Advanced Engineering Forum *Online: 2012-06-04*
ISSN: 2234-991X, Vol. 4, pp 38-42
doi:10.4028/www.scientific.net/AEF.4.38

Application of Filtration System in the Network Security

Jinliang Bian

QingDao Hismile College, Qingdao Shandong Province，266100

Keywords: Network security; filtration system; network monitoring.

Abstract. With the Internet spread and deepening of the application, the development of enterprises Internet and e-commerce business has become an inevitable trend. Network security has drawn more and more attention. Therefore, it is necessary to monitor network information real-timely by using an effective information filtering systems in the network security. So the paper discussed the system structure, the estimation methods and key technology of the Network Information Filtering System in the general and in detail.

Introduction

With the rapid development of IT around the world, Internet has used and developed adequately, and the enterprises Internet and e-commerce business have attracted more and more attention, which are the main feature of the information age [1]. In a network, the political, economic and commercial activities can be realized with the characteristics of the convenient high-speed and low transaction cost, and it has brought huge benefits to the economy and society. But many criminals try to steal important business or economic information from the network for personal gain, or disrupt the normal order of society by spreading reactionary or yellow information on the network. Thus, in order to improve real-time tracking of network activity, it is quite necessary to increase the information filtering in network security. At present, WWW and email are the main two types of services in the internet, which play an absolute dominant position in information accessing and transmitting, and they are the main communication methods between the Internet and intranet boundary. Therefore, it is very important to implement the real-time monitoring to the information between the two parts [2].

In fact, the existing yellow information in Internet has caused great harm to juveniles, so it shows that a safe and effective information filtration system has become more necessary for the development of network security. Existing traditional approach used for network security mainly is the firewall packet filtering technology; however, there are still nearly 250,000 pornographic websites, so it is bound to have affected to network performance by relying on the traditional firewall technology only [3]. In order to solve this problem, the filtration system used in network security was researched and discussed in detail in this paper.

Filtration Systems

The main function of the network information filtering system designed in the paper are: if there are some not interested information showed in a web page, the user can simple shield of bad information by using some customized settings, which can realize the bad information filtering [4]. The system is designed based on the current system model, and its system architecture shown in Figure 1.

The first three steps of the filtration system are introduced as following.

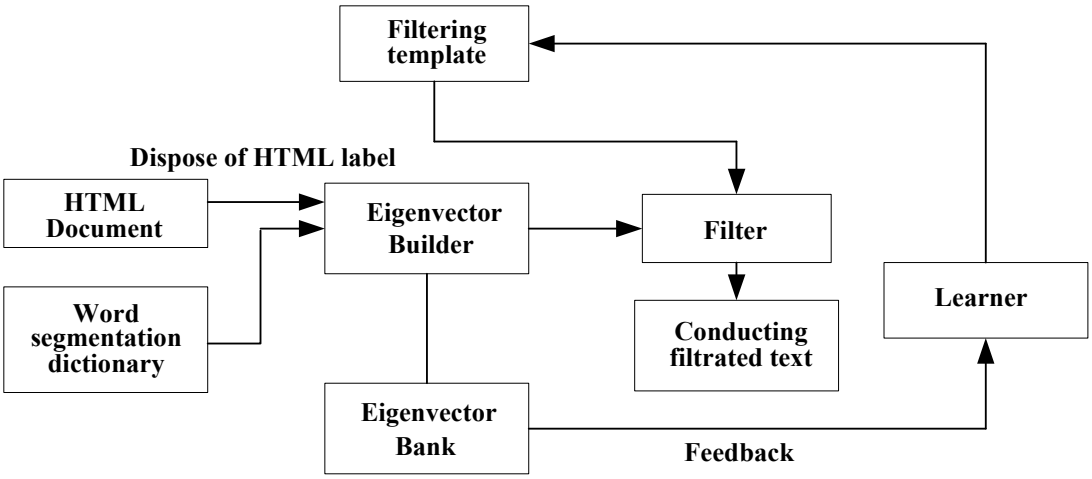

Fig. 1 System structure of filtering system

Building filtering template. Keyword vector $C=(u_1,\cdots u_i,\cdots u_n)$ was used to express filtering template in vector space model, in it n represents the dimension of vector, and u_i represents the corresponding weight of the keyword. We choose a real text from the Internet as an example text, and get the filtering template of the filtering system from the text.

The first thing is to carry out word segmentation conducting of the example text by using word segmentation dictionary, then calculate the weight using the weight function (formula 1).

$$w_{jk}=\frac{tf_{jk}\cdot\log_2(1+n_{jk})^l}{\beta\sqrt{\sum_{i=1}^{m_j}(tf_{ji}\cdot\log_2(1+n_{jk})^l)^2}}(1+\alpha) \quad (1)$$

Where w_{jk} is the weight of the word k in the example text of j, tf_{jk} is the appearing frequency of the word k in the example text of j, m_j is the appearing number in the example text of j, n_{jk} is the appearing paragraph frequency of the word k in the example text of j, β is the scaling factor, l is the word length , $\alpha=0.5$ represent the word appears in the front, title or end, or $\alpha=0$. The designed weight function has considered the appearing position and the frequency in the text.

If the word vector $C_j=(w_{j1},\cdots w_{jk},\cdots)$ was extracted from n-article sample text, where the w_{jk} means the weight of the word k in the example text of j, then the gravity of the N-word vector can be calculated to improve the stability of the system characteristics.

(1) If the importance of N-article sample text is different, the gravity of the word vector can be represented as $C_0^{'}=(u_1,\cdots u_i,\cdots u_n)$, where $u_i=\frac{1}{N}\sum_{j=1}^{N}\lambda_j w_{ji}$, and $\lambda_j(0\le\lambda_j\le 1,j=1,2,\cdots N)$ is the importance ratio.

(2) If the importance of N-article sample text is the same, the gravity of the word vector can be represented as $C_0^{'}=(u_1,\cdots u_i,\cdots u_n)$, where $u_i=\frac{1}{N}\sum_{j=1}^{N}w_{ji}$.

Vector gravity $C_0^{'}$ is a high-dimensional vector, and there is a certain amount of "noise" (the word has a smaller weight values). $C_0=(u_1,\cdots u_i,\cdots u_n)$ is the keyword vector filtrating the "noise", and the vector can be used as the characteristics of the filter information. As filtering accuracy increasing need to constantly change the keyword vector, so the initial filter template is defined as C_0.

Feature vector extraction of been filtrating text. As the word segmentation is the premise basis of feature vector extraction, the extraction rate of feature vector mainly depended on the speed of word segmentation, and it is also the main bottleneck in carrying out the content real-time filtering, so the

main purpose of designing word segmentation algorithm is to improve the speed of segmentation [5]. In the traditional word segmentation algorithm, the change of words needs to re-match character string, which is less efficient. In response to this situation, this paper uses the direct matching method, which is a special word segmentation algorithm and can cut down the re-match work of repeating word string in the traditional word segmentation algorithm and show higher efficiency.

On the assumption of word segmentation, then re-calculated the weight according to equation (1), and then can obtain the text to be filtered feature vectors generated after the word vector de-noising.

Algorithm matching. Algorithm matching is the similarity between the texts of to be filtered and the filter template, that is using the cosine of the angle between two vectors, such as formula (2) as shown.

$$sim(C,D) = \cos\theta = \frac{C \cdot D}{\|C\| \cdot \|D\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i-1}^{n} u_i^{\,2} \sum_{i=1}^{n} v_i^{\,2}}} \qquad (2)$$

From where can be seen that the cosine of the angle and the angle between two vectors is inversely proportional to the size, the smaller the angle, the cosine of the greater, indicating that the greater the degree of similarity between two vectors, increasing the waiting filter text to meet the possibility of filtering requirements. Once again given a filter threshold ψ , and satisfy with $sim(C,D) \geq \psi$, indicating that the corresponding content of feature vector D can meet the filtering requirements, and the diffusion and transmission in the network will be prohibited.

Analysis of major technology

To design the needful template for users. To get needful information for users. Different in accordance with the active side, the user needs is roughly divided into three ways: First, users can get demand information by filling in keywords. This way is affordable, quick and easy, small feature of the system overhead; but it also increases the burden on the user, and the effectiveness and timeliness of service is no guarantee if the user can't give out a clear information demand. Second, the system can track the users of no explicitly involved in the user tracking system; the system would get the needs of users based on user behavior that is to use implicit feedback to achieve the user's interest. Third, the evaluation given by the user through the display of information can use to learn the needs of users, and that is the explicit feedback learning method to get the user's interest. This method can effectively avoid the difficult of choosing the keywords, and is clearer in showing information demand for the potential users.

To describe the needful template for users. The analysis methods of discovery rules, keywords or classification can be used to describe the needful template for users. Usually, the description of the needful template for users is related with the matching algorithm and the description of network document, and every user's needs template can be used as an information document that can organize in specific way and store in the client side, server-side, and proxy-side.

Matching techniques. Information filtering system based on the users' access the network using the appropriate matching algorithm to compare the information document and the user needs template. The user needs information in existing systems typically use keywords, classifying method and regulation to describe, and a different description method has a different matching algorithm. Such as the system of based on keywords description commonly used the matching techniques of Vector Space Model, Boolean Logical Model or Probabilistic Inferential Model; for the system of based on classification describe is suitable for automatic classification Bayes classifier and TFIDF classification methods to match; and for the system of regulation describe can be projected by the rules of the user which even if not seen the information but may be interest to it.

Describing the network information document. There are there existing models in describing the network information document: the Boolean Logical Model, the Probabilistic Inferential Model and the Vector Space Model. These three models have been widely used that is mainly due to the computational complexity is less:

1) Boolean Logical Model. In this model, the relationship that exists between the keywords and Boolean operators described in the document features.

2) Probabilistic Inferential Model. The model achieves information filtering through similarity calculation between the page and the document of user requests.

3) Vector Space Model. In this model, the smallest unit of the document describes is the features and the document is a collection of a series of features. So a vector can be used to represent a document in which the number of features is the dimension of the vector, and it is effective to transfer the document information matching and change into the space vector expressed. According to the characteristics of the space vector, the angle between two vectors can be used to represent the similarity between two documents, and the smaller the angle indicates a greater the similarity.

Feedback mechanism. The user needs to have a gradual clear process, and has a dynamic filter change, so it is necessary to have a feedback mechanism that can track changes in user information needs, and which can timely changes the user needs template. This process is the learning process of the user needs template. The learning method in the information filtering system has the following four main ways:

1) Direct learning. Based on the filter results, the user can directly add, delete, or add some rules to modify the demand templates according to their need beyond the current technical conditions, which is one of the most effective learning methods.

2) Semi-direct learning. According to the evaluation of user filter results, the system can carry out adjustments to the needs template.

3) Indirect learning. System can obtain information in a user browsing behavior without requiring the user to provide information.

4) Collaborative learning. It will be interested in similar or the same user to constitute a group, in which group the changing needs of any member of the group will play a catalytic role.

Now scholars at home and abroad also study and introduce new learning methods. Such as the machine learning methods and artificial intelligence were introduced into the information filtering systems, and some method was use to determine the similarity of the document and user information needs, such as the use of genetic algorithms, nearest neighbor method, neural network and support vector machine, etc. All of this can real-timely give feedback to the changing needs of users and effectively improve the efficiency of the filter.

Evaluation methods and results analysis of network information filtering system

Experimental evaluation. First, analytical evaluation It should be built an analytical model for the performance system. Losee had pointed out that the filtration system behavior was described by analytical system model using a set of equations, and the characteristics quantities of the system was deduced by mathematically. Efficiency and performance of such systems was predicted by the analytical model (such as the reliability of a method can directly determine the speed of learning). However, the increase in the complexity of the filtering system also increased the difficulty of developing analytical models.

Second, simulation-based evaluation Comparative analysis also can be used to evaluate other methods. It was determined the filtration method and other systems using the performance of different system, or comparing the performance of the system between the expected results and the best results, the simulation-based evaluation can evaluate the system fairly and objectively but in order to obtain conclusions, it will be summarized the results and reduced its accuracy.

Analysis of results . This paper implements the filter by choosing the Asian financial crisis as the theme and selected the sample text and filter template on the Internet, and then 500 of the different length article was detected in this paper. The results were showed that if the real-time certain, the Filtering accuracy was up to 69.7% (such as figure 2), it was indicated that the initial filter template was fully comply with the requirements of the content filtering.

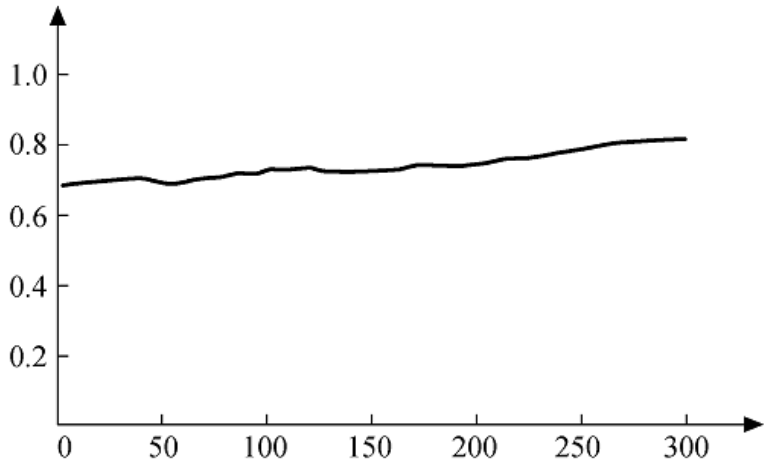

Fig. 2 Relation between filter templates and filtering accuracy

Figure 2 showed the changing relationship of the filter templates and filtering accuracy. In the Figure, the abscissa was the number of filter template cases, and the vertical axis was the filtration accuracy. It can be seen that the figure of filtration accuracy was proportional to filter template cases. It will increase with increase of the positive cases of the filter template. It also showed that the filter template will be closer to the real template if the vector value method was adopted to adjust the filter template.

Conclusion

Filtration system that was used in the network security was a real-time monitoring of network information system or a monitoring system that implemented the network information flow to achieve a particular purpose, which was a key tool to manage and maintain network security for network administrators. For the shortage of conventional filtration systems, the filtration systems used in this paper were completed the initial design goals after the actual testing and running. Filtration systems had important application value and significance for deleting and selecting of network information in the days that network security is increasingly important.

References:

[1] Huang Xiaobin, Qiu Minghui. Study on Network Information Filtering System[J]. JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION 2004, 23(3): 326-332.

[2] Luo Dongyun, Sun Xiaoming, Xu Qin. Netword platform technology research and development[J]. Manufacturing Automation, 2010, (04): 114-115.

[3] Cheng Ni, Cui Jianhai, Wang Jun. Overview of Research on Foreign Information Filtering Systems[J]. NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE, 2005, (6): 30-38.

[4] Dai Yijia. Computer network security[J]. Manufacturing Automation, 2011, (01): 171-172.

[5] Huang Xiaobin, Yue Minghui. A Comparative Study On Network Information Filtering Methods[J]. Journal of Academic Libraries, 2005, (10): 42-48.