International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com
Reference ID: IJCS-411
Volume 10, Issue 1, No 6, 2022
ISSN: 2348-6600
PAGE NO: 2791-2794

MINING USER AWARE RARE SEQUENTIAL TOPIC PATTERNS IN DOCUMENT STREAM

Vignesh P[#1], Dr.T.Velumani[*2]

[#]*Student, M.Sc Computer Science,*
Rathinam College of Arts and Science, Coimbatore,
Tamil Nadu, India -641021. vitvicky1234@gmail.com

**Assistant Professor, Department of Computer Science,*
Rathinam College of Arts and Science, Coimbatore,
Tamil Nadu, India -641021. velumani.cs@rathinam.in

Abstract - Among various multi keyword and sequential topic semantics, we choose the efficient similarity measure of "coordinate matching", i.e., as many matches as possible, to capture the relevance of data documents to the search query. 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. We implement 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. This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pinpoint where the topic emergence is; the proposed approach can detect changes in the communication patterns of users even in a realistic setting when only some part of the users reacts to the emerging topic.

Keywords - Sequential Topic Pattern, probability model, Secret Key, Search Data.

I. INTRODUCTION

Modules:

Admin

Admin Login:

In this module, the admin login to his page and view the word details, upload details, and user details. Through this module, the admin can view and insert the needed details.

Word Details:

Root Word:

In this module, the admin inserts the meaning for the word. With the help of this word, the user can view the searching content.

Related Word:

In this module, the admin can modify the details of the root words. Through this related word, the user can view the no. of words for the root word.

Upload Details:

Upload File details:

In this module, the admin can upload the file details related to the word details. The user can view the file details.

Upload User Details:

In this module, the admin can view the user details and their secret key request detail. They can send response to the authorized user to download the file.

II. SYSTEM ANALYSIS

EXISTING SYSTEM

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.

Disadvantages:

1. Design corresponding not algorithms to support it.
2. To topic modeling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored.
3. Do not provide a trade-off between not accuracy and efficiency.

PROPOSED SYSTEM

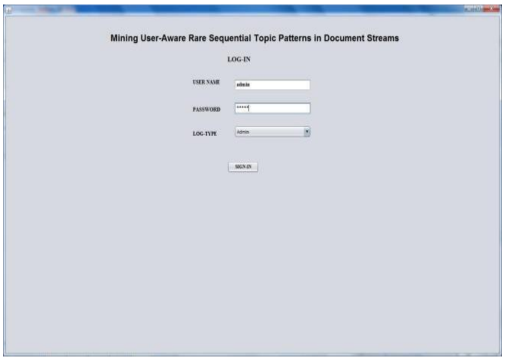
We propose a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Then this model is used to measure the anomaly of future user behavior. Using the proposed probability 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.

Advantages:

1. A framework to pragmatically solve this problem, and design corresponding algorithms to support it.
2. We give preprocessing procedures with heuristic methods for topic extraction and session identification.
3. 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.
4. That provides a trade-off between accuracy and efficiency.

III. OUTPUT SCREENS

Login



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

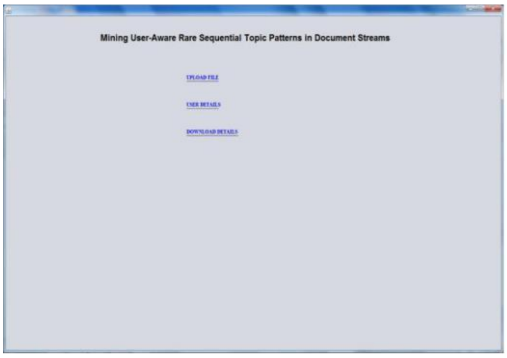
LOG-IN

USER NAME admin

PASSWORD *****

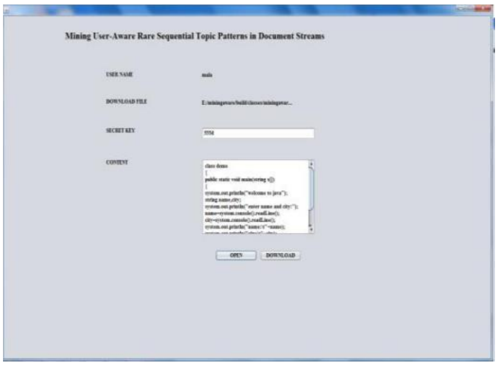
LOG-TYPE Admin

Admin Page



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

User Page



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

USER NAME mala

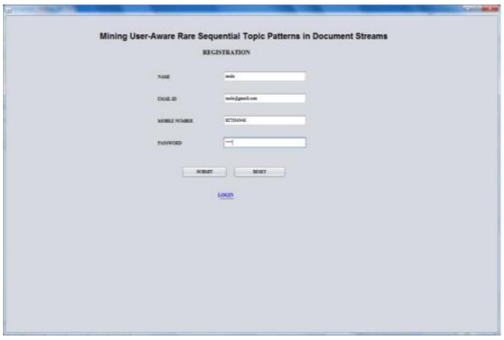
DOWNLOAD FILE E:\miningaware\build\classes\miningawar...

SECRET KEY 5554

CONTENT

```
class demo
{
public static void main(string s[])
{
system.out.println("welcome to java");
string name,city;
system.out.println("enter name and city:");
name=system.console().readLine();
city=system.console().readLine();
system.out.println("name:\t"+name);
```

Registration



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

REGISTRATION

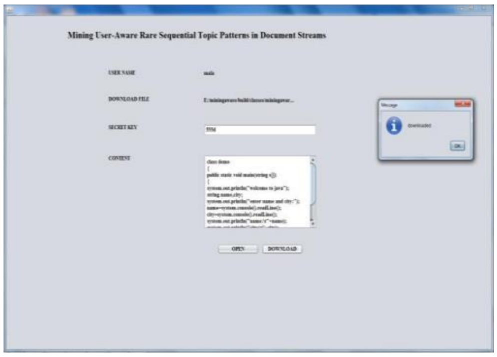
NAME: mala

EMAIL-ID: mala@gmail.com

MOBILE NUMBER: [illegible]

PASSWORD: ***

Download a File



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

USER NAME: mala

DOWNLOAD FILE: E:\miningaware\build\classes\miningawar...

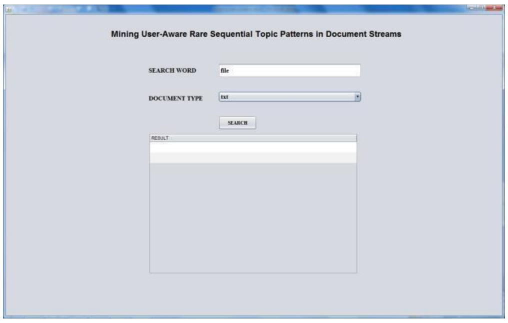
SECRET KEY: 5504

CONTENT:

```
class demo
{
public static void main(string s[])
{
system.out.println("welcome to java");
string name,city;
system.out.println("enter name and city:");
name=system.console().readLine();
city=system.console().readLine();
system.out.println("name:\t"+name);
```

Message: downloaded

File Search



Mining User-Aware Rare Sequential Topic Patterns in Document Streams

SEARCH WORD: file

DOCUMENT TYPE: txt

IV. CONCLUSION

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 reallife tasks of user behavior analysis on the Internet.

REFERENCES

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc. IEEE ICDE'95, 1995, pp. 3–14.

[2] J. Allan, R. Papka, and V. Lavrenko, "On line new event detection and tracking," in Proc. ACM SIGIR'98, 1998, pp. 37–45. [4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD'09, 2009, pp. 119–128.

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

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc.

[5] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," J. 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 VAST'12, 2012, pp. 143–152. [9] 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.

[7] C. K. Chui and B. Kao, “A decremental approach for mining frequent itemsets from uncertain data,” in Proc. PAKDD’08, 2008, pp. 64–75.

[8] 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.

[9] 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.