ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Analysis of Efficient Way to Identify User Aware Rare Sequential Pattern in Document Stream

Swati V.Mengje, Prof. Rajeshri R. Shelke

ME Second Yr (CSE) H.V.P.M's COET, Amravati, S.G.B Amt University, Maharashtra, India

Assistant Professor, ME (CSE), Ph.D (Pursuing) H.V.P.M's COET, Amravati, S.G.B Amt University, Maharashtra, India

ABSTRACT : 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. In order to characterize and detect personalized and abnormal behaviours of Internet users, we propose 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 behaviours. Here present solutions to solve this innovative mining problem through three phases: pre-processing 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.

KEYWORDS: Web mining, sequential patterns, document streams, rare events, pattern-growth, dynamic programming.

I.INTRODUCTION

Sequential Pattern Mining is the method of finding interesting sequential patterns among the large databases. It also finds out frequent sub sequences as patterns from a sequence database. Enormous amounts of data are continuously being collected and stored in many industries and they are showing interests in mining sequential patterns from their database. Sequential pattern mining has broad applications including web-log analysis, client purchase behaviour analysis and medical record analysis [2].

Sequential or sequence pattern mining is the task of finding patterns which are present in a certain number of instances of data. The identified patterns are expressed in terms of sub sequences of the data sequences and expressed in an order that is the order of the elements of the pattern should be respected in all instances where it appears. If the pattern is considered to be frequent if it appears in a number of instances above a given threshold value, usually defined by the user, then it is considered to be frequent.

There may be huge number of possible sequential patterns in a large database. Sequential pattern mining identifies whether any relationship occurs in between the sequential events. The sequential patterns that occur in particular individual items can be found and also the sequential patterns between different items can be found. The number of sequences can be very large, and also the users have different interests and requirements. If the most interesting sequential patterns are to be obtained, usually a minimum support is pre-defined by the users. In this paper, we focus on the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in sequence of sets. Mining sequential patterns has become an important data mining task with broad applications [9].

For example, supermarkets often collect customer purchase records in sequence databases in which a sequential pattern would indicate a customer's buying habit. Sequential pattern mining is commonly defined as finding the complete set of frequent subsequence's in a set of sequences [1]. Much research has been done to efficiently find such

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

patterns. But to the best of our knowledge, no research has examined in detail what patterns are actually generated from such a definition. In this paper, we examined the results of the support framework closely to evaluate whether it in fact generates interesting patterns[4].

II. EXISTING SYSTEM

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. Taking advantage of these extracted topics in document streams, most of existing works analyzed the evolution of individual topics to detect and predict social events as well as user behaviours[3]. 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. 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. 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[10].

III. LITERATURE REVIEW & RELATED WORK ON STRING PATTERN MATCHING

Topic mining in document collections has been extensively studied in the literature. Topic Detection and Tracking (TDT) task [3], [9] aimed to detect and track topics (events) in news streams with clustering-based techniques on keywords. Considering the co-occurrence of words and their semantic associations, a lot of probabilistic generative models for extracting topics from documents were also proposed, such as PLSI , LDA [7] and their extensions integrating different features of documents [5] as well as models for short texts like Twitter-LDA . In many real applications, document collections generally carry temporal information and can thus be considered as document streams. Various dynamic topic modelling methods have been proposed to discover topics over time in document streams [6], and then to predict offline social events [8]. However, these methods were designed to construct the evolution model of individual topics from a document stream, rather than to analyze the correlations among multiple topics extracted from successive documents for specific users.

Sequential pattern mining is an important problem in data mining, and has also been well studied so far. In the context of deterministic data, a comprehensive survey can be found. The concept support is the most popular measure for evaluating the frequency of a sequential pattern, and is defined as the number or proportion of data sequences containing the pattern in the target database. Many mining algorithms have been proposed based on support, such as *PrefixSpan* [15], *FreeSpan* [12] 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.

Topic mining has been extensively studied in the literature. Topic Detection and Tracking (TDT) task [3] aimed to detect and track topics (events) in news streams with clustering-based techniques. Many generative topic models

were also proposed, such as Probabilistic Latent Semantic Analysis (PLSA) [11], Latent Dirichlet Allocation (LDA) [5] and their extensions.

In many real applications, text collections carry generic temporal information and therefore can be considered as a text stream. To obtain the temporal dynamics of topics, various dynamic topic modeling methods have been proposed to discover topics over time in document streams [6]. However, these methods were designed to extract the evolution model of individual topics from a document stream, rather than to analyze the relationship among extracted topics in successive documents for specific users.

Sequential pattern mining has been well studied in the literature in the context of deterministic data, but not for topics with uncertainty. The concept support is the most popular criteria for mining sequential patterns. It evaluates frequency of a pattern and can be interpreted as occurrence probability of the pattern.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Many methods have been proposed to solve the problem of sequential pattern mining based on *support*, such as PrefixSpan [16], FreeSpan [9] and SPADE. These methods were designed to discover frequent sequential patterns whose supports are not less than a user-defined threshold minsupp. However, the obtained patterns are not always interesting, because those rare but significant patterns are pruned for their low supports. Furthermore, the frequent sequential pattern mining from deterministic databases is completely different from the STP mining that handles uncertainty of topics. Few researches addressed the problem of sequential pattern mining on uncertain data. Muzammal and Raman [10] proposed a method to discover frequent sequential patterns from probabilistic databases and evaluated the frequency of a pattern based on the expected support. However, the data model cannot be applied to topic sequences. In addition, they focused on the frequent pattern mining and failed to discover interesting rare patterns for some users.

In this section, we propose a novel approach to mining URSTPs in document streams..It consists of three phases. At first, 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 behaviours. 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.

IV. PROPOSED WORK AND OBJECTIVES

The proposed method is outlined in Fig. 1 and comprises three main components: pattern extraction, alias extraction and ranking. Using a seed list of name-alias pairs, we will first extract lexical patterns that are frequently used to convey information related to aliases on the web. The extracted patterns are then used to find candidate aliases for a given name. We will define various ranking scores using the hyperlink structure on the web and page counts retrieved from a search engine to identify the correct aliases among the extracted candidates[12].

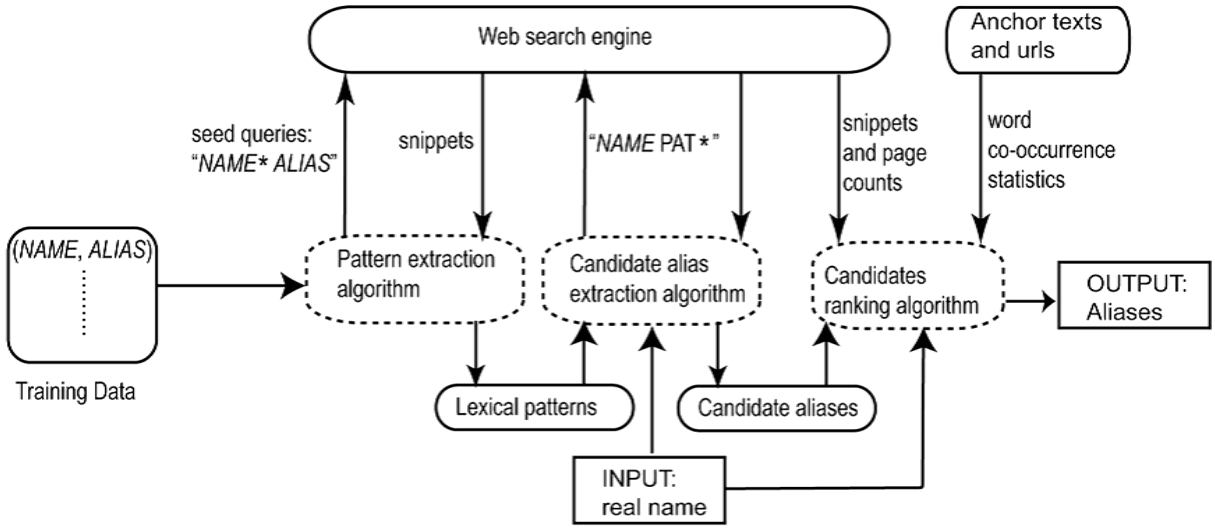

Fig. 1 Proposed Method

We proposed a fully automatic method to discover aliases of a given personal name from the web. Our contributions can be summarized as follows:

- We propose a lexical pattern-based approach to extract aliases of a given name using snippets returned by a web search engine. The lexical patterns are generated automatically using a set of real world name alias data. We evaluate the confidence of extracted lexical patterns and retain the patterns that can accurately discover

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

aliases for various personal names[8]. Our pattern extraction algorithm does not assume any language specific pre-processing such as part-of-speech tagging or dependency parsing, etc., which can be both inaccurate and computationally costly in web-scale data processing.

- To select the best aliases among the extracted candidates, we propose numerous ranking scores based upon three approaches: lexical pattern frequency, word co-occurrences in an anchor text graph, and page counts on the web. Moreover, using real-world name alias data, we train a ranking support vector machine to learn the optimal combination of individual ranking scores to construct a robust alias extraction method[15].
- We conduct a series of experiments to evaluate the various components of the proposed method. We compare the proposed method against numerous baselines and previously proposed name alias extraction methods on three data sets: an English personal names data set, an English place names data set, and a Japanese personal names data set. Moreover, we will evaluate the aliases extracted by the proposed method in an information retrieval task and a relation extraction task[6].

V. DESIRED IMPLICATIONS

-It can be useful tool for analyzing the name of person from his alias name.
-In this it can be predictable step for detection of any kind of cyber crime.

We will utilize both lexical patterns extracted from snippets retrieved from a web search engine as well as anchor texts and links in a web crawl. Lexical patterns can only be matched within the same document. In contrast, anchor texts can be used to identify aliases of names across documents. The use of lexical patterns and anchor texts, respectively, can be considered as an approximation of within document and cross-document alias references. By combining both lexical patterns based features and anchor text-based features, better performance in alias extraction

will be achieved [5]. It can only incorporate first order co-occurrences. An alias might not always uniquely identify a person. For example, the alias Bill is used to refer many individuals who has the first name William. The namesake disambiguation problem focuses on identifying the different individuals who have the same name. The existing namesake disambiguation algorithms assume the real name of a person to be given and does not attempt to disambiguate people who are referred only by aliases. The knowledge of aliases is helpful to identify a particular person from his or her namesakes on the web[7]. Aliases are one of the many attributes of a person that can be useful to identify that person on the web. Extracting common attributes such as date of birth, affiliation, occupation, and nationality have been shown to be useful for namesake disambiguation on the web.

VI. CONCLUSION

Mining user-related rare Sequential Topic Patterns (STPs) in document streams on the Internet is an innovative challenging problem. It formulates a new kind of patterns for uncertain complex event detection and inference, and has wide potential application fields, such as personalized context-aware recommendation and real-time monitoring on abnormal user behaviours on the Internet. Due to the continuous addition of large amount of data in the databases, the idea of sequential pattern mining is becoming popular.As this paper puts forward an innovative research direction on Web data mining; much work can be built on it in the future. At first, the problem and the approach can also be applied in other fields and scenarios. Especially for browsed document streams, we can regard readers of documents as personalized users and make context-aware recommendation for them.

REFERENCES

[1] [Guha & Garg, 2004] R. Guha and A. Garg, "Disambiguating People in Search," *Technical report, Stanford University*, 2004.
[2] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for PeopleSearching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570,2005.
[3] G. Mann and D. Yarowsky, "Unsupervised Personal NameDisambiguation," Proc. Conf. Computational Natural LanguageLearning (CoNLL '03), pp. 33-40, 2003.
[4] R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of People in a Social Network," Proc. Int'l World Wide WebConf. (WWW '05), pp. 463-470, 2005.
[5] P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04),2004.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

[6] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H.Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An Advanced Social Network Extraction System," Proc. WWW '06, 2006.
[7] C. Galvez and F. Moya-Anegon, "Approximate Personal Name-Matching through Finite-State Graphs," J. Am. Soc. for Information Science and Technology, vol. 58, pp. 1-17, 2007.
[8] T. Hokama and H. Kitagawa, "Extracting Mnemonic Names of People from the Web," Proc. Ninth Int'l Conf. Asian Digital Libraries(ICADL '06), pp. 121-130, 2006.
[9] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
[10] A. Bagga and B. Baldwin, "Entity-Based Cross-Document Coreferencing
Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), pp. 79-85, 1998.
[11] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," Proc.Conf. Empirical Methods in Natural Language (EMNLP '04), 2004.
[12] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," Proc. Int'l Semantic Web Conf. (ISWC '05), 2005.
[13] S. Sekine and J. Artiles, "Weps2 Evaluation Campaign: Overview of the Web People Search Attribute Extraction Task," Proc. Second Web People Search Evaluation Workshop (WePS '09) at 18th Int'l World Wide Web Conf., 2009.
[14] G. Salton and M. McGill, Introduction to Modern, Information Retreival. McGraw-Hill Inc., 1986.
[15] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, pp. 206-214, 1998.
[16] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining sequential patterns by prefixprojected growth," in *Proc. IEEE ICDE'01*, 2001, pp. 215–224.l of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012