Research and Implementation of BHO-Based Content Filtering System

Ruibo Li,Xingwei Hao
School of Computer Science and Technology
Shandong University
Jinan, China
liruibo@mail.sdu.edu.cn, hxw@sdu.edu.cn

***Abstract*—Content filtering technology is a hot research topic in the field of Internet application. The traditional filtering algorithms, such as Error Back Propagation algorithm and KMP algorithm, in which sensitive words will be matched with the content to be retrieved one by one, and this reduces the efficiency when massive data is filtered. To solve the shortcoming above, a prototype of content filtering system based on BHO (Browser Helper Object) is proposed in this paper. The system includes URL filtering and content filtering. It stores sensitive words by way of Hash function to improve the retrieval speed, and matches them by using Longest Prefix Match algorithm and Binary Search algorithm. At last, the system is tested in two ways (accuracy and filtering time) and performs well.**

Keywords-content filtering; BHO; URL filtering;

I. INTRODUCTION

With the rapid development of Internet, the information on the Internet increases with a high growth rate, and some sensitive content involving reactionary, obscenity and violence spread widely. In order to get the required information from the Internet quickly and accurately, information filtering technology comes into being. Information filtering is based on the user's information requirements, using some tools to filter the information automatically from the large-scale dynamic information flow [1]. Information filtering includes non-text information filtering and text information filtering, and the latter is the major research direction. The text information is contained in the webpage, so text information filtering is also webpage filtering.

II. RELATED RESEARCH

A. Content Filtering Methods

Traditional filtering sensitive word methods include KMP algorithm [2] and Error Back Propagation algorithm [3]. Firstly, in such algorithms, HTML tags, JavaScript statements and CSS statements are removed from the webpage, and then the rest content is regarded as primary string. Secondly, the primary string will be matched with sensitive words retrieved from the database one by one. If the text content length of the webpage is L, and the amount of sensitive words to be retrieved is N, then the time of filtering process is O (L*N). These methods above are simple and easy to implement, but get poor efficiency when massive data is filtered. In addition, when the number or the content of sensitive word is changed, much modification work will be done with the source code, resulting in wasting much time.

B. BHO Technology

Browser Helper Object (BHO) is a DLL module designed as a plugin for web browsers to provide added functionality [4]. Most BHOs are loaded once by each new instance of browser, and a new instance is launched for each window. Each time a new instance of browser starts, it will check the Windows Registry for the Browser Helper Objects key[5]. If browser finds this key in the registry, then it looks for the CLSID key listed below the key. The CLSID key below Browser Helper Objects tells the browser which BHO to load and the BHO will be prevented from being loaded if the registry key is removed. For each CLSID listed below the BHO key, browser calls CoCreateInstance to start the instance of the BHO in the same process space as the browser. If the BHO gets started and implements the IObjectWithSite interface, then it can control and receive events from the browser. BHO can be created in any language that supports COM [6].

To solve the shortcomings of traditional methods, a prototype of content filtering system based on BHO technology is proposed in this paper. The system stores sensitive words by way of Hash function to improve the retrieval speed, and matches them using Binary Search algorithm and Longest Prefix Match algorithm. A Chinese webpage content filter is designed in this system, which real-time monitors users browsing webpages and shields sensitive information.

III. SYSTEM DESIGN

The content filtering system based on BHO technology uses the mixed operating mode of C/S structure and B/S structure [7]. Its architecture is shown in *Fig. 1*.

Published by Atlantis Press, Paris, France.

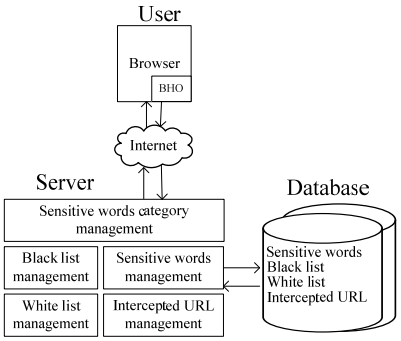

Figure 1. Architecture of content filtering system.

A. *Server design*

The server is designed by the B/S structure, and the main functional modules are shown as follows.

- Managing the black list and the white list. The system provides the default black list and the default white list. Users can also update and maintain them according to their needs. The system automatically updates and maintains them based on the frequency of sensitive words appeared on webpages.
- Managing sensitive word. The system supports CRUD operations for sensitive words and their categories. In addition, it also supports import and export operations for sensitive words and their categories.
- Managing the intercepted URL. The system will automatically record the intercepted URL, the frequency being intercepted, the time of last being intercepted and so on. When the frequency being intercepted of URL reaches a certain level and the system will automatically add them to the black list.

B. *Database design*

The database mainly includes the black list table, the white list table, the sensitive word table, the sensitive word category table and the intercepted URL table. The black list table contains some common malicious websites and URLs, and the white list table contains some well-known websites and URLs.Sensitive words are stored according to their categories, and the category is designed in accordance with The self-discipline norms to prohibit the dissemination of obscenity, pornography and other unflattering information on the Internet published by China Internet Illegal Information Reporting Centre，and it is made up of feudal superstition, ethnic, rumor, national security, terrorism, pornography, gambling, politics, abnormal, insult, violence and religion. Each sensitive word belongs to one of the 13 categories. The sensitive word table keeps in contact with the sensitive word category table by using a foreign key, and the intercepted URL table contains the intercepted URLs, the frequency being intercepted, the time of last being intercepted and so on.

C. *Client design*

The client is designed by the C/S structure, and it filters sensitive words by the webpage content filter. This module is mainly used to check whether if the webpage browsed by users contains sensitive content, so such a process should be in real time and rapid, otherwise, if the filtering operation occurs after the end of the user's browsing or it takes too long, then the filter will be of no significance. So before the user browses the webpage, the filter should filter web content. The filter structure model is shown in *Fig. 2*.

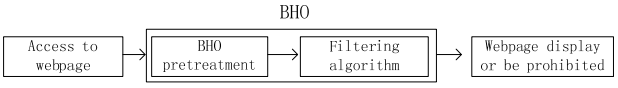

Figure 2. Structure model of the filter based on BHO technology.

When the user requests access to a URL, firstly the browser gets the webpage text from the server, secondly it does some BHO preprocessing with the current webpage content, such as removing HTML tags, JavaScript statements and CSS statements which will not be displayed on the webpage, then matches them using Longest Prefix Match algorithm and Binary Search algorithm, lastly it displays the webpage or prohibits it.

IV. FILTERING ALGORITHM AND SYSTEM IMPLEMENTATION

A. *Filtering process based on BHO technology*

The system includes URL filtering and content filtering, and the process is shown in *Fig. 3*.

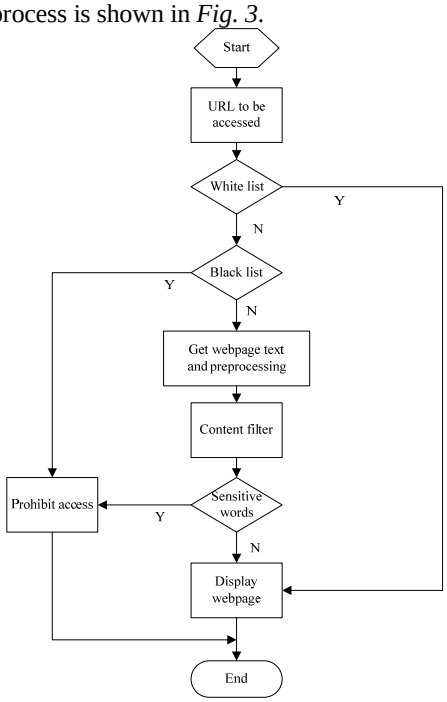

Figure 3. Filtering process of the filter based on BHO technology.

Firstly, matching the black and the white lists with the URL to be accessed, then requesting the server for the

webpage, after getting the webpage content, doing some preprocessing with it by using BHO.Secondly,matching them using Longest Prefix Match algorithm and Binary Search algorithm, if match is successful, then prohibiting accessing, otherwise displaying the webpage.

B. URL filtering

The content filtering system collects harmful URLs from http://webscan.360.cn/url,then stores them into the black list table. The white list table includes Alexa top 500 sites on the web which contain a huge number of safe webpages. The system sorts the black list table and the white list table, then matches them with the URL to be accessed by using Binary Search algorithm, and the time consumed will be O (log(N)), namely, the filtering speed is proportional to the URL’s number of the black list and the white list [8].

C. Access to the webpage and BHO preprocessing

Firstly, the browser requests the server for accessing to the URL’s content, then downloading starts, when it is complete, the DocumentComplete event will be triggered, and after that we can access to webpage content by using the HTML Document Object Model. To determine whether the webpage downloading is complete, we need to get the IUnknown interface pointer to the IDispatch parameter, then comparing the instance of the browser with the pointer pointing to IUnkown interface, if the two pointers are the same, then downloading is complete. After getting the document object, querying the IHTMLDocument2 interface which packages HTML Document Object Model showing all the features of HTML page. Finally, using the get_body method to get the HTML code contained between <BODY> and </ BODY> [9].

After accessing to the webpage content, we should do some preprocessing. There are a large number of noises such as HTML tags, JavaScript statements and CSS statements in the initial webpage and these noises will have a great impact on the webpage content filtering accuracy and speed[10]. We can remove the noises by using the regular expression, and then get the plain webpage content[11].

D. Content filtering

After the preprocessing, we should filter the webpage content. In order to improve the efficiency of retrieving, a special data structure for sensitive word is designed in this paper. Sensitive word is stored in the three-dimensional array dirty_word[i][j][k] by using Hash function, and the data type of array element is character. Dirty_word[i] contains the collection of words or phrases including the same first Chinese character. The subscript i is the same as the Unicode character encoding of the first character. Character data type occupies two bytes in the program, which is the same as the Unicode encoding, so all of sensitive words can be contained in this array. Dirty_word[i][j] contains the sensitive word. Dirty_word[i][j][k] is the k-th word of the sensitive word. Filtering principle is shown in *Fig. 4.*

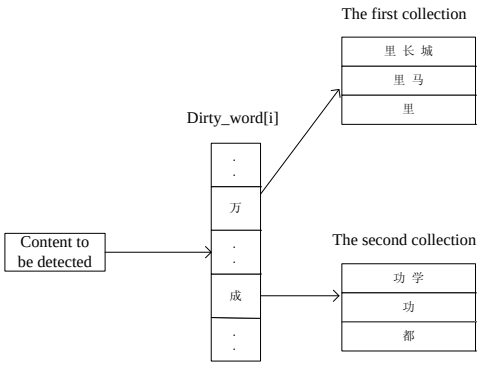

Figure 4. Sensitive word filtering based on Hash function.

For example, The Unicode character encoding of "成" is 0x6210 which is greater than the encoding of "万”, so "万" is in front of "成",and its subscript is smaller. Similarly, "功" is in front of "都".

Sensitive word filtering algorithm based on Hash function is as follows.

1) Load dirty_word[i][j][k] with sensitive words.

a) Read sensitive word records from the database.

b) Store sensitive words into dirty_word[i] by the Unicode character encoding of the first Chinese character. If all sensitive words are read, then proceed the next step, otherwise, jump to the first step.

2) Read a character from the plain webpage content preprocessed by BHO, and find its Unicode encoding, then match it with dirty_word[i].If the character is the last one, the matching fails, and then proceed the step 7. Otherwise, jump to the next step.

3) If there is no sensitive word in dirty_word [i], then jump to the previous step. Otherwise, proceed the next step.

4) Read the next character from the plain webpage content and find its Unicode encoding, then match it with the first Unicode encoding of dirty_word[i][j]. If the matching fails, then jump to the previous step, otherwise, proceed the next step.

5) Determine whether the current sensitive word is end, if it is over, the matching succeeds, then proceed the step 7, otherwise, jump to the next step.

6) Get the next character, and then build up a string with it and the previous characters to be matched. Determine whether it is a prefix of certain a sensitive word, if it is not, then jump to the step 2, otherwise, jump to the previous step.

7) End the matching.

V. RESULTS AND ANALYSIS

The system is mainly designed for Chinese information, and it is tested in two ways in this paper.

- Choose safe webpages and malicious webpages as test data sources to determine whether the system has the capability of intelligent identifying sensitive information accurately.

- Choose massive information webpages as test data sources to determine whether the system can complete the filtering fast.

To get the recognition rate of the system, more than 400 webpages from the Internet are collected in this paper. The amount of malicious webpages is 39, and the amount of each category is 3.The test result is shown in Table Ⅰ.

TABLE I. ACCURACY OF THE BHO-BASED CONTENT FILTER

Webpage Category	Webpage Number	Success	Fail	Recognition Rate
Safe page	412	347	65	84.22%
Malicious webpage	39	34	5	87.18%

To test whether the system can complete the filtering fast, multiple massive information webpages from the Internet are collected in this paper. The test result is shown in Table Ⅱ.

TABLE II. TIME OF FILTERING BASED ON BHO TECHNOLOGY

Sensitive Word Number	Number of Characters	Average Filtering Time
About 1800	9000-10000	654ms
	14000-15000	957ms
	19000-20000	1352ms

We can get the conclusion that the system performs well in accuracy and time-consumed.

VI. CONCLUSION

As the Internet develops rapidly,people can get more and more information through surfing it.However,some lawbreakers disseminate sensitive information on the Internet.In order to shield these information, a content filtering system based on BHO technology is proposed in this paper.The system filters sensitive words by using Hash function,so it speeds up the retrieval process.The tests show that the system takes up less resource with the advantages of high reliability,less time consuming and accuracy of sampling and it provides users with a safe and reliable access to the Internet.

ACKNOWLEDGMENT

Thanks all the people who helped me,especially my tutor,professor Xingwei Hao.

REFERENCES

[1] Yuan Wang. Key Technology Research on Content-Based Text Filtering [D]. Changchun : Northeast Normal University, 2006.

[2] Yuesheng Tan, Ruichun Gu. Application of Improved KMP Algorithm in Deep Packet Filtering Technology [J]. Journal of Computer Applications, 2007, 27(B06) : 217-218, 222.

[3] Yafeng Yao, Xianjin Fang. Research of New Firewall for Content Filtering [J]. Computer Technology and Development, 2010, 20(11): 158-161.

[4] Juan Wang, Yongchong Dan. Research of Network Covert Channel Based on BHO [J]. Computer Engineering, 2009, 35(5): 159-161, 164.

[5] Xiaoxia Rong, Jindong Wang, Shengyuan Wu. Implementation of BHO and Co-operation-based MMC IE [J]. Computer Engineering, 2004, 30(2): 42-44.

[6] Qingbing Sang,Xiaojun Wu. Research and Implementation of BHO-based Website filtering system [J]. Computer Engineering and Applications, 2009, 45(31): 18-21.

[7] Liya Jiang,Hongtao Huo. Filter of Erotic Images Based on IE [J]. Application Research of Computers, 2009, 26(3):1180-1182.

[8] Xiang Rao,Huaiming Wang. Noisy Template Skip List Based Log Filtering in Cloud Systems[J]. Journal on Communications, 2011, 32(7) : 103-113.

[9] Zhenghua Shuai,Xueguang Zhou.Research and Implementation of Content Flexible Filter in Chinese Webpage[J]. Computer and Digital Engineering, 2009, 37(11) : 108-110.

[10] Yijun Gu, Rong Wang, Jianhua Wang. Word-text matrix feature selection in Chinese Text Classfication based on LSI. First International Workshop on Education Technology and Computer Science, 2009, 3: 808-811.

[11] Dan Ma, Hanhu Wang, Mei Chen. A Two-level KNN based Teaching Web Pages Classification Model. International Conference on Networking and Digital Society. 2009, 1: 190-193.