Biomedical Research 2017; Special Issue: *ISSN 0970-938X*
www.biomedres.info

Data mining based on integrated network behaviour analysis of different biological groups.

Hejun Zhu, Liehuang Zhu*

School of Computer Science, Beijing Institute of Technology, Beijing, PR China

Abstract

With the fast development of big data and the encryption of network behavior was increasing, it was not enough to use single network behavior for big data analysis and mining in encrypted network environment. In this paper, a large number of network behaviors were analyzed in the process of handling cases, and the features of network behavior were extracted in order to identify different kinds of biological groups, and then the information mining scoring model which based on the integrated network behavior of big data was proposed, and the knowledge of frontline intelligence analysts were processed into expert model, then the key suspects were automatically and exactly found through the expert knowledge and scoring model, and it can also be used in the biological areas to find special objects. Experiments and practical project applications showed that this method could not only solve the difficult in the practical work, but also improve the efficiency of the work.

Keywords: Biological groups, Sub-standard drugs, Data mining, Network behavior, Expert model.

Accepted on August 28, 2017

Introduction

The technology of big data and data mining is the power of development of information technology, there is a lot of valuable information hidden behind the big data in the social, economic, biological and other fields. With recent advances in bioinformatics and genomics, handling and processing of biological data has become more complicated. Gene sequencing technologies and other experimental innovations has demanded the "big data revolution" among biologists [1]. Researchers desperately need to translate the biological data existing in various databases and information libraries to solve important queries in science. Biological data mining has been considered as a broad application prospective in the challenging research area of biotechnology. Now, the network space is easy to be attacked from outside, it will take a big threaten to property and personal safety, therefore, It is the key research object of large data mining [2].

There are various biological data encryption models used to hide the data from un-authorized users. DNA encryption model is used to protect sensitive data based on bio-molecular properties. This utilizes DNA computing in securing the data [3]. With the rapid development of big data [2] and the improving of network behavior encryption [4], it is hard to get useful information only through a single network behavior. So, we need to use integrated network behavior analysis which is based on the big data. As we know, the network behaviors reflect the work nature, hobby and interest etc., there are also many generalities for Internet users, it considers as a group behavior on the network [5]. In this paper, a large number of network behaviors are analyzed in the process of handling cases, and the features of network behavior are extracted in order to identify different kinds biological groups, and then the information mining scoring model which based on the integrated network behavior of big data is proposed, and this method can also be used in various biological areas to find special objects.

Methodology

According to different behaviors and five factors that include person, location, object, organization and event, network behaviors can be measured and quantified the element description of network behaviors can be realized. Then, the unified behavior is identified by means of scenes, and the representativeness of the topic and the relevance of the clues are analyzed. Starting from the typical network behavior, we explore the internal relations of different clues, and then quantify the typical network behavior.

Topic models [5] include two parts:

(1) Accumulation and use of experience, the establishment of expert knowledge base.

The expert knowledge library records all kinds of features, rules and ideas that generated in the process of topic mining, and then their mutual contacts were established. The knowledge library is made up of base library, process library and policy library. The base library includes topic organization library, topic population database, topic location library, topic keywords library and topic Internet behavior library. The

process library includes topic time series description library and topic Internet behavior space sequence description library. The policy library includes all kinds of special features, such as online time library, online frequency library, network behavior and expert base database.

(2) Scene analysis of network behavior, identification of representative behavior, and establishment of topic index system and establishment of topic description model.

The behavior elements in the knowledge library are extracted and formed the scene, the network behavior and organization, personnel, online activity sites, contents, keywords and other elements are organized according to the scene, and then compared with the knowledge in the knowledge base. The model is as shown in Figure 1.

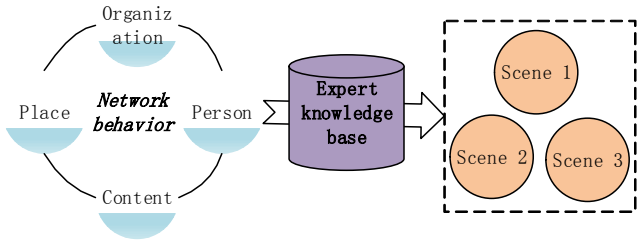

Figure 1. *Score and mining model.*

As we seen in Figure 1, for the special service of network, we can classify them by people, things, cases and organizations, and give the different weights for each of them. According to the experience of information analysis experts, the behavior of each kind of object activity is summarized. According to different behavior characteristics, the corresponding model is constructed to refine it into behavior analysis model. For example, substandard drugs agency often use hack tools, send PGP encrypted attachments, and send mail with Gmail, gamblers often login gamble websites and so on. The work of the intelligence analyst is to discover these representative actions and compare them synthetically, and then produce intelligence. In the actual work project, the labels should be assigned to different sensitive categories, and the weight value is set separately. This action that extracts specific behavior and sets values of different quantized weights is called the corresponding weight index according to the sensitive category. Each index corresponds to a sensitive class group of Internet users.

According to the scoring model, user behaviors and weight index, the formula is recorded as Equation (1).

$$r = \sum_{i=1}^{N} \lambda_i \cdot \eta_i \rightarrow (1)$$

Among Equation (1),

The r represents the evaluation results for the score model,

The N represents the total number of user behavior characteristics,

$\lambda_i (i=1,2,3,...N)$, it represents the specific characteristics of user behavior,

$\eta_i (i=1,2,3,...N)$ it represents the weight index of the corresponding user behavior specific characteristics.

We can get the score result by Equation (1), and according to the values of the result, the system will make a decision to give us the key suspect automatically.

Taking substandard drugs as an example, this paper discusses how to set up a classified special model of Internet users based on comprehensive network behavior analysis, which is based on the actual experience of handling cases, the behaviors of substandard drugs agency include using hack tools, downloading substandard drugs files, sending email with encrypted mail server, posting information on Webbbs, sharing information with IM application, talking online with Skype or Viber and so on. Of course, these users usually do some normal network behaviors including Internet news browsing, online games or online shopping. Then, how the substandard drugs agency can be automatically excavated from the big data that are encrypted and scattered?

According to extensive analysis and experimentation by intelligence experts, the internet users involved in substandard drugs agency usually include at least 7 kinds of network behavior, and the network tools and network protocols used by substandard drugs agency are analyzed and listed in Table 1.

Table 1. *Agency network behavior.*

Network tools	Network protocols
Use hack tools	Non-standard protocol
Browse or download file	Http/P2P
Login encrypted mail	Https
Send files by mail or BBS	Http/Https/SMTP/POP3
Use IM application	MSN/QQ
Use Skype or Viber	Non-standard protocol
Browse news	Http
Online games	Online game protocol
Online shopping	Http

The classification and identification model will be established according to these network behaviors of Table 1. Firstly, the network protocols involved in 7 kinds of network behavior are extracted, and the network protocols are more than 10 kinds including Httpget, Httppost, P2P, SMTP, POP3, IM, Https, Webmail, non-standard protocol, Webbbs and so on.

Secondly, there are 10 kinds of massive structured data in these network protocols, involving 7 kinds of big unstructured data. According to the knowledge library that we have created, the topic model of integrated network analysis is created, as shown in Table 2.

According to the classified model, the index system and the scoring model, it can be concluded that the model characteristics of the internet can be considered to be involved in substandard drugs agency and according to the definition of the comprehensive analysis of network behavior, a class of users based on the analysis of the classification model is related to substandard drugs class of Internet users. In this paper, substandard drugs topic is an example to create the classification model based on comprehensive analysis of network behavior, for other classes, such as online gambling, Internet fraud, drug category, and other network events that concerned by other departments, can also be built in accordance with the classification model in the similar method.

Table 2. *Agency topic model.*

Network tools	Characteristic data
Httpget/Webbbs/Httppost	URL/IP library, keywords library
SMTP/POP3/Webmail	Mail address library, Encrypted mailbox and keywords library
IM	Account and keywords library
P2P	Torrent name and MD5 value library, URL/IP address library
Non-standard protocol	Hack tools type and encrypted IP phone type library

Results

After the model is refined and different indexes are set up, the system automatically quantifies each object according to all defined tags, and finds the object that meets the requirements. And with the rich experience of experts, the system can change the exponential model at any time and adapt itself to change.

Table 3. *The comprehensive intelligence index of substandard drugs suspect within one week.*

Behavior	Frequency	Weight	Value	Index
IM	10	3	30	
Mail	3	5	15	
P2P	2	15	30	
BBS	3	15	45	726
Skype	15	10	150	
Gmail	20	10	200	
Encrypted data	32	8	256	

When displaying the results of the key suspects, each score is indicated by different color warning signals, and red, orange and yellow warning strategies are defined. In the excavation of substandard drugs, according to expert experience, the comprehensive index is defined as follows: 108-400 are yellow alert, 401-660 are orange alert, and 661 and above are red alert. As shown in Table 3, a comprehensive intelligence index of a substandard drugs suspect in a week, who has exceeded the threshold of the red alert, has been recommended as a key suspect.

When it is red alert in this model, the suspect is automatically recommended as the key suspects. When orange and yellow warning, the suspects need to manually be determined whether the key suspects.

Conclusion

This paper studies the information mining based on integrated network behavior analysis, mainly based on expert knowledge topic mining and substandard drugs topic mining. After a lot of analyses and experiments, the features of network behavior are extracted in order to identify different kinds of groups, and the comprehensive network behavior mining and expert scoring model is put forward, and the knowledge of frontline intelligence analysts are processed into expert model, then the key suspects are automatically and exactly found through the expert knowledge and scoring model, and this method can also be used in the biological areas to find special objects. Experiments and practical project applications show that this method can not only solve the difficult in the practical work, but also improve the efficiency of the work.

Health care informatics is rapidly growing as a new arena in personalized medicine and public health. The big data application tools are one of the powerful co-factors that help the data processing and retrieval easy [7]. The technology is being considered as a potential savior in public health management system.

References

1. Tetko IV, Engkvist O, Koch U, Reymond J, Chen H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. Mol Informat 2016.
2. Nguyen TT, Nguyen MH. Zing database: high-performance key-value store for large-scale storage service. Vietnam J Computer Sci 2015; 2: 13-23.
3. Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big data bioinformatics. J Cell Physiol 2014; 229: 1896-900.
4. Saabith ALS, Sundararajan E, Bakar AA. Parallel implementation of apriori algorithms on the hadoop-mapreduce platform - an evaluation of literature. J Theor Appl Info Technology 2016; 8: 321-351.
5. Neumeyer L, Robbins B, Nair A, Kesari A. S4: Distributed Stream Computing Platform. IEEE International Conf. on Data Mining Workshops 2010.
6. Sherry J, Lan C, Popa RA, Ratnasamy S. Blindbox:deep packet inspection over encrypted traffic. Acm Sigcomm Comput Commun Rev 2015; 45: 213-226.
7. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Informat Insights 2016; 8: 1-10.

***Correspondence to**

Liehuang Zhu

School of Computer Science
Beijing Institute of Technology
PR China