IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

USER BEHAVIOUR ANALYSIS USING SEQUENCE OF DOCUMENT ON INTERNET OF STREAM

Mr.P.Vijayaragavan Associate.Prof, S.Swetha student, M.Selva Udhaya student.
BE-CSE,DHANALAKSHMI COLLEGE OF ENGINEERING,MANIMANGALAM,CHENNAI,TAMIL NADU

DOI: 10.5281/zenodo.496088

ABSTRACT

Textual documents designed and divided on the Internet are ever changing in various forms. The aim of this project is to characterize and detect personalized and abnormal behaviours of Internet users.It can be applied in many real-life scenarios, such as real-time monitoring on abnormal user behaviours. The existing system of our project works are devoted to topic modelling and the evolution of individual topics, while sequential relations of topics in successive documents published by a specific user are ignored. Hence the users activity monitoring doesn't feasibly and effectively. We proposed our system to extract the user's activity on real time web application data set on Twitter and Gmail. Using our technique can monitor the user's sequential topic pattern based on their session identification on multiple applications with single sign on email id and their session id

KEYWORDS: sequential patterns,Web mining,rare events.

INTRODUCTION

Documents created and distributed on the Internet are ever changing in variety of forms. In this paper, in order to characterize and monitoring 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. We deliver 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 create special users and interpretable URSTPs effectively and efficiently, which significantly reflect users' characteristics

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). Each of them records the complete and repeated behaviour of a user when she is publishing a series of documents, and are suitable for inferring users' intrinsic characteristics and psychological statuses. First, difference 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. Association order analysis is one of the most important fields in data mining. It is commonly applied to market-basket databases for analysis of consumer purchasing behaviour. Such databases consist of a set of transactions, The most important and computationally intensive step in the mining process is the extraction of frequent itemsets sets of items that occur in at least minSup transactions.It is generally assumed that the items occurring in a transaction are known for certain. However, this is not always the case. For instance.In many applications the data is inherently noisy, such as data collected by sensors or in satellite images.In privacy protection applications, articial noise can be added deliberately [4]. Finding patterns despite this noise is a challenging problem.

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. Hence the users activity monitoring doesn't feasibly and effectively. And due to the static content monitoring makes the false alert on the sequential and individual topic extraction. Monitoring individual users' activity in single web application doesn't give the effective dataset of topic extraction about the user. So the users' intention and interest are extracted with ambiguous and suspicious manner due to uncertain data set. Hence the user's activity management cannot provide the effective guidance and feasible detection. Existing techniques of sequential pattern mining for probabilistic databases. So the content identification is huge to handle.

PROPOSED SYSTEM

In our proposed system, Users rare and sequential activities can be monitored using sequence of document streams on multiple web application. We proposed our system to extract the user's activity on real time web application data set on Twitter and Gmail. Using our technique can monitor the user's sequential topic pattern based on their session identification on multiple applications with single sign on email id and their session id. We used the documents of inbox and send box mail of Gmail contents and twitter's tweet and individual chats to extract the topic and mining the user's activity. We extract the topic of document stream content using Stanford Natural Language Processing. Using this NLP processing and Monitoring dynamic user's different activities can be extracted and monitored effectively.

It is worth noting that the ideas above 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.

MODULES

User's registration and creating dataset for user rare topics:

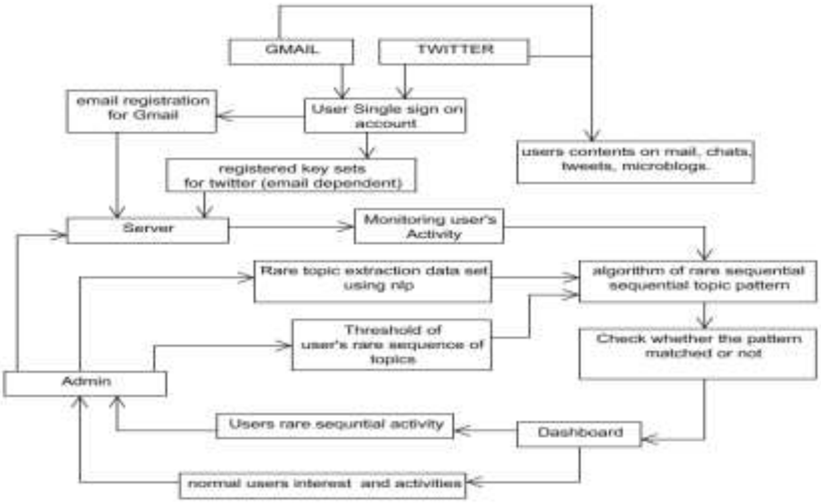
In this module the users have to register their email id and twitter key with our application. The email id and regarded twitter key's id should be a single sign on Gmail and Twitter account. Our application using users details make threshold for every users account by admin process. The data set of user's sequential topic extraction has to provide to the application. We build Stanford NLP algorithm to mining the user's activity. The data has been maintained and customized in the server. The user's details are stored in server database in the encrypted format because of the security purpose. To implement the effective rare topic extraction on sequential of document stream of the user's activity we used deserved data set of data mining process using Stanford NLP. In this API we implements pos tagging, chunking processing, stemming, spell checking and word net connection. We can feasibly extract the content of the user's rare topics using above mentioned NLP processing.

NLP processing on Gmail and twitter content:

The user's details can be extracted and monitored from the Gmail and Twitter to our local server database. Because of the huge amount of data set we create threshold based data retrieving from the Social Medias content. Before proceeding to the content retrieving has been make sure of single sign on id for Twitter and Gmail. Using twitters key and email id the mail content and twitter content can be extracted using Java Mail API and Twitter4j API. The type of data set can be categorized like inbox, sent items, mail chats, user's tweets, twitter chats and micro blogs maintained in our local server database. These social media contents are mined and extracted using Stanford NLP processing. The extracted topics of the user's contents are monitored in the server. Pos tagging create the parts of speech of the each content of the user's data set. Stemming process grouping the similar types of words of the content like calling, call, called and callable, etc. Chunking process removes the common words filtering on the content like is, was, the, of, etc.

Monitoring user's activity using Gmail and Twitter dataset:

The Server monitors every user's activity on Gmail and Twitter. Single user activity on the two different web applications can be identified and extracted using single sign on email ids. The sequential topic extraction on sequence of documents are extracted and grouped. The evolution of individual topics, while sequential relations of topics in successive documents published by a specific user can be grasped by our application. 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.


GMAIL
TWITTER
email registration for Gmail
User Single sign on account
users contents on mail, chats, tweets, microblogs
registered key sets for twitter (email dependent)
Server
Monitoring user's Activity
Rare topic extraction data set using nlp
algorithm of rare sequential sequential topic pattern
Threshold of user's rare sequence of topics
Check whether the pattern matched or not
Admin
Users rare sequntial activity
Dashboard
normal users interest and activities

ARCHITECTURE DIAGRAM

Mining rare user sequential activities :

While monitoring and extraction of the users sequential topics , if illegal behaviors are involved, detecting and monitoring them is particularly significant for social security surveillance. We can still expose them by URSTPs, as long as they satisfy the properties of both global rareness and local frequentness. That can be regarded as important clues for suspicion and will trigger targeted investigations. Therefore, mining URSTPs is a good means for real-time user behavior monitoring on the Internet. The ideas above 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. We implement the aware recommendation on admin dashboard. We highlight the rare user's activity and normal user's interest based on their social network**.**

CONCLUSION

Mining URSTPs in published document is streams on the Internet is a significant and challenging problem. It calculates a new kind of complex event patterns based on document topics, and has wide potential application scenarios, such as real-time monitoring on abnormal behaviours 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 solve the 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 behaviours and characteristics.

As this paper puts forward an innovative research direction on web data mining.we propose two new algorithm one is Key phrases and Aspect extraction and another one is Rare Pattern domain analysis.

REFERENCES

1. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in Proc. ACM SIGKDD, 2009, pp. 29–38.
2. R. Agrawal and R. Srikant, "Mining sequential patterns," in Proc. IEEE Int. Conf. Data Eng., 1995, pp. 314.
3. 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.
4. 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.
5. D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.
6. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 113–120.
7. D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 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 Conf. Vis. Anal. Sci. Technol., 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, Aug. 2007.
10. C. K. Chui and B. Kao, "A decremental approach for mining frequent itemsets from uncertain data," in Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2008, pp. 64–75.
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 Conf. Vis. Anal. Sci.Technol., 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. 31st Int. Conf. Very Large Data Bases, 2005, pp. 181–192.
13. J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, "FreeSpan: frequent pattern-projected sequential pattern mining," in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2000, pp. 355–359.
14. N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latent topic sequential patterns," in Proc. 6th ACM Conf. Recommender Syst., 2012, pp. 131–138.
15. T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999,pp. 50–57.
16. L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. 1st Workshop Soc. Media Anal., 2010, pp. 80–88.
17. Z. Hu, H. Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, "Discovery of rare sequential topic patterns in document stream," in Proc.SIAM Int. Conf. Data Mining, 2014, pp. 533–541.

18. A. Krause, J. Leskovec, and C. Guestrin, “Data association for topic intensity tracking,” in Proc. ACM Int. Conf. Mach. Learn.,2006, pp. 497–504.
19. W. Li and A. McCallum, “Pachinko allocation: DAG-structured mixture models of topic correlations,” in Proc. ACM Int. Conf. Mach. Learn., 2006, vol. 148, pp. 577–584.
20. 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 11th Int. Conf. Data Mining, 2013, pp. 448–457.