Research on Adaptive Filtering Model of Illegal Webpages

Ronghua Lu
Jingdezhen Ceramic Institute, Jingdezhen 333403, China
lrhlucy2003@163.com

Abstract
In order to enhance the information security of campus network, this paper researches on the key technology of web based adaptive and intelligent webpage filtering. The model is constructed by adaptive sample library illegal category feature character, illegal class model based on intelligent, and eventually realize the illegal webpage filtering. The experimental results show that the proposed model can filter the illegal webpages in the document collection with fast speed and high accuracy.

Keywords
Text filtering; Frequency statistics; Adaptive; Character.

1. Introduction
The construction and implementation of campus net, greatly facilitate the construction of school of network education, graphic information center. This is good for teachers and students to work and study, enrich their life. However, all kinds of unhealthy bad text (pornography, violence, reactionary) has been rampant in the network. How to effectively filter the illegal Webpages quickly, build healthy campus network environment, has become one of the main contents in the research of the campus network security.
In this paper, the illegal webpage filtering as the main research object, creatively put forward the concept of vector fonts, and designed an adaptive illegal webpage filtering model using frequency statistics and data mining technology.

2. Adaptive Illegal webpage Filtering Model
Adaptive illegal webpage filtering mode is divided into two stages: training and filtering. In the training stage, the author uses the adaptive sample library to construct the illegal class type pattern, the sex medicine category pattern and the sex literature category mode, thus forming an intelligent filtering mode; The filtering stage is to quickly classify and predict the test document set through the created filtering mode and user set threshold.

3. Text Representation and Vector Font
This paper uses the vector space model, each document will be regarded as vector composed of Chinese characters of characters, the total number of each document (all appear not to repeat the Chinese characters) is the dimension of document vector, the weight of the character is stored in a number of Chinese characters (such as 80) in the frequency of Chinese characters. Therefore, a document can be represented as \( D = (T_1, W_1; T_2, W_2; \ldots; T_n, W_n) \). Among them, \( T_k \) is Chinese characters character, \( W_k \) is the frequency statistics of the character in the document value, where \( 1 \leq k \leq N \).
For each document, which in accordance with the order of frequency Chinese characters character has a unique arrangement, "Chinese characters, character vector frequency, frequency ranking" as the main
information in this paper is called the word vector. In each category of training samples for all sets of documents, statistics of each category respectively Chinese characters appear in the document and all documents under the category of concentrated frequency, frequency ranking and other information, resulting in each category word vector font.

4. Adaptive Classification Model

4.1 Adaptive classification algorithm

1) From the illegal training sample set generation illegal vector font;
2) To generate vector font from the general category of the training sample set;
3) Statistical illegal vector font and ordinary vector font, and generate illegal feature library;
Operation method is: The calculation of all two font in the same characters in different fonts in the frequency of absolute difference, and based on the absolute difference between the high frequency sequence, The bigger the difference is, The bigger the difference is, the stronger the legitimacy of the subject is. In this experiment, this kind is defined as the absolute difference frequency word for word, illegal construction features from these feature character font can and frequency value, frequency of absolute difference, ranking and other information.
You can use the following query to obtain illegal feature characters:
Select top 350 a.word, (a.frequence-b.frequence)as frequency difference,
a.id as '[Illegal class:id', a.rank as rank, a.frequence as frequency, a.type as 'type ]',
b.id as '[Ordinary class:id', b.rank as rank, b.frequence as frequency, b.type as 'type ]'
From (select * from docsetfrq where word not in (select word from st_stopword))as a , docsetfrq as b
Where (a.type=1 and b.type=0 and a.word=b.word and (a.frequence-b.frequence)>0.02 )
order by (a.frequence-b.frequence) desc
4) Using the feature threshold feature extraction method from font before several feature words and word frequency value as legal standard mode.

4.2 Adaptive Classification Algorithm

1) Construct the feature vector of the test document
2) The test documents and the illegal class standard mode comparison, calculate the similarity, according to the threshold to filter judgment
In this paper, the following 3 methods are used to calculate the similarity between the standard and the test document.
Minimum two difference method: In this paper, a minimum of two difference is illegitimate standard test model and document feature vectors in the same character corresponding to different frequency absolute difference between the cumulative values.
The characteristics of character frequency of addition: In this paper, the characteristics of character frequency is the accumulated value of each character in the test document feature vector corresponding to the different frequency of the accumulated value.
Cosine metric method: In this paper, we calculate the similarity between Illegal document vector and test vector.

5. Experimental Results and Analysis

In the experiment, we constructed an illegal class classification model. The results of the experiment are shown in Table 1. Where 120 documents as a test document, which contains 25 pieces of illegal documents, the text of the general category of the document, the 10 sex medicine category, the 5 sex literature.
Table 1. Experimental result

<table>
<thead>
<tr>
<th></th>
<th>The default threshold</th>
<th>Statistics of the number of illegal documents</th>
<th>Statistics of legal documents</th>
<th>Number of documents to be judged by mistake</th>
<th>accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real situation</td>
<td></td>
<td>25</td>
<td>95</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>The characteristics of character frequency of addition</td>
<td></td>
<td>20</td>
<td>36</td>
<td>84</td>
<td>13</td>
</tr>
<tr>
<td>Minimum two difference method</td>
<td></td>
<td>5</td>
<td>64</td>
<td>56</td>
<td>71</td>
</tr>
<tr>
<td>Cosine metric method</td>
<td></td>
<td>0.6</td>
<td>28</td>
<td>92</td>
<td>12</td>
</tr>
</tbody>
</table>

Thus it can be seen in the classification of the prediction of the 3 algorithms, the cosine metric algorithm is the best, basically can be very high accuracy filtering illegal webpages; The characteristics of character frequency of addition while filtering illegal "the highest accuracy, but the identification of medical" than the cosine measure method. The least two difference method is the worst.

6. Conclusion

In this experiment, the adaptive classification model can be used to filter the illegal webpages with satisfactory accuracy, But the use of such a filter mode based on word frequency statistics and data mining technology in the processing of related literature , low accuracy, often mistaken for illegal webpage filtering will. In the future, it can be used to construct the classification model of medical and sexual literature, and is used for two judgments, so as to improve the accuracy of the system.

Acknowledgements

This paper was supported by Humanities and Social Science Research Project in Colleges and Universities of Jiangxi Province --“Research on Webpage filtering technology based on Data Mining (NO.JC1557) ”, China and 2016 Jingdezhen Social Science Planning Project--“Research on the Countermeasures of network group events based on Web mining”, China.

References