Design and Realization of Topic Search Engine for Chinese Internet Fraud Information

Hu Liang, Ding Ai-chun, Zhu Yu-chi
Department of humanities and management,
Jiangxi Police College
Nanchang, China
huliang_thu@163.com

***Abstract*—According to the China Internet fraud criminal and victim information asymmetry, this paper makes Internet fraud information resources as the research object, constructing a cross platform Internet fraud database by data collection and extraction technology, the realization of the network anti fraud vertical search engine. The study can provide the Chinese public users with online fraud data retrieval, disclosure of new forms of Internet fraud crime, reduce the risk of fraud, and enhance the network security.**

Keywords—topic search engine, Internet fraud, information retrieve

I. INTRODUCTION

The equality and non-centrality of China's network leads to more and more forms of Internet fraud. Therefore, it is necessary to conduct a careful analysis of the Internet fraud, and its methods and characteristics, to find out the countermeasures against the crime of Internet fraud. According to the research, it is found that the asymmetry of the knowledge between the criminal and the victim is one of the important factors of the Internet fraud. Network crime of fraud is indolent, greedy, adventure lucky psychology, but also their higher level of knowledge and innovation ability to learn, more comprehensive understanding of the characteristics of computer and network, has a strong computer and network skills, they set meter with high technology content of fraudulent activities so that the relative lack of basic knowledge of computer and Internet technology, people easily deceived. The part of the victim is younger and inexperienced users, due to less social experience and easily deceived. Another part of the person is elderly, they to newly emerging things lack of understanding, vulnerable to criminals to deceive and fall into the trap of Internet fraud [1]. In this paper, from the angle of the offender and the victim knowledge asymmetry in the Internet fraud crime of fraud of network information resources as the research object, building a cross platform, cross institutional large network public opinion of fraud database by data collection and extraction technology, intends to research and implement based on anti fraud vertical search engine technology and the network public opinion data retrieval and warning service platform [2].

II. SYSTEM ARCHITECTURE DESIGN

Internet fraud information vertical search engine system structure as shown in Figure 1, consists of information collection, indexing and query of three parts. Information acquisition module is responsible for from network collection contains network bilk information page, parsing, extraction and filtering of the text on the page; index module on the web page for sorting, classification and indexing; query module according to the requirements of the user's query, from the index database retrieve the relevant information feedback to the user.

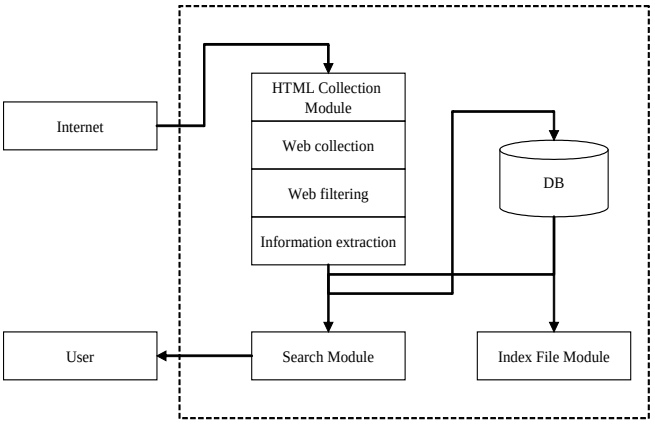

Fig. 1. System architecture of Topic Search Engine for Chinese Internet fraud information

III. SOFTWARE ARCHITECTURE

In order to make the system has good scalability and availability, the model is divided into three layers: application layer is responsible for providing user access interface, mainly HTTP, API and WinSock, etc.. Users can directly access through the HTTP, you can also use API or WinSock to write applications; virtual layer is responsible for providing metadata services, data synchronization services, retrieval services and resource management, etc.. The layer accepts the virtual data attribute described by the application layer, and returns the global address of the virtual data according to the property of the metadata service. Data synchronization services can be replicated in the area of virtual data; resource layer is

responsible for the definition of the communication protocol between the various search nodes, including document management protocol, election protocol, data forwarding protocol.

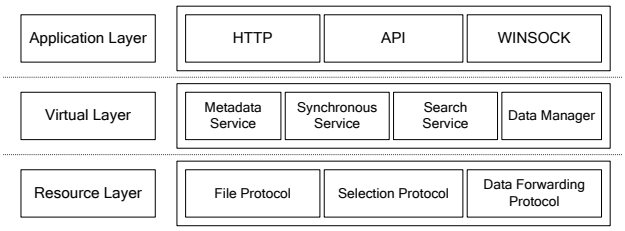

Fig. 2. Software architecture

IV. INTERNET FRAUD INFORMATION COLLECTION

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 public opinion database. According to the model estimates, about media reports of fraud news data volume of about 300 million, Internet Crime Complaint related information is more, probably orders of magnitude in 800 million, large-scale public opinion of fraud databases, public security system have more than 100, the total amount of data about 150 to 200 million, there are other some data from some public fake website.

Comparison of general data collection system, the two have a lot in common, with automatic information collection function, task scheduling, multi thread design and scalability. But the two have their own characteristics, general data collection data objects, including all the contents of the web pages and other files, and the system can be used to collect the data object is contain fraud information content of the data, many of the size of the data is relatively small, so in the design strategy consider collecting update frequency can than general-purpose data acquisition high.

Step1: create a new WebDB;

Step2: the beginning of the grab WebDB into the root URL;

Step3: the new segment from WebDB to generate fetchlist;

Step4: collect web according to the contents of the fetchlist;

Step5: grab back the page link and update WebDB;

V. INTERNET FRAUD INFORMATION EXTRACTION

Fraud information usually contains fraud of host and guest body, date of fraud, fraud tools, process of fraud and some properties description, and the general page there are obviously different. In order to make the computer understand the page meaning must be to organize the data of the non semantic, to filter and extraction technology were devoted to the analysis of the optimization, to strengthen some unrelated words selecting, purifying, fire, further improve the extraction efficiency, for the next step index query to create the conditions.

Fraud information usually contains fraud of host and guest body, date of fraud, fraud tools, process of fraud and some properties description, and the general page there are significantly different, so it is necessary to filter and extraction technology were devoted to the analysis of the optimization, to strengthen some unrelated words selecting, purifying, fire, further improve the extraction efficiency, for the next step index query to create the conditions. This project intends to study the automatic generation of template data filtering and extraction model. After data semantic annotation of additional properties become template instance, by calculating the similarity between the template instances using cluster template examples are divided into different categories, each corresponding to a template, in the same category pattern instance screening information extraction templates.

Evaluation for data filtering and extraction effect, the recall rate and precision to be used as the evaluation criteria, the overall accuracy of the total accuracy is used to describe a source of information containing a plurality of slots. In this study, all the extracted number of correct information expressed by c t, said the number of correct information is not extracted, f extracted error message number is defined, the calculation formula is as follows:

$$GP = \frac{\sum_{solt} c}{\sum_{slot} (c+t)} \times 100\% \qquad (1)$$

$$R = \frac{c}{c+t} \times 100\%, P = \frac{c}{c+f} \times 100\%$$

VI. INTERNET FRAUD INFORMATION CLASSIFICATION

Classification is based on the public opinion data has been mastered by each type of fraud, summed up the rules of classification and establish a formula, and then in the face of new data, according to the formula to determine the type of data.

K nearest neighbor algorithm is used in this paper. The K nearest neighbor algorithm is proposed by Cover and Hart in 1968. K nearest neighbor is to examine the text with the most similar to the K text to be classified, according to the category of the text of this K to determine the category value of the text to be classified. Similar values can be used to determine the Euler distance, or cosine similarity, etc.. And the most similar K text according to the similarity of the text to be classified and the classification value is weighted average, so as to predict the category value of the text to be classified. The formula used in this paper is as follows:

$$D(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad (2)$$

This paper is divided into two parts: the training set and the test set. The training set is a set of text that has been classified, and the test set is a set of text used to test the effect of classification.

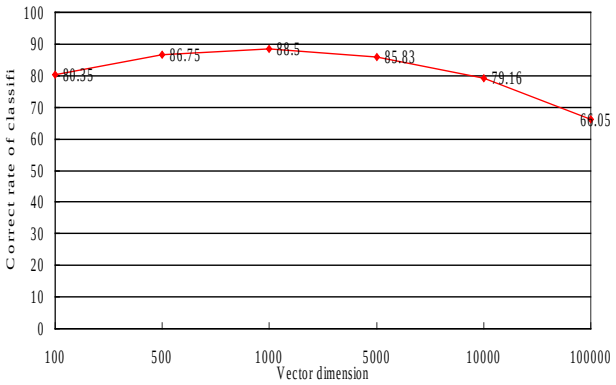

Fig. 3. Classification of KNN algorithm

VII. DISCUSSION AND FUTURE WORK

This paper designs a vertical search engine for Internet fraud information, analyzes the key technologies involved, and discusses the specific implementation methods. The results of this study will not only be able for common user provide Internet fraud data retrieval, disclosure of new forms of crime, reduces the user is fraud risk, but also can through the statistical analysis of data of fraud, the social security administration provides Internet fraud analysis report and auxiliary decision support, in order to improve the working efficiency of the crime of fraud prevention.

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] Li Yude, Xin. Legal thinking about the legal system of Internet fraud[J]. legal system and society, 2008.

[2] Peng Z. On group events and network public opinion[J]. Journal of Shanghai Police College, 2008(1): 46-50. (in Chinese)

[3] Wang Zhihong. The crime of fraud on the network[J]. Journal of Shanxi Police Officers Academy, 2009, 17 (3): 68-70.

[4] Yang Xiejiao, Wei Bin, Zhao Xue. The construction of the Internet fraud status and prevention system[J]. Administration and Law, 2011, (8): 55-60.

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