The Sensitive Information Identification for Internet

Wenqian Shang

The school of computer, Communication University of China, Beijing, China

shangwenqian@cuc.edu.cn

Keywords: Surface web, Deep web, Sensitive information, Text preprocessing, Text classification.

Abstract. With the rapid development of Internet, online information has greatly enriched. The Internet becomes a vast treasure of information, but simultaneously it is also flooding with various trash information, such as: viruses, Trojans, violence, pornography, gambling and so on. The hostile forces outside of country and criminal elements are using the Internet to engage in illegal activities that endanger national security. So how to recognize this information to find the corresponding website and to carry on the effective supervision has become an urgent problem. For these reasons, this paper presents the related technology and solution.

Introduction

With the dissemination of information convenient, efficient, its randomness becomes increasingly prominent. There are a lot of progress, health and useful information and there are a lot of reactionary, superstitious, pornographic and other negative content. The accuracy of media information and dissemination range cannot be effectively controlled. Anyone can publish their expressions and opinions on BBS forums, message boards or self-website. Publishers often do not consider the authenticity of publishing statements and the social impact. As the rapid development of China's Internet, users are immature, the network is lack of monitoring and the related laws are not perfect, so the network is more prone to vulgar, preliminary, violent, pornographic, false statements and so on. This highlights the importance of sensitive information identification, monitoring and filtering. At first, we must get the sensitive information, and then we can deal with them.

At present, the common used techniques of web site information supervision are: (1) search engine; (2) finding of sensitive information based on data packet reassembly; (3) finding of sensitive information based on web page information extraction.

The search engine is also known as information retrieval. It comes into being with the rapid increase of the web page information. Its purpose is to facilitate users to find the information they need from the vast amounts of data [1]. After the user enters the keyword, search engine collects and finds information according to a certain degree of strategy, and then it extracts and organizes information, at last returns to the user in the form of URL. The user browses the web pages that map the URL to determine whether it is the required information [2]. To some extent, this technique can find the sensitive information existing in the web page, but it has some problems as follows: (1) Using artificial means to locate information is lack of intelligence; (2) The search results have low accuracy, it returns large number of irrelevant search content and increases the workload of the user; (3) Keyword-oriented search is only the simple matching keyword. It cannot expand the user query and cannot meet further requirements.

Sensitive information discovery based on data packet reassembly is: In the process of network communication, through the research of the underlying network packet capture technology and application layer protocol for packet reassembly technology, analyzing session characteristics of captured packets, based on session characteristics, analyzing the session layers of protocol, reorganization, splicing, removing consultation, response, retransmission, head of packet and other network information to obtain a complete record of session-based, and then according to the restored session content, using different keyword matching strategy for the storage matching, it can realize the intelligent alarm of the violations and speech for sensitive information. This way can realize intelligent to a certain extent, but has the following problems: (1) it needs to design from the ground

Published by Atlantis Press, Paris, France.

to intercept the data stream and analyze the protocol layer to layer. The hardware costs high and it has low flexibility. (2) In the process of data transfer, packet fragmentation are more, it is easy to cause packet loss and unable to complete the reproduce of the whole session.

Because of deficiencies of the above two techniques, in recent years it appears a new sensitive information discovery technology based on web page information extraction. Using web page information collection technology, it can extract and analyze the web page and its links content to find the sensitive information [3,4,5]. This technology has the following advantages: (1) Adopting intelligent method to locate sensitive information, it has high flexibility. (2) It has high retrieval precision of web information and can increase the recall rate of sensitive content.

With the development of Web 2.0 and the widely using of Ajax technology, more and more valuable information is hidden behind web pages. According to their existence, internet web pages can be divided into surface web and deep web. Surface web refers to the static web pages that hyperlinks can reach and the traditional search engine can index. The deep web refers to those resource that is stored in the network database, hyperlinks cannot access. Deep web was defined by Dr. Jill Ellsworth in 1994 as the web pages that its information content was difficult to find by the general search engines [6]. According to the survey: the deep web containing information is about 50 times of the surface web information. The entire internet has about 307,000 web database site, 450,000 web databases, 1,258,000 query interfaces. The information content of deep web is in the rapid growth, it increases by 3-7 times compared with year 2000. Accessing deep web data is higher 50% than surface web. 95% information on the deep web can be publicly visited, i.e., free of charge. In addition, Deep web has many advantages, such as high quality, rich content, theme specific, and so on. For this, full using of the information in the deep web becomes great significance. Studying deep web becomes an effective way to improve accessing internet information quality and quantity.

In summary, applying extraction and analysis of information technology on web page to monitor the site cannot only improve the recall rate on the web page, but also improve the accuracy of its intelligent recognition of sensitive information. It has high flexibility, easy expansion and maintenance. Therefore, this paper uses a web based structural analysis, web information extraction and analysis techniques, so as to achieve the purpose of identification and regulation of sensitive information.

Sensitive information collection

This section mainly has two parts: one is information collection of static web pages; the other is information collection of dynamic web pages.

The information collection of static web pages can be described as Fig. 1:

This part mainly includes the following major modules and database:

(1) Module of pages download
(2) Module of URL extraction
(3) Pages structure recognition module
(4) Decision tree learning module
(5) Module of extraction rules generation
(6) Text extraction module
(7) Module of page analysis
(8) URL database
(9) Initialize database of pages
(10) Page structure features-extraction rules database
(11) Text extraction results database

The module of page download is used to download the contents of specified page. The module of URL extraction is responsible for extracting the hyperlink in the pages and saving them into the database, so that the web extraction system can make the next processing. Pages structure recognition module is used to analyze and generate the structure features of the pages, then searching a match in the page structure features-extraction rules database. The decision tree learning module executes

related operations based on the matching results of pages structure recognition module. Module of extraction rules generation is responsible for generating the extraction rules for HTMLParser based on the results of the decision tree determine the node of the text. Text extraction module is responsible for reading the rules and extracting the text. Module of page analysis is responsible for invoking the module of URL extraction, pages structure recognition module, decision tree learning module, the module of extraction rules generation and text extraction module. Page structure features-extraction rules database mainly consists of three elements: the string expressing the web structure, the decision tree corresponding to the structure features of pages and the extraction rules generated by the decision tree.

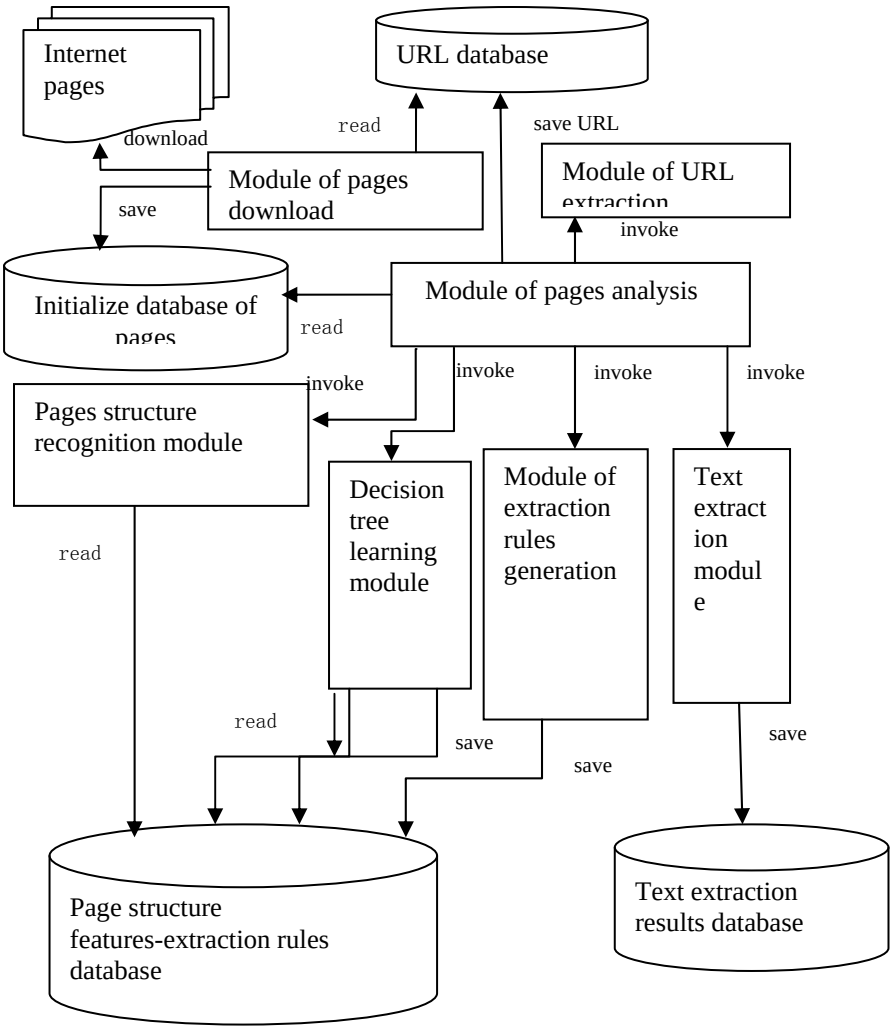

Figure. 1 The structure of static web information extraction part

The information collection of dynamic web pages mainly contains two parts: one is based on server side resources, usually for hidden resources after the form, such as the common database resources by filling out a form to visible part. The research of this part mainly consists of the following modules: page analysis module, automatically fill out form and submit module, analysis results and save module. The detailed description can be seen as Fig. 2. The other is based on the client's resources. i.e., the client side script code dynamically operates the web page. Accessing such resources needs to resolve the problem of script analysis. It mainly contains the following three parts: (1) The JS code is between a pair of marks; (2) The JS code is placed in an external file specified by the src attribute of <script>; (3) The JS code is placed in the event handler, such as onclick, etc. specified by the html attribute.

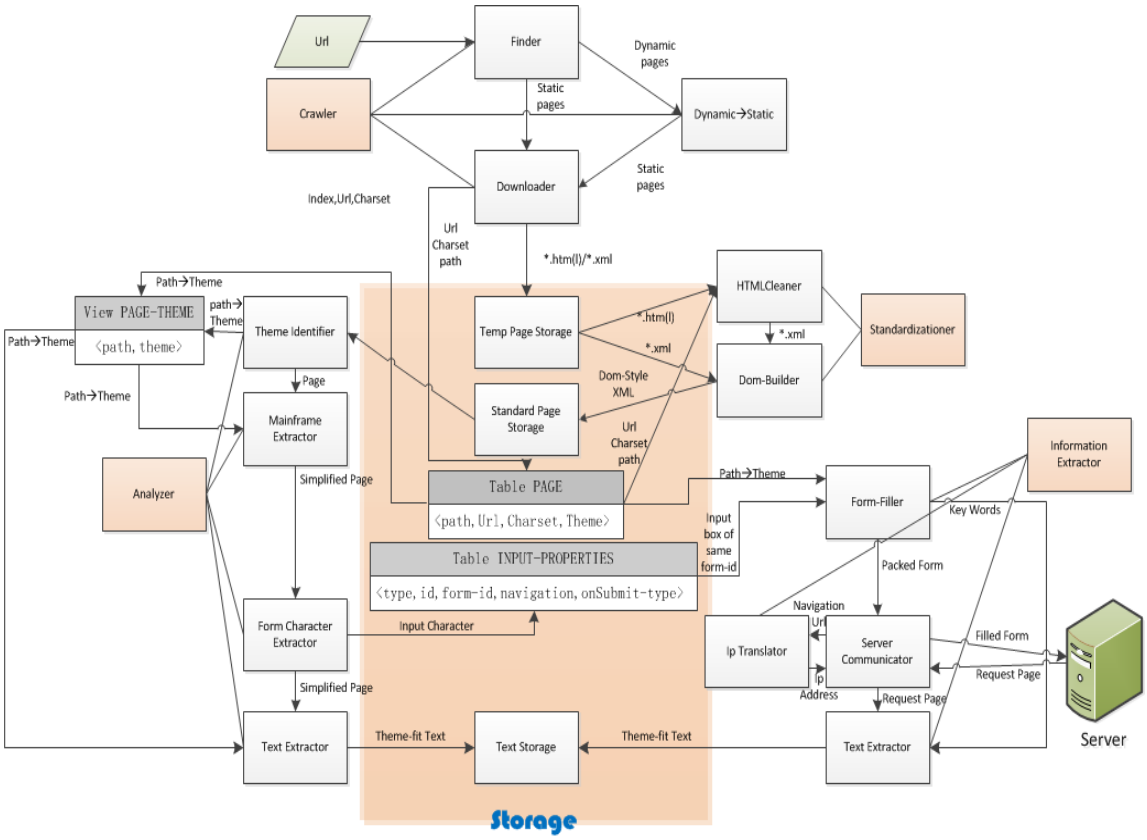

Figure. 2 The structure of dynamic web information extraction part

Sensitive information extraction

This section mainly adopts Dom related technologies, analyzing web page structure, automatically removing unwanted advertising, copyright, columns and other useless information, accurately capturing the main body of sensitive information contents.

Sensitive information preprocessing

Sensitive information preprocessing mainly contains word segmentation, getting rid of stop words, feature selection, feature weight and generation of vector space model. In this paper, we adopt an improvement feature selection algorithm. It can be described as follows:

Charu C. Aggarwal studied the Gini index on feature selection of text categorization, but they used the Gini index of the hybrid degree. Our method is completely different from his, we construct a new measure function of Gini index through in-depth analysis of the Gini index. We use the Gini index of purity for not only the categorization by centroid but also other categorization methods.

The initial form of Gini index is to measure a "hybrid degree", i.e., the property for categorization, namely, the smaller "hybrid degree" the better property. If we use the following form:

$$Gini(W) = \sum_{i=1}^{|C|} P_i^2 \tag{1}$$

It is to measure a "purity" that is the property for categorization, namely, the larger "purity" the better property. This "purity" form of the Gini index can be further changed as follows:

$$Gini(W) = \sum_{i=1}^{|C|} \sqrt{P(W \mid C_i)} \tag{2}$$

Sensitive information classification

After Sensitive information preprocessing, we can classify them into different types, specifically identify the pornography, reactionary, irregularities or vulgar, etc., in order to take different measures. In this paper, we adopt an improvement classification algorithm. It can be described as follows:

The Bayesian approach is a commonly supervised classification algorithm. It bases on the Bayes theorem. It is a kind of pattern recognition method when the prior and conditional probabilities are known.

The decision rule of classical Bayesian classifier is as follows:

$$P(c_j \mid d) \propto P(c_j) \prod_{t=1}^{n} P(x_t \mid c_j) \tag{3}$$

This formula only considers the effect of probability but not considers the effect of feature weight, so we give the improvement algorithm of Bayes. The new algorithm not only considers the effect of probability to classification result but also considers the effect of feature weight to classification result. The new decision rule of classification is as follows, the detailed information can consult [7]:

$$P(c_j \mid d) = \log[P(c_j)] + \sum_{t=1}^{n} TF - Gini(x_t) \times \log[P(x_t \mid c_j)] \tag{4}$$

Summary

This paper gives the whole process of sensitive information identification and at the same time gives the key algorithm of sensitive information identification. In the future, we will improve it further.

Acknowledgement

This work was supported in part by the project “National Science and Technology Support Program (2009BAH40B04)” and partly supported by the project “48th Postdoctoral science foundation of China” (20100480357)

References

[1] K. Jiang and G. S. Wu, in: Overview of Information Retrieval Technology for Web. Computer Engineering, Vol.31(24): 33-836 (2005).

[2] Y. Wu, in: Research on Vertical Search Engine of Recency-sensitive Objects. Zhejiang University doctor thesis, Zhejiang (2011).

[3] M. Wei, M. L. Xi and Y. H. Zhou, in: A Network Data Analysis System Based on WinPcap. Computer Security, Vol. 11: 49-51 (2010).

[4] C. X. Jin, W. N. Qian and A. Y. Zhou, in: Data Flow Analysis and Management, Journal of Software, Vol. 15(8): 1172-1181 (2004).

[5] A. Miure, N. Fujihara and K. Yamashita, in: Retrieving Information on the World Wide Web: Effects of Domain Specific Knowledge, AI & Society, Vol. 20(2): 221-231 (2006)

[6] Y. L. Qu, H. Yu and G. W. Xu, in: Automatic Segmentation of Web Information Block, Journal of Chinese Information Processing, Vol. 18(1): 6-13 (2004).

[7] T. Dong, W. Q. Shang and H. B. Zhu, in: An Improved Algorithm of Bayesian Text Categorization, Journal of Software, Vol. 6(No. 9): 1837-1843(2011).