A New Network Forensics System for Chinese Text Content

Cui Yimin, Jin Qi, Zhang Lufeng, Zou Tao
National Key Laboratory of Science and Technology on Information System Security
Anxiang Beili 10#, Chaoyang District, Beijing, China
tsui-min@163.com

Abstract: - This paper presents an implementation technical solution of network forensics system for Chinese text content. The technical solution utilizes Bloom filter algorithm and Chinese word segmentation and meta-aggregation algorithm(CWSMA) to preprocess and effectively store contents of the text aiming at technical challenges caused by characteristics of "unpredictability of the event features" and "unpredictability of forensics operation", information related with the events such as 'where', 'who', 'when' and the like can be provided for investigators through member query, network verification analysis can be carried out under the condition without predefining event characteristics, the forensics analysis time traceability can be prolonged from several days of existing technique to several months, it is particularly suitable for network forensics of network secret disclosure events and illegal content propagation events with sensitive content analysis.

Key-Words: - Network forensics; Bloom filter; Chinese word segmentation

1 Introduction

Network secret disclosure and illegal content propagation events cause serious threats to national security and social stability. Chinese text is the most important information carrier of current network secret disclosure events and illegal content propagation events in China. The paper refers confidential content and illegal content as sensitive contents uniformly. The network secret disclosure and illegal content diffusion events occurring in network are traced for forensics, and it is not only a technical means imperative for network regulatory authorities and the judiciary, but also a technical difficulty which is unsolved in the network safety technical field currently. The forensics of network transmission contents is difficult mainly due to two major aspects of unpredictability in its application.

- Unpredictability of event characteristics. Namely: forensic system and the regulators cannot predict the concrete contents transmitted by network secret disclosure events and illegal content propagation events. Moreover, the illegal events have no prominent difference on behavior operation except the difference from other normal network users in transmitted contents, therefore the network secrete disclosure and illegal content propagation events cannot be monitored through setting one character or one characteristic in advance.
- Unpredictability of forensic operations. Forensic operation behavior is generally started after the result of one illegal act is discovered (for example, somebody discovers that some secret disclosure information appears on the network). However the time interval between the event occurrence and event discovery cannot be predicted, which may be several days as short as possible, and several months as long as possible [1]. Thereby it is required that the system should have strong event analysis traceability.

Intrusion detection [2], network content monitoring [3, 4] and existing network forensics technology [5, 6] cannot effectively support the evidence collection of network sensitive content propagation events because of the unpredictability of its application in network forensics. In addition, network content forensic application also has special "knowledge apriority" at the same time, namely, when the investigators start network forensic analysis operation, a part or whole sensitive contents have been mastered generally through event discovery and report. Questions on three aspects should be further answered through network forensics, namely: "WHERE: the network in which sensitive contents are transmitted", "WHO: the distributor and the receiver?", "WHEN: the time of propagation?", thereby providing evidence for finding out the person in charge, and helping the investigator to confirm the suspicion host machine at the same time, and the host machine can be further started for forensics.

2 System Structure and Work Flow

ISBN: 978-1-61804-051-0

2.1 System Composition Block Diagram

Through preceding analysis, it can be known that the unpredictability of network content forensics decide that forensic analysis ability necessary for network forensics can be provided only through realizing "global capture and long-term storage" (i.e.: all text contents transmitted in the network should be captured, and should be stored for long time). However, the "knowledge apriority" only enables the system to provide answers of "WHERE", "WHO" and "WHEN" under the condition of providing 'member query' rather than 'original text browse', thereby greatly simplifying system design.

Network Forensics System for Chinese Text Content (NFSCTC) utilizes Chinese Word Segmentation and Meta-Aggregations algorithms(CWSMA) and Bloom filtering techniques [7] to realize long term storage of network transmission text content information without omission, and answers of "WHERE", "WHO" and "WHEN" can be provided for investigators through member query for many times.

NFSCTC system is composed of network data capture interface, network data preprocessing module, data storage module, and query and analysis module as shown in Fig. 1.

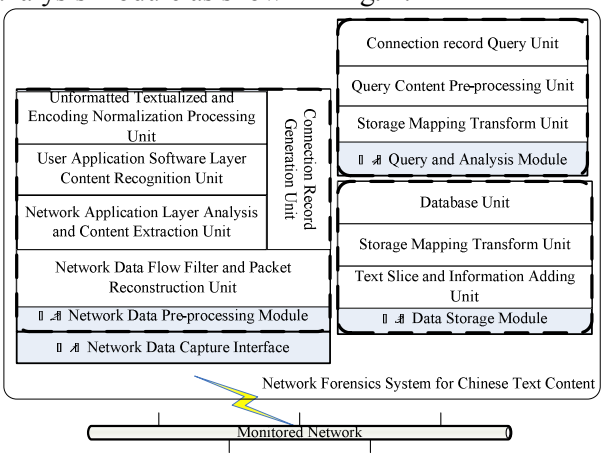

Fig. 1 System Composition Block Figure

2.1 Work Flow

System work flow mainly comprises forensic analysis basic data accumulation phase (hereinafter referred to as the accumulation phase) and the forensic analysis query phase. In the accumulation phase, the network data capture interface firstly captures all data flows passing through the network from the monitored network, then, the connection record with the shape like <source IP, target IP> is extracted from the network data preprocessing module, meanwhile, network data are analyzed layer by layer to extract the text content of the user layer (namely: contents which are really transmitted by the users, such as word document), and are processed into unformatted text with uniform encoding mode. The complete text paragraphs which are output in the condition (such as complete unformatted text content extracted from one word document) can be cut into text slices with smaller length after CWSMA algorithm processing, and are added with affiliated connection record information for bloom filter mapping transform together. The transform results are stored in the database unit.

Query and analysis module works in the forensic analysis query phase. One typical query forensic analysis process is shown in Fig. 2. For example: we assume that the text contents related with one network secrete disclosure event is discovered to be "This document is not allowed to be transmitted on network", and then the investigation and forensic process of the secret disclosure event can be roughly divided into the following steps.

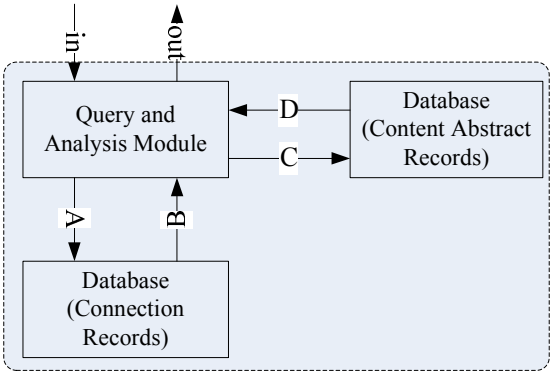

Fig. 2 Typical Query Analysis Process

- The investigators largely determine the possible time slices $t_1 \sim t_2$ of the secret disclosure event according to prior knowledge obtained from event investigation, and the query information is input into the forensic system;
- The connection record query unit in the query analysis module sends out query A 'which connections are generated among time slices $t_1 \sim t_2$?' to connection record storage unit of record connection information in the forensic system;
- The connection record query unit returns the query result B: All sources, target IP to set IPCs appearing among $t_1 \sim t_2$;
- Slicing, connection information addition and mapping transform are carried out by the query content pre-processing unit on forensic

contents “This document is not allowed to be transmitted on network” according to the data accumulation phase mode, and member query is carried out in the database-query C;

- After member query is carried out on all IPs in *IPCs,* the system returns the query result D.

If all member queries are not hit, users can adjust the time slices $t_1 \sim t_2$ to continue query; or explain that the secret disclosure event does not occur in the network. If all text slice queries of one pair of IPs are hit, it can be explained that ‘ the network has transmitted the content “This document is not allowed to transmitted on network”’ and ‘the content is transmitted by the source IP in the IP pair to the target IP’, and ‘the transmission time is about ×××’ (namely the storage time of the hit record in the database), thereby leading the investigator to start host machine forensics on the source IP and target host machine.

3 Key Algorithm Analysis

The key technique point of NFSCTC to realize “global capture and long-term storage” lies in that the adopted data storage mode is not direct storage of text content, but improvement of system space storage efficiency through bloom filter mapping transform. The essence of bloom filtering is to utilize the abstract to express original information through Hash transform. Bloom filter algorithm can be described as follows [8]: the bit vector V with the length m is used for expressing element set $A = \{a_1, a_2, \cdots, a_n\}$, k Hash functions h_i with uniform distribution features are set, and $\forall x \in A, h_i(x) \in \{1, 2, \cdots, m\}$, namely:

```
BitString BF_reprent(Set A, int m) {

// bit vector V with the length m is used for expressing data set A.
    BitString V[1 ··· m]=0;

    for (j=1; j<n+1; j++)
```

Fig. 3 Set representation algorithm.

As for NFSCTC, set element is text slice obtained after Chinese word segmentation. We assume that the average length of text slices is g characters, a two-byte Chinese character encoding method is adopted, the storage space of $g \times n \times 16$ bits is needed for directly storing set containing n text slices, and only the storage space of m bits is needed through adopting bloom filter. $m << g \times n \times 16$ in practical condition. The expense of bloom filter to obtain space efficiency is to introduce the false positive, and the false positive rate can be calculated according to the formula (1):

$$FP = (1 - (1 - \frac{1}{m})^{kn})^k \approx (1 - e^{-kn/m})^k \quad (1)$$

It can be deduced that the system storage space efficiency improvement multiple after the use of Bloom filter algorithm is: $\frac{g \times n \times 16}{m}$. When $k = \ln 2 \times (\frac{m}{n})$, FP obtains the minimum value $FP_{\min} \approx 0.6185^{m/n}$. The false positive rate, storage space compression rate and other system parameters of Bloom filter algorithm are concretely analyzed in the following section.

The member query method adopted by the bloom filter decides that the minimum analysis granularity provided by the NFSCTC system is granularity of set element in Bloom filter algorithm. Too coarse granularity will reduce the forensics analysis ability of the system.

The NFSCTC cuts longer text paragraphs into text slices with smaller granularity through designing a novel CWSMA algorithm, thereby guaranteeing that the system not only has stronger error adaptability of contents to be inquired, but also has forensic analysis ability of thinner granularity at the same time. CWSMA algorithm is introduced in details in the article [8].

4 Performance Analysis

Compared with other technologies, NFSCTC system has the greatest advantage that it can provide network forensics analysis support on events with sensitive contents long ago under the condition without pre-defining features. The change relation among the false positive rate FP, parameter k and m/n can be obtained according to the formula (1), as shown in fig. 4 and 5.

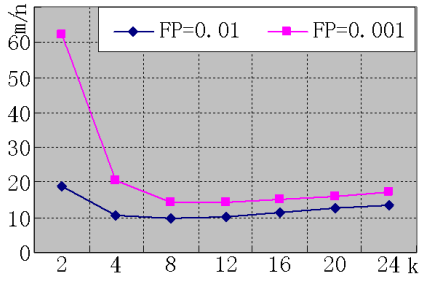

Fig. 4 Change curve of m/n on different FP s with k value

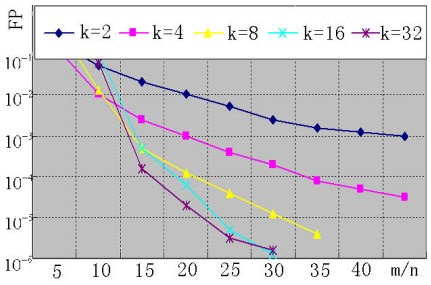

Fig. 5 Change figure of FP on different k values with m/n

Curve in Fig. 4 shows that when 8 Hash functions are used, and the system false positive rate is 0.01, m/n is about 10. It is assumed that the average length of the text slice is 8 characters, and the improvement multiple of storage space efficiency when n non-repeated text slices are stored is about 12.8 times. The condition that repeated text slices are always contained in actual situation is taken into account, and CWSMA algorithm can filter out useless Chinese auxiliary words such as 'de', 'le', 'ma' and the like for forensic analysis, the practical storage space efficiency of the system can be improved to about 15 times, namely, the forensic analysis tracing ability of the system can be improved by 15 times under the same storage condition. It is assumed that traditional technique can support forensics of events within one week, and the NFSCTC system can realize forensics of events within four months.

NFSCTC system forensic analysis traceability is improved with the expense of introducing false positive rate. But modest false positive rate does not affect the normal use of the system. It is assumed that the to-be-inquired contents of once network forensics analysis are sliced into r text slices by CWSMA algorithm, the final false positive rate of the query analysis is FP^{r}. As for the former example, when r is 5, the final false positive rate of query forensics is only $10^{-2\times5} = 10^{-10}$, which can be almost negligible.

Curve in fig. 5 shows that: the difference of FP curves is not great when k selects 8, 16, and 32 under the condition of given m/n, thereby smaller algorithm complexity O(k) can be obtained through selecting smaller k value.

5 Conclusion

This paper presents implementation technical solution of network forensics system for Chinese text content. The technical solution fully utilizes information advantage of 'knowledge apriority' in network forensics applications, and simplifies the design of network forensics system; Bloom filter algorithm and Chinese word segmentation and meta-aggregation algorithm are utilized to solve the contradictions between forensics analysis foundation data quantity and forensics analysis time scope aiming at technical challenges caused by characteristics of "unpredictability of the event features" and " unpredictability of evidence collection operation" at the same time. The network forensics system realized through adopting the technique can analyze and verify any text contents or text slices transmitted in the network under the condition that characteristic character series are not pre-defined, the forensics analysis time traceability of the system can be prolonged from several days to several months, the performance of network forensics is greatly improved, and it is particularly suitable for forensics analysis of network secret disclosure events and illegal content propagation events with sensitive content analysis.

References:

[1] K. Shanmugasundaram, N. Memon, A. Savant, and H. Bronnimann. Fornet: A distributed forensics system. The Second International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security, May2003.

[2] S. Axelsson. Research in Intrusion Detection Systems: A Survey. Technical Report. 1999.

[3] Zhu Huaxing, Dai Guanzhong and so on. Network Safety Research Based on Content Examination Filtration. *Computer Application* 2006 Vol.23 no.10 130-132.

[4] Wang Jigang, Gu Guochang and so on. Research and Design of High-speed Backbone Network Content Monitoring System. *Computer Research and Development* 2006 Z2 343-348.

[5] Xu Xiaoqin, Gong Jian, Zhou Peng. Analysis of Network Forensics System and Tools. *Microcomputer Development* 2005 Vol.15 no.5 139-141.

[6] Yang Zeming, Research and Realization of Computer and Network Forensics System [D] Institute of High Energy Physics, Chinese Academy of Sciences 2007.

[7] Bloom B. Space/time tradeoffs in hash coding with allowable errors. Communications of the ACM, 1970, 13(7): 422~426.

[8] Zhao Qian, Distributed Network Forensics System Based on Bloom Filter Engine. *Transaction of China Science and Technology University*. 2009.

ISBN: 978-1-61804-051-0