **Multiple scans of the database**: This feature entails scanning the original database to ascertain whether a long list of generated candidate sequences is frequent or not. It is a very undesirable characteristic of most apriori-based algorithms and requires a lot of processing time and I/O cost. A solution to this limitation is to scan the database only once or twice to create a temporary data structure, which holds support information used during mining. PSP [Masseglia et al. 1999] uses a prefix tree to hold candidate sequences along with the support count for each sequence at the end of each branch that represents it. This method (along with GSP, from which it is adopted) becomes very inefficient when the support threshold is very low, making it a poor choice for web log mining. Apriori-GST [Tanasa 2005], on the other hand, employs a generalized suffix tree as a hash index to calculate the support count of stored subsequences.

An attentive investigation of Apriori-GST makes one wonder why the author did not

use the suffix tree for mining instead of just holding the support information and depending on a variation of apriori for mining. This helps reduce I/O access costs or extra storage requirements. Table.1 lists some top words as well as simple descriptions of involved topics.

TABLE 1: Comparison of approximate match type and proposed keyword match type

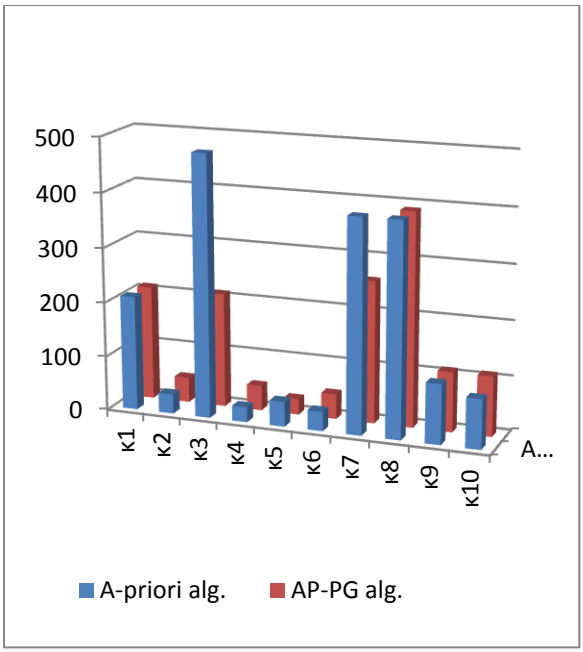
Query (κ)	A-priori approximate match type(α)	APG algorithm keyword match type(β)	Absolute difference (η)	Error %
κ1	**210**	**210**	**0**	**0.00%**
κ2	**36**	**46**	**10**	**27.7%**
κ3	**480**	**210**	**270**	**56.25 %**
κ4	**28**	**46**	**18**	**64.29 %**
κ5	**46**	**29**	**18**	**39.13 %**
κ6	**36**	**46**	**10**	**27.77 %**
κ7	**390**	**260**	**130**	**33.33 %**
κ8	**390**	**390**	**0**	**0.00%**
κ9	**110**	**110**	**0**	**0.00%**
κ10	**91**	**110**	**19**	**20.97 %**

TABLE 3: top words and simple & description of topics

S.NO	Top words	Description
1	Disease prediction,treatment,health care, nutrient diet,proverb wealth, balanced diet, fitness centre,beauty tips, nutrition loss, proteins minerals,vitamins upplements,nutritional life	Health
2	Web browser,famous search engine, Larry page,sergy brin,Sundar Pichai,American multinational company providing internet related services	Google

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS:

We have evaluated the efficiency of the approach on synthetic datasets, and compare the two alternative search queries of STP candidate discovery to demonstrate the tradeoff between accuracy and efficiency.



The obtained sets of element-level probability sequences conform to the topic-level document stream defined in this paper. Notice that the stream has already been divided into sessions, so the preprocessing phase is not required here, and the test will thus concentrate on the mining algorithms.

CONCLUSION AND FUTURE WORK:

In our work we have implemented, several new concepts and the mining problem are formally defined, and a group of algorithms are designed and combined to systematically solve this problem. The experiments conducted on both real (Twitter) and synthetic datasets demonstrate that the proposed approach is very effective

Our paper also provides a framework to pragmatically solve this problem, and design corresponding algorithms to support it. We give preprocessing procedures with heuristic methods for topic extraction and session identification. Then, borrowing the ideas of pattern-growth in uncertain environment, two alternative algorithms are designed to discover all the STP candidates with support values for each user. That provides a trade-off between accuracy and efficiency.

REFERENCES:

[1]Y. Li, Et al. "Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases", in Proc. IEEE ICDM'13, 2013

[2] C. H. Mooney and J. F. Roddick in ACM Comput. "Sequential pattern mining - approaches and algorithms", Surv., vol. 45, no. 2, pp. 19:1–19:39, 2013

[3] Z. Hu, Et al."Discovery of rare sequential topic patterns in document stream", in Proc. SIAM SDM'14, 2014

[4] X. Yan, Et al. "A biterm topic model for short texts" in Proc. ACM WWW'13, 2013

[5] Z. Zhao, Et al. "Mining probabilistically frequent sequential patterns in large uncertain databases" , IEEE Trans. Knowl. Data Eng., vol. 26, no. 5, pp. 1171–1184, 2014

[6] N. Hariri, B. Mobasher, and R. Burke "Context-aware music recommendation based on latent topic sequential patterns", Proc. ACM RecSys'12, 2012

[7] Wenwen Dou, Et al. "Interactive Visual Analysis of Text Data through Event Identification and Exploration", vol. 36, no. 3, pp. 20:1-13,2011.

[8] M. Muzammal, "Mining sequential patterns from probabilistic databases by pattern-growth," inProc. BNCOD, 2011, pp. 118–127.

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

[10] Z. Zhang, Et al. "Mining evolutionary topic patterns in community question answeringsystems," IEEE Trans. Syst., Man, Cybern. A, vol. 41, no. 5, pp. 828–833, Sep. 2011.

[11] Jiaqi Zhu, Et al."Mining user aware rare sequential topic patterns in document streams",IEEE Transactions on knowledge and Data Engineering, 2016.

[12]Suvrit Sra,Member, IEEE, ICDM, "Incremental aspects models for mining document streams",in proc.ACM WWW'13, 2013.

[13] Domingos P. And Hulten G."Mining high-speed data streams", in Proc of the ACM KDD Conference. vol. 41, no. 5, pp. 828-833, Sep. 2000

[14] Dasu T.,Krishnana S.,"An information-theoretic approach to detecting changes in multidimensional data streams", in vol. 26, no. 5, pp. 1154–1184, 2011.

[15] Kifer D.,David S.B., Gehrke J."Detecting change in data streams" in surv., vol. 45, no. 2, pp. 19:1-19:39,2013.

[16] J.Chae,Et Al."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

[17] K. Chen,L.Lueskprasert Et Al."Hot topic extaction basedon timeline analysisand multimensional sentence modeling,"IEEE Trans. Knowl. Data Eng., Vol. 19, No. 8, pp. 1016–1025, Aug. 2007

[18] C.K.Chui And B.Kao,"A determental approach for mining drequent itemset from uncertain data",In Proc.12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2008, pp. 64-75

[19] T. Bernecker,Et Al."Probabilistic Frequent Itemset Mining In Uncertain Databases" In Proc.ACM SIGKDD, 2009, pp. 119-128.

[20] D. Blei And J. Lafferty,"Correlated Topic Models," Adv. Neural Inf. Process. Syst., Vol. 18, pp.147-154, 2006.

[21]S.Chidambaram,S.Esakkiraj,"A predictive approach for fraud detection using hidden markov model", international jurnal of engineering research and technology, vol.2,issue 1, 2013.

[22]S.Chidambaram,S.Esakkiraj,"Comprehe nsive survey on decision tree alagorithm in datamining",int. conf. on E-Goernance and cloud computing services,vol.1,2012.