Research of File Index Model for Chinese Internet Fraud Information Management Platform

Hu Liang[1, a *], Ding AiChun[2, b], Zhu YuChi[1, c]

[1] Department of Humanities and Management, JiangXi Police College, NanChang City, JiangXi Province, P.R.China;

[2] Scientific Research Department, JiangXi Police College, NanChang City, JiangXi Province, P.R.China;

[a*]huliang_thu@163.com, [b]acting2312@163.com, [c]zhuyuchi_jxga@163.com

Keywords: File index model, internet fraud, information management platform

Abstract. According to the Chinese Internet fraud is becoming more and more complex, this paper studies to search engine technology as the foundation, through the computer collected public opinion data fraud, then use the information extracting and processing these data, the semi-structured web data into specific structured information, to establish to double byte inverted index technology based database files to index, construct a provide network bilk information retrieval of the vertical search engine public platform.

Introduction

Network fraud is a new type of crime, and its research is in the initial stage. With the development of network technology, the type of Internet fraud is increasing, the understanding of network fraud also exist different views, so the definition of network fraud did not form a unified understanding [1,2]. Such as some researchers think that Internet fraud is not strictly a legal concept, but criminology meaning of a category of crimes; studies that define the concept of network crime should first clear the research from the perspective of the localization problem, namely from the criminology angle or from the criminal law perspective to explore the concept of network crime. In the science of criminal law of network crime, or criminal jurisprudence on the network crime is a statutory crime, accord with the criminal law clearly stipulates the crime constitution and should bear criminal responsibility behavior. Although definitions of network fraud researchers have different opinions, but also a unified concept, namely Internet fraud is to the illegal possession for the purpose, the use of the Internet as an important tool for crime, cheat the behavior of a large amount of public or private property.

This paper studies using information technology construction of the database of network public opinion of fraud, through a unified database will be all kinds of fraud information gathered in the database, and to take relevant measures to prevent consumers become victims of similar fraud. On the one hand can provide clues for the judiciary to detection, on the other hand can be to the greatest extent so that users from being cheated. In this paper, the research object is network, there are a lot of fraudulent information, therefore should be solved firstly in this study is data from where and how data data acquisition and whether you need to finish machining, namely data sources, data collection and data filtering research.

Internet fraud information sources

The main source of fraud information is the media coverage of fraud news and online fraud complaints related information, followed by the public security system fraud database. It is estimated, about media reports of fraud news data volume of about 300 million, Internet Crime Complaint related information is more, according to the prediction model about 800 million in orders of magnitude, public security system, large-scale fraud database has more than 100, the total amount of data about 150 to 200 million between, but also have some data from some public fake website.

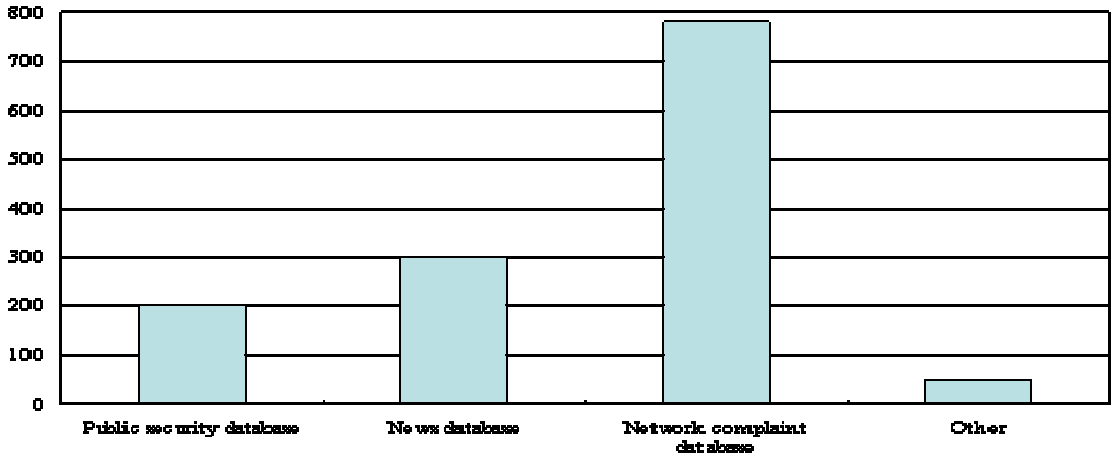

Fig.1 Internet fraud information sources

Inverted file index model of Internet fraud information

Because of Chinese network information fraud accounted for the proportion of more than 80%, and the proportion of English and other types of characters seldom and not more than 15% and Chinese character set commonly used Chinese characters number more than English letters, so both the encoded in different ways, English is single byte, and Chinese with two bytes to store, and double byte index performance than the single byte to high, so we use double byte inverted index technology, to file name per two bytes to establish inverted index table. The inverted file is divided into two parts: the first part is made up of words index, the second part is all the documentation of each word corresponds to the set, called the record file. Each data item index file is composed of pointer keyword and point to the record file, records each data record and a corresponding word document list. When all the documents (TJ, Di,...) According to the above two level structure to organize according to the word of the index, inverted file is built up, the word TJ corresponding Lists Posting is {(DI, fi, a*) + (di+k, fi+k, a*) +... }, FI is the number of TJ in di.

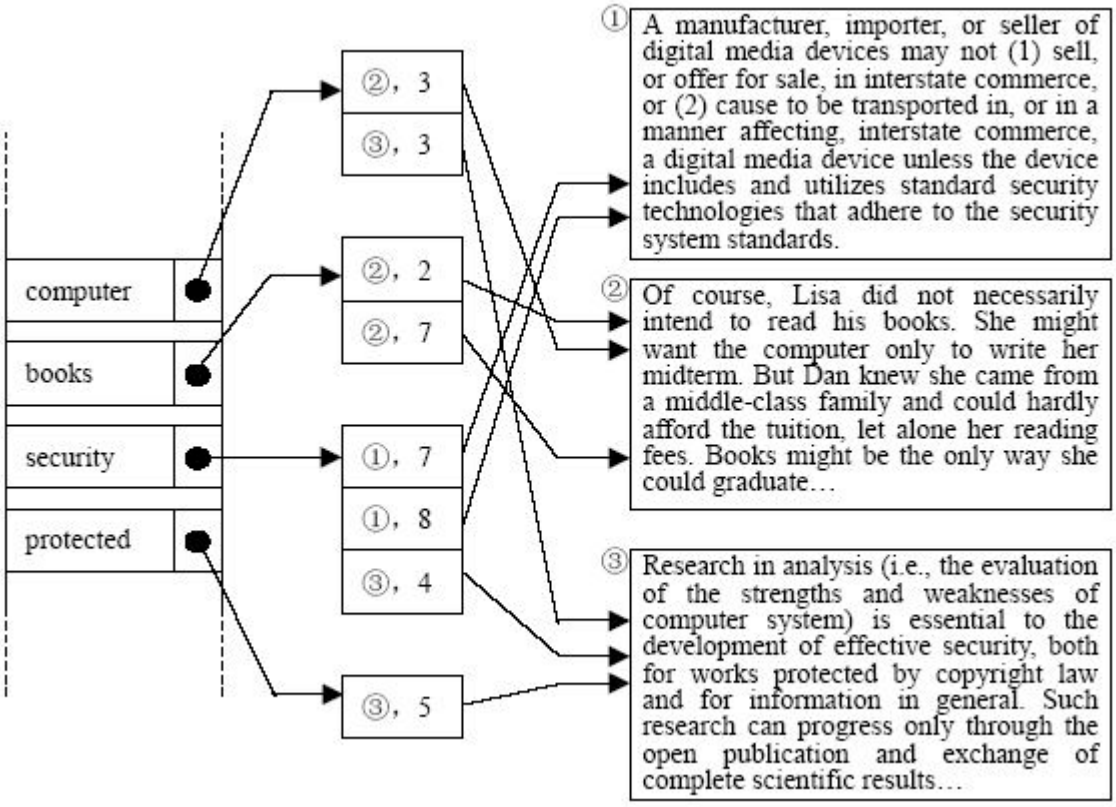

Fig.2 Inverted index

Performance testing and evaluation

In this study, we use the system throughput and average query response time to evaluate the performance of the system in two indexes. System throughput rate refers to the system per second of processing a number of queries, the reaction of the retrieval system's query processing capabilities; the average query response time refers to received from the query request to return the query results, the average time required, the reaction of the retrieval system efficiency.

In the evaluation experiment, the system is used to make the actual measurement of the retrieval system. During the experiment, on the other machines in the same local area network, a number of virtual client threads are set up to run the system. By adjusting the number of virtual client threads, the load of the retrieval system is changed, the performance of the retrieval system under different load is changed, and the influence of each parameter in the system is analyzed.

1) System throughput

In order to system analysis and system throughput rate of load, we compare under different load conditions the system throughput changes. The results as shown in Figure 3, which the X-axis said system load is ready process queue length, the Y-axis said per second the number of request processing.

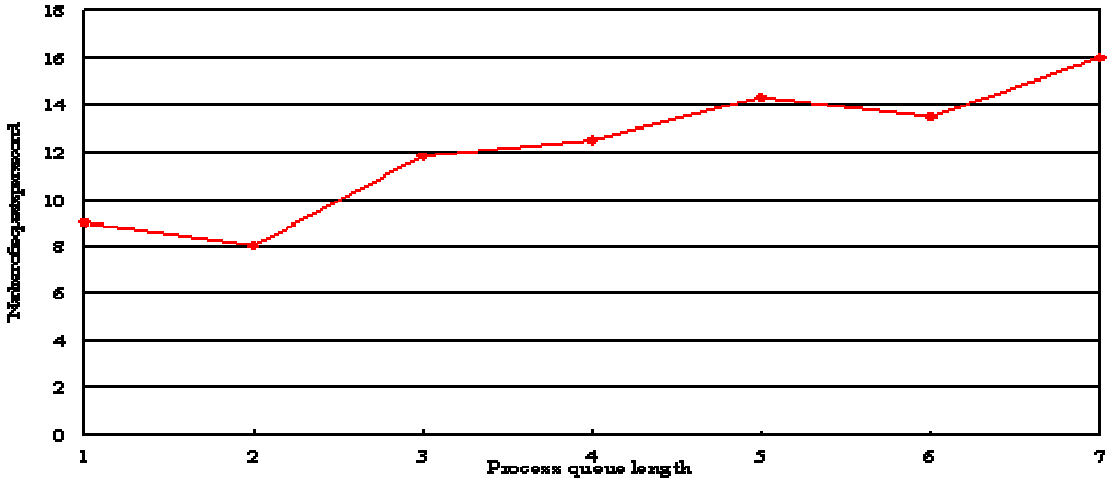

Fig.3 Comparison of system throughput under different load conditions

When the load is small, low throughput, this time failed to fully exert the ability of retrieval system. When the load increases slowly, and the calculation of CPU disk read and write overlap, the system throughput increased gradually, when the load is close to 6, the throughput rate reaches the maximum, which is the largest search capability of the system. When the simulation experiment continues to improve the number of concurrent virtual clients, some requests are discarded, and the number of threads is limited, and the load is not increased. At this time the host's CPU usage is less than 60%, we can see that the computing power of CPU has not been fully utilized, so that the main bottleneck of the single machine system is still disk access.

2) Average query response time

To analysis system load and the average query response time, we compare in different load conditions, the average query response time change. The results as shown in Figure 4, which the X-axis said the system load is ready process queue length, the Y-axis said system the average query response time.

Fig 4. Comparison of average query response time under different load conditions

Compared to the system throughput, the average query response time is better. When the load is 3-6, the average query response time is between 500 and 650 Ms.

Discussion and future work

Considering the limitations of the traditional anti fraud methods, this paper from a variety of heterogeneous data sources to collect information construct a cross agency network fraud database, for the social security administration to provide multi angle, multi-level query and analysis data of the functions and the network fraud warning decision support, also for practical applications, building a public platform for the Chinese public to provide Internet fraud public opinion data retrieval.

Acknowledgment

This author’s work is supported by JiangXi Research on teaching reform of higher education(JXJG-14-19-1, JXJG-15-19-3), JiangXi Science and technology research project of Education Department(GJJ151193) , JiangXi Social Science Planning Projects during the 12th Five-Year Plan(14TQ05) and JiangXi Police College Scientific Research Project(2014QN001).

References

1. Wu Xiao. Based on the user's personalized comprehensive inverted index[J]. Journal of Hangzhou Normal University, 2008(03). (in Chinese)

2. Deng Pan. An efficient inverted index storage structure for[J]. Computer engineering and applications, 2008(31). (in Chinese)

3. Wang Qian. Improvement of DNF algorithm based on inverted index [J]. Information technology, 2014(08). (in Chinese)

4. Zhang Xudong. The inverted index compression algorithm based on 64 bit system [J]. Computer engineering. 2014 (02). (in Chinese)

5. Lin Junhong. Inverted index query processing technology[J]. Computer engineering and design, 2015(03). (in Chinese)