Research of Information Extraction 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: information extraction, internet fraud, information management platform

Abstract. With the number of China's Internet fraud and complexity increasing, from a large number of network fraud data discovery and summarize the regularity of a variety of Internet fraud formulated is conducive to social stability and development of the decision, and take corresponding measures is becoming more and more important. In this paper, starting from the angle of the offender and the victim knowledge asymmetry China Network in the crime of fraud, fraud information in network as the research object, using data extraction technology to build a cross platform network public opinion of fraud database. Through the statistical analysis of network fraud data contained in the fraud to provide support.

Introduction

Network fraud is the illegal possession for the purpose, the use of virtual reality or the use of the Internet to conceal the truth of the method, the larger the amount of property to cheat behavior. Internet fraud has become the main form of cyber crime, and showed a group and modus operandi of hacker of, harm degree surge of, affected groups to expand, involves the character of the VW range [1,2]. According to China Internet Network Information Center statistics, Internet fraud and other industry tens of thousands of practitioners, which relates to the Internet field segments, 74.1% of Internet users in the past six months met network fraud, the total number reached 4.38 billion. China Electronic Commerce Association released the 2012 China website credible verification industry development report "shows that the total number of Internet users in China reached 5.13 billion, in Internet users online shopping experience, 31.8% have direct encounter fraud website, every year because of Internet fraud caused by the loss of not less than 30.8 billion yuan. Visible, Internet fraud crime is becoming more and more serious in our country, to the state and society caused great loss, serious damage to the interests of the masses of the people, greatly reduces the network integrity, great harm to society[3,4].

Internet fraud information extraction

The information extraction model sends a query request to a specific database to retrieve the corresponding web page, and then the wrapper extracts the information from the web page and maps it into the corresponding tag information, as shown in Figure 1.

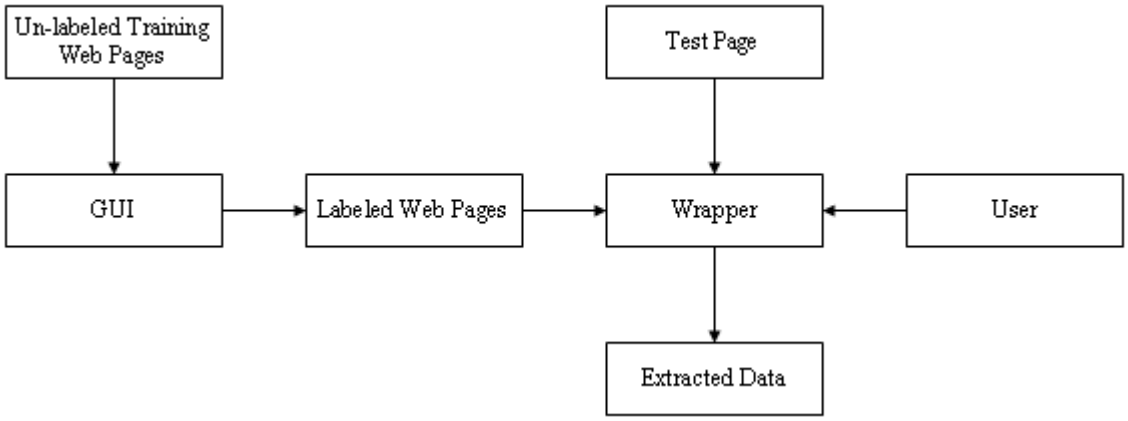

Fig.1 Information extraction model

Search web page contains |L|>0 string unit, each string unit with K attributes, the integer 1 and less than or equal to K is less than or equal to K is attribute index, and 1 and less than or equal to m is less than or equal to |L| said cable quotes page element string retrieval, is each <PBmk, PEmk> said for a single attribute set. PBmk the M string unit of the kth attribute in the search page in the starting position, PEmk is where it ends. Therefore, the value of the K attribute of the first m of the second string unit is in the character content between the PBmk and the PEmk two location points in the search page.

W represents a wrapper that is a function of a set of L mapped to the P of a web page, using a formula that is:

$$L = f(W, P) \tag{1}$$

Algorithm design

For search sub node x grammar space can find consistent with the child nodes, but a NP complete problem. Following the heuristic method and web annotation, punctuation and separators and other tagging sequence sequential feature, in the collection E1{S11, S21,..., Sm1} according to sequential feature selection the longest string, according to the order of reasons is web page annotation is decided attribute, the beginning and the end of the most important sign, followed by punctuation and separate annotation. E1 features the longest string for Si1, Si=Si1xSi2=a1a2. AkxSi2, a x node first marked b, the algorithm is divided into the following two steps:

Step 1: Select s with respect to x of a meta grammar r1=(*.*ak) or r2 = (*...*u(b), r1 or r2 to match s, if both early matching s, is according to the characteristics of sequence from ak-1 to a1 selected a tagging nd improved grammar r1 and r2 r1=r1+ (*.*aj*...*), r2=r2+(*.*aj*...*), so cycle until r1 or r2 recognition x1.

Step 2. Obtained in the first step of r to match the example set E, for each example s, according to the characteristics of sequence from ak-1 to a1 select an annotation, if does not have this mark, is the example set E does not exist in the set of child nodes x consistent grammar, otherwise set the label, the grammar r improvement for r=r+ (*.*aj*...*).

Evaluation formula

The information recall rate R and information accuracy P as the evaluation standard, information accuracy GP for containing a description information source accuracy of multiple attributes, representing all extract the right information using C and t denotes not pulled out of the number of correct information, f represents the extracted error message number is defined as the formula is:

$$GP = \frac{\sum_{solt} c}{\sum_{slot}(c+t)} \times 100\%, R = \frac{c}{c+t} \times 100\%, P = \frac{c}{c+f} \times 100\% \tag{2}$$

Among them, there is an inverse relationship between P and R, if the R increases when P will decrease, while the P minus R will increase. Therefore, in the evaluation of performance will be considered at the same time P and R, the more commonly used indicators for F:

$$F = \frac{(b^2 + 1) \times P \times R}{b^2 \times P + R} \tag{3}$$

Performance evaluation

In this paper, the 300 sites of the web site contains information fraud information for the test set, a total of 10 times, each time selecting 100 samples of the page to mark, and then the information extraction, the results shown in Fig.2.

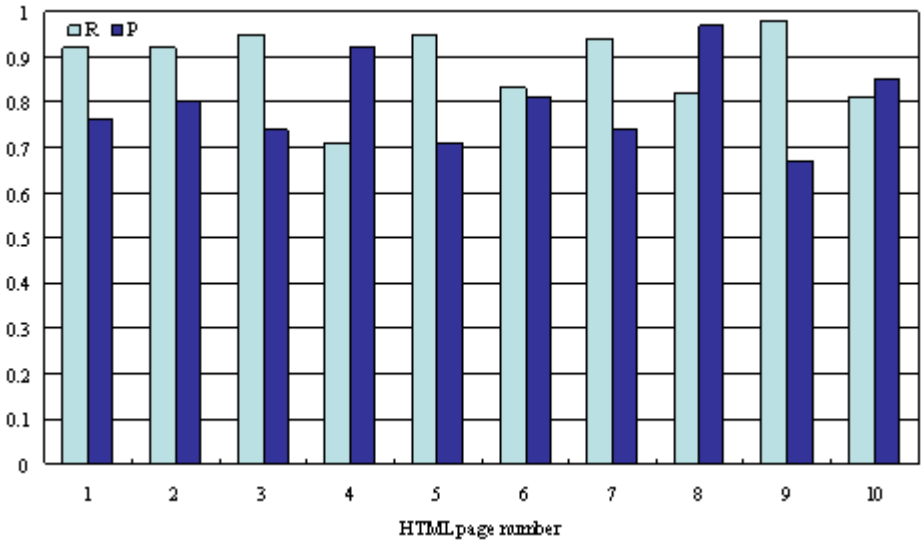

Fig.2 Information recall test

Algorithm of information recall rate is relatively good, the average value of 0.88 and in the algorithm under the condition of sample web page selection of R value impact is relatively small; information precision P average of 0.8, basically and information recall rate R is inversely proportional to the relationship. In order to study the influence of the number of training sample pages on P and R, P and R were tested under different number of conditions, as shown in Fig.3.

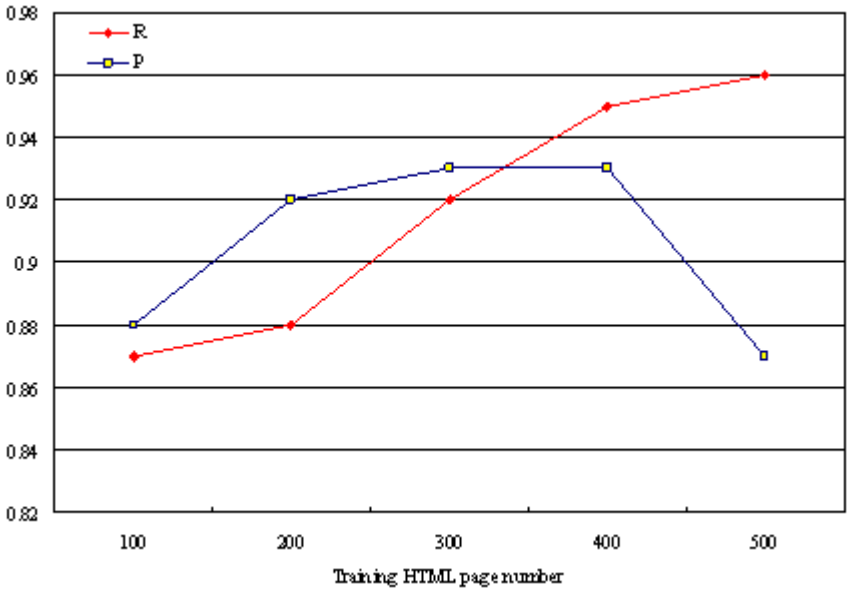

Fig.3 Information recall ratio and precision comparison

With the increase in the sample web page, the R value also increased simultaneously, P value while also increased but in the sample page number reaches a certain value decreased, suggesting that the sample page number is not the more the better, because the information recall rate despite the increase, but the accuracy is lower, so like a sample page number in the 300 or so can be achieved a good balance.

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. Chu. The Internet fraud crime means and features[C]. Chinese criminology research will set of the thirteenth session of the symposium, 2004.

2. Dai Yongwei, Si Zhigang, Fei Huaping. Design of public security decision support system based on data warehouse[J]. Micro Computer Information, 2007, 23(6):179-180.

3. Yang Zhiyong. Internet fraud crime characteristics and measures[C]. National Computer Security Conference, 2008.

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

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

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

7. Lu Xu. On the crime of network fraud and its preventive measures[J]. Journal of Heilongjiang Institute of political science and law management, 2012.

8. Wang Yuan. The network fraud[J]. Merchant 2013, (17): 234-235.

9. Sun Jingjing. Study on the mechanism of intelligence synthesis in the investigation of network fraud[J]. Journal of Railway Police College, 2013, 23 (4): 31-34.

10. Lv Yan. Analysis on the crime of network fraud[J]. Chinese Journal of Leshan municipal Party school, 2013, 15 (1): 102-105.