Document Steams Mining By Using Sequential Topic Pattern

Sayyad Zulfin A. [1], Kaute Poonam H.[2], Avhad Pallavi R.[3], Tajanpure Vaishnavi S.[4], Prof. P.V.Waje [5]

[1,2,3,4] *BE Student, Information Technology Department, SVIT Chincholi, Nashik*
[5] *Professor, Information Technology Department, SVIT Chincholi, Nashik*

Abstract

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. In this paper, in order to characterize and detect personalized and abnormal behaviors 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 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.

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

---

1. Introduction

1.1 How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

2. Literature Survey:

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. 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 and their extensions integrating different features of documents 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 modeling methods have been proposed to discover topics over time in document streams and then to predict offline social events .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. 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. 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 and their extensions integrating different features of documents 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 modeling methods have been proposed to discover topics over time in document streams and then to predict offline social events .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.

3. System Architecture:

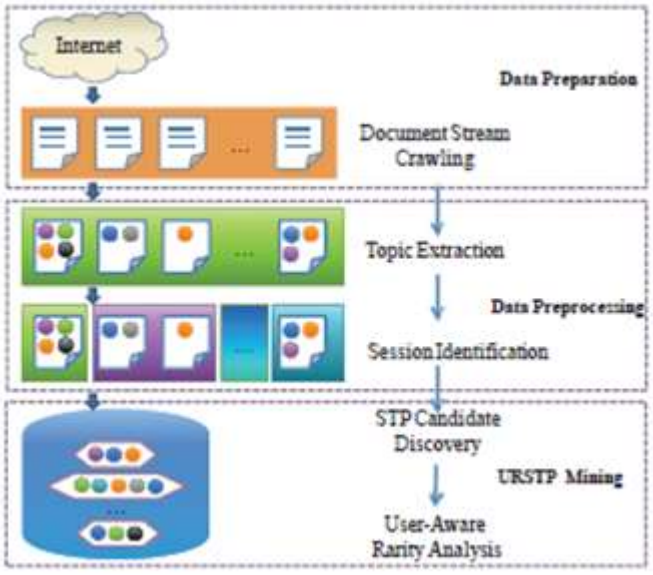

Fig: System Architecture

4. Objectives :

- Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

- It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

- When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

5. Scope Of Proposed System :

- In order to characterize user behaviors in published document streams, we study on the correlations among topics extracted from these documents, especially the sequential relations, and specify them as Sequential Topic Patterns (STPs).
- To solve the innovative and significant problem of mining URSTPs in document streams, many new technical challenges are raised and will be tackled in this paper.
- Firstly, 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.
- Secondly, 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.
- Thirdly, 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, un-supervised mining algorithms for this kind of rare patterns need to be designed in a manner different from existing frequent pattern mining algorithms.

6. Software, Hardware &Test Data Requirements:

6.1 Hardware Requirement:

1.System: Pentium IV 2.4GHz

2.Hard disk: 40 GB

3.Floopy Drive:1.44mb

4.Monitor:15 VGA Color

5.Ram:512 MB

6.2 Software Requirements:

1.Operating system: Windows XP/7

2.Coding language: JAVA/J2EE

3.IDE:Netbeans 7.4

4.Database:MYSQL

7. Conclusion:

Mining URSTPs in published document streams on the Internet is a significant and challenging problem. It formulates a new kind of complex event patterns based on document topics, and has wide potential application scenarios, such as real time monitoring on abnormal behaviors of Internet users. In this paper, 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 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. 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. Also, we will refine the measures of user-aware rarity to accommodate different requirements, improve the mining algorithms mainly on the degree of parallelism, and study on-the-fly algorithms aiming at real time document streams. Moreover, based on STPs, we will try to define more complex event patterns, such as imposing timing constraints on sequential topics, and design corresponding efficient mining algorithms. We are also interested in the dual problem, i.e., discovering STPs occurring frequently on the whole, but relatively rare for specific users. What's more, we will develop some practical tools for real life tasks of user behavior analysis on the Internet.

7. References:

[1] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, Frequent pattern mining with uncertain data, in Proc. ACM SIGKDD09, 2009, pp. 2938.
[2] R. Agrawal and R. Srikant, Mining sequential patterns, in Proc. IEEE ICDE95,1995, pp. 314.
[3] J. Allan, R. Papka, and V. Lavrenko, On-line new event detection and tracking, in Proc. ACM SIGIR98, 1998, pp. 3745.
[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuee, Probabilistic frequent item set mining in uncertain databases, in Proc. ACM SIGKDD09, 2009, pp.119128.
[5] D. Blei and J. La_erty, Correlated topic models, Adv. Neural Inf. Process. Syst.,vol. 18, pp. 147154, 2006.
[6] D. M. Blei and J. D. Lafferty, Dynamic topic models, in Proc. ACM ICML06, 2006,pp. 113120.
[7] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res.,vol. 3, pp. 9931022, 2003.
[8] 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 VAST12, 2012, pp. 143152.
[9] K. Chen, L. Luesukprasert, and S. T. Chou, Hot topic extraction based on timelineanalysis and multidimensional sentence modeling, IEEE Trans. Knowl. Data Eng.,vol. 19, no. 8, pp. 10161025, 2007.
[10] C. K. Chui and B. Kao, A decrement approach for mining frequent item sets fromuncertain data, in Proc. PAKDD08, 2008, pp. 6475.
[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, LeadLine: Interactivevisual analysis of text data through event identification and exploration, in Proc.IEEE VAST12, 2012, pp. 93102.
[12] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, Parameter free bursty events detection in text streams, in Proc. VLDB05, 2005, pp. 181192.
[13] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, FreeSpan: frequent pattern-projected sequential pattern mining, in Proc. ACM SIGKDD00, 2000, pp.355359.
[14] N. Hariri, B. Mobasher, and R. Burke, Context-aware music recommendation based on latent topic sequential patterns, in Proc. ACM RecSys12, 2012, pp. 131138.
[15] T. Hofmann, Probabilistic latent semantic indexing, in Proc. ACM SIGIR99, 1999,pp. 5057.
[16] L. Hong and B. D. Davison, Empirical study of topic modeling in Twitter, in Proc.ACM SOMA10, 2010, pp. 8088.
[17] Z. Hu, H.Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, Discovery of rare sequentialtopic patterns in document stream, in Proc. SIAM SDM14, 2014, pp. 533541.
[18] A. Krause, J. Leskovec, and C. Guestrin, Data association for topic intensity tracking, in Proc. ACM ICML06, 2006, pp. 497504.