Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 3, No. 3, 2022

Construction of Unhealthy Webpage Filtering Mode Based on Data Mining Technology

Ronghua Lu

Jingdezhen Ceramic University, Jiangxi, 3334403, China

Abstract: The unhealthy category eigenvector library is constructed by adaptive sample library, and the unhealthy category model is constructed on this basis to realize the filtering of unhealthy webpages. Our expermient proves that this mode can filter unhealthy webpage at higher speed and satisfying precision.

Keywords: Data mining, Web filtering, Feature word, Vector space model, Characteristic vector library.

1. Introduction

Internet technology has made the sharing and release of information across the limitations of time and space. Internet has become an important source of information for people to obtain information. While Internet has brought us information explosion and rapid economy, it has also brought a lot of unhealthy information that is not conducive to economic development and people's lives. For example, various unhealthy texts (pornography,violence,reactionary) on Internet are flooding. How to effectively filter unhealthy webpages and give network users a ‘green’space is an urgent problem to be solved.

2. Text Representation and Characteristic Vector Libraries

The current widely used text representation models are: Boolean model, probability model and vector space model. In this paper, we use the vector space model, each document is regarded as a vector composed of items (Composed of seperate word, words or phrases), these items are called the dimension of the vector; the document set is regarded as a vector space composed of documents. In the vector space model, each document is represented by vector, and the dimension of the vector is the number of elements in these item sets. Document D can be expressed as $D=D(T_1,T_2,\ldots\ldots,T_n)$, where T_k is a character item,$1\leq k\leq n$.

For a document with n items, the item T_k is often given a certain weight W_k, indicating its importance in the document, that is, $D=D(T_1,W_1;\ T_2,W_2;\ldots\ldots;\ T_n,W_n)$,abbreviated as $D=D(W_1,W_2,\ldots\ldots,W_n)$,$W_k$ is also called the word frequency statistic of the character in the document.

In this paper, the Chinese character items in each document are formed into a sequence according to the order of word frequency, and the vector with the main information of ‘Chinese character items, word frequency and word frequency ranking’is called ‘characteristic vector word’. That is to say, characteristic vector word is a word with multi-dimensional information such as word frequency obtained after the word frequency statistics of the document set, expressed as: characteristic vector word =(character items, word frequency, word frequency ranking, other field information).

The training corpus of each category is used as the document set, and all of Chinese characters appearing in the document set of each category, and word frequency and word frequency ranking of all of document sets under the category are counted respectively, so as to obtain characteristic vector library of each category.

3. Creation of Unhealthy Class Characteristic Vector Word

If the word frequency value of a characteristic vector word in the unhealthy class Characteristic Vector library is larger than the absolute value of the word frequency value in the common class characteristic vector word, and it belongs to the non-stop vector word, it is considered that the vector word has a strong unhealthy tendency, which is called the unhealthy class characteristic vector word.

The construction steps of unhealthy class characteristic vector library can be summarized as:

1)Construct the characteristic vector library of common classes;

2)Construct the characteristic vector library of unhealthy classes;

3)By comparing common class characteristic vector library and unhealthy class characteristic vector library, extracting the characteristic vector with large absolute difference of word frequency and it's not belong to stop vector word;

After constructing common class characteristic vector library and unhealthy class characteristic vector library, the creation of unhealthy class characteristic vector word can be obtained using the following query statements:

Select top 350 a.word, (a.frequence-b.frequence) as frequency difference.

a.id as ‘[unhealthy class: id’, a.rank as rank, a.frequence as frequence, a.type as ‘type] ’.

b.id as ‘[common class: id’, a.rank as rank, a.frequence as frequence, a.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

Note: The above thresholds can be adjusted according to actual needs. In the above statements, a.type=1 and (a.frequence-b.frequence)>0.02 ensures that the frequency of type=1 is at least above 0.02, that is, a.frequence>0.02.

4. Creation of Stop Vector Word

If a character has a large frequency value in common class characteristic vector library and unhealthy class characteristic vector library, and both rank relatively high, it is considered that the classification tendency is very weak. If it often appears in the document set. After removing the character, it does not affect the theme analysis of the text, and this kind of vector word is called stop vector word.

The construction of stop vector word library has a premise: given a sufficient amount of training document set, build common class characteristic vector library and unhealthy class characteristic vector library.

Steps of create library of stop vector word:

1)common class characteristic vector library is generated by the common class adaptive training sample library;

2)unhealthy class characteristic vector library is generated by the unhealthy class adaptive training sample library;

3)Using the rule of stop vector word and comparing the unhealthy class characteristic vector library, plus manual judgment, library of stop vector word is generated. This kind of stop vector word is generally small, generally called library of stop vector word.

Based on both of characteristic vector ranking and word frequency two fields, set a threshold, and build a rough library of stop vector word, for example: We can use the following query to obtain stop vector word from the common class characteristic vector library “docsetfrq”:

Select top 100 * from docsetfrq d

Where (d.type=0 and d.rank<100 and d.frequence>0.2) order by d.rank

5. Two rules Obtained from Experimental

Rule 1(rule of stop vector word): After constructing the common class characteristic vector library, according to the word frequency ranking of the characteristic vector, the top-ranked feature vectors are generally meaningless, and the elimination of such characteristic vectors does not affect the theme of the document. Therefore, such characteristic vectors can be used as stop vector word.

In the actual construction of rule of stop vector word, it is also necessary to determine an appropriate size of stop vector word based on the experience of rule of stop vector word.

Rule 2(rule of feature word): The corresponding word frequency of the same character in common class training corpus and unhealty class training corpus is usually different. If the absolute value of the word frequency difference of the same character in common class characteristic vector library and unhealty class characteristic vector library is sorted, the larger the difference, the greater the tendency of the unhealty class category, indicating that the higher the feature strength of the word, it can be called "unhealty class characteristic vector word".

6. Construction and Prediction of Unhealty Category Pattern

1) unhealty category pattern building algorithm, the steps are as follows :

① Using characteristic vector library and generating unhealthy class characteristic vector library according to the rule of characteristic vector ;

②using the feature threshold to extract the first several characteristic vectors from the unhealthy class characteristic vector library as the unhealthy class standard pattern ;

Methods : The feature threshold is set (the absolute difference of word frequency is specified for filtering, which can be adjusted according to efficiency and filtering accuracy). From the unhealthy class characteristic vector library, according to the absolute difference of word frequency, it is compared with the given threshold respectively. The higher is selected and the lower is eliminated. Finally, a set of unhealthy class characteristic vector word which meet the requirements can be obtained. After vectorizing these feature words and their frequency values, "unhealty category pattern" is obtained.

2)Unhealthy category pattern filtering algorithm, the steps are as follows :

①Construct the characteristic vector of the test document ;

The text content of the webpage is extracted from the test document, then the word frequency of the unhealthy category characteristic words appearing in the document is counted, and a multidimensional characteristic vector is constructed with these characters and their word frequency in the document.

Assumption: $D=D(T_1,W_1;T_2,W_2;.....;T_n,W_n)$, where T_k is the feature character term, W_k is the word frequency statistics in the test document,$1\leq k\leq n$.

②Compare test document characteristic vectors with unhealthy category patterns, calculate the similarity, in order to filter according to the threshold.

Precondition: Suppose the unhealthy class pattern string extracted from the database is (a1,a2,...,an) with frequencies of (f1,f2,...,fn), their frequencies in the test document are (tf1,tf2,...,tfn).

In this paper, the feature word frequency accumulation method and cosine metric method are used to calculate the similarity between unhealthy category pattern and test document characteristic vector which are described as follows:

①Feature word frequency accumulation: cumulative value of each character in test document characteristic vector is called feature word frequency accumulation value.

unhealthy category mode character frequency cumulative value:$Weight_1=f_1+f_2+...+f_n$;

test document vector character frequency cumulative value: $Weight_2= tf_1+tf_2+...+tf_n$;

difference $Weight=|Weight_1-Weight_2|$;

Compare Weight with threshold to make predictions or calculate similarity.

②Cosine metric: Let v_1 and v_2 represent a document vector respectively, then the similarity(cosine metric) of the two documents is $sim(v_1,v_2)= \frac{v_1.v_2}{|v_1||v_2|}$, where the inner product $v_1 \cdot v_2$ is the standard vector dot multiplication, that is$\sum_{i=1}^{n} v_{1i}v_{2i}$, $|v_i|=\sqrt{v_i \cdot v_i}$.

③The calculated similarity is compared with the filtering threshold, and the filtering judgment is made accordingly. There are three types of judgment results: unhealthy text class, which needs to be filtered; non-unhealthy text class, need to pass; suspected text class, need further judgment.

The algorithm is described in detail in the form of pseudo-code below:

File name: Modeljudge.java

Description: “Unhealthy Category Pattern Filter”, that is, unhealthy class prediction model to determine whether a webpage is an unhealthy webpage

Input: the path to the text file or web file

Output: Judging results

Presupposed conditions:

The characteristic vector class St_character=(word, f_difference, frequency1, frequency0), where word represents the characteristic vector character, f_difference represents the word frequency difference, frequency1 represents the word frequency value in unhealthy class characteristic vector library, and frequency0 represents the word frequency value in common class characteristic vector library.

Algorithm steps :

Step 1:Initialize :

step 1.1 File type file=Get the test file through the given path ;

step 1.2 Feature class vList=Get the unhealthy class feature character set sorted by frequency value from the database ;

step 2: Perform unhealthy class feature word frequency statistics on the test file"file", and put the results into vList :

if (file is not empty and file is a file) {

string type fileText=reads file contents from file ;

if (vList is not empty) {

for (int i=0 ; i< size of vList; i++) {

set the frequency0 domain of the i-th element of vList to 0;

set the f_difference1 domain of the i-th element of vList to 0;

the character corresponding to the i-th element of the character variable word=vList ;

count word frequency in the test document fileText and store it in the frequency0 domain of the i-th element of vList ;

}

}

}

Step 3:Using the stored values in the feature word class variable vList, a prediction algorithm is used to achieve classification prediction.

Method 1: Feature word frequency accumulation, the algorithm is :

floating point number type frqSum=0; //cumulative value of unhealthy class character frequency in test sample .

floating point number type frqSum1=0; //cumulative value of unhealthy class feature word frequency in unhealthy characteristic vector library.

for(int i=0; i<size of vList; i++) {

if (the i-th character of vList is Chinese) {

frqSum=frqSum+ the domain value of the frequency0 domain of the i-th element of vList;

frqSum1=frqSum1+ the domain value of the frequency1 domain of the i-th element of vList;

}

}

if (|frqSum-frqSum1|> changing the threshold)

determine that the document is an unhealthy class webpage, need to filter;

else

determine that the document is a non-unhealthy webpage, passable;

Method 2: Cosine metric method, the algorithm is:

Initialize floating point number type: sim=0, v0v1=0, v0v0=0, v1v1=0, sqrt_v0v0=0, sqrt_v1v1=0;

for(int i=0; i<size of vList; i++){

if(the i-th character of vList is Chinese){

v0v1+=the domain value of the frequency0 domain of the i-th element of vList * the domain value of the frequency1 domain of the i-th element of vList;

v0v0+=the domain value of the frequency0 domain of the i-th element of vList * the domain value of the frequency0 domain of the i-th element of vList;

v1v1+=the domain value of the frequency1 domain of the i-th element of vList * the domain value of the frequency1 domain of the i-th element of vList;

}

}

sqrt_v0v0=square(v0v0);

sqrt_v1v1=square(v1v1);

sim=v0v1/(sqrt_v0v0*sqrt_v1v1);

if(sim> changing the threshold)

determine that the document is an unhealthy class webpage, need to filter; else determine that the document is a non-unhealthy webpage, passable;

7. Experimental Results and Analysis

This paper selects 100 documents as test documents, including 15 unhealthy documents, 70 common documents, 10 sexual medicine documents and 5 sexual literature documents. The experimental results are shown in Table 1:

It is concluded that word frequency accumulation and cosine measure algorithm can filter unhealthy webpages with very high accuracy from Table 1 ; although the feature word frequency accumulation method has higher accuracy in filtering unhealthy webpages, it is worse than cosine measure method in identifying sexual medical webpages.

Table 1. Experimental results

	default threshold	Statistics of the number of unhealthy class documents	Statistics of the number of non-unhealthy class documents	Misjudgments (Upper limit of misjudged documents)					accuracy
				Total number of misjudged documents(100)	Number of Misjudged Common Class Documents(70)	Number of Misjudged sex medical documents(10)	Number of Misjudged Sex Literature Documents(5)	Number of unhealthy class documents(15)	
the real situation		15	85	0	0	0	0	0	
Characteristic vector word frequency accumulation method	12	18	82	12	2	5	4	1	88%
Cosine metric method	0.6	13	87	14	4	3	4	3	86%

Acknowledgment

This work was supported by the Science and Technology Project of Jiangxi Provincial Department of Education, and the project number is GJJ170797.

References

[1] Zhu Guojie.Research and Implementation of Bad Webpage Detection System Based on Text Features [D].Zhejiang University. 2020.

[2] Wang Lei.Research on double filtering method of bad web pages based on content recognition [D]. Jilin University. 2012.

[3] Si Derui. Research on Web Filtering Technology Based on Text Content [D]. Lanzhou University. 2008.

[4] Jiawei Han, Fan Ming&Meng Xiaofeng. Data Mining Concept and Technology (3rd Edition). Machinery Industry Press. 2012.

[5] Tang Jiangang, Xiong Guoping. Research and Application about Unhealthy Webpage Filter Model Based on Words' Frequency and DataMining Technology[J]. Journal of Xiamen University (Natural Science), 2007(11).

[6] Ronghua Lu. Design of Bad Information Filtering System for Web Pages[J]. Frontiers in Computing and Intelligent Systems, 2022(3).

[7] Yao Mei. Research on Key Technologies Based on Web Content Filtering [J]. Information and Computer (Theoretical Edition . 2022 (14).