*International Journal of Innovative Research in Electronics and Communications (IJIREC)*
*Volume 2, Issue 6, August 2015, PP 1-6*
*ISSN 2349-4042 (Print) & ISSN 2349-4050 (Online)*
*www.arcjournals.org*

# Research on the Pre-processing Methods of Bad Text Information Filter

**ZHU Zhenfang**

School of Information Science and Electric Engineering,
Shandong Jiaotong University,
Jinan, China
*zhuzhfyt@163.com*

**Abstract:** *As the internet becomes more and more popular, a variety of bad information appears on the network such as eroticism, reaction, violence, etc. which seriously disrupts the order of the internet. The text information with bad contents diversifies on the internet. Targeting at the features changes of the emerged sensitive words included in bad text information, this paper puts forward a text pre-processing plan for filtering bad text information and a kind of filter system. In addition, it emphasizes the discussion on the methods for identifying and disposing of sensitive information arisen from the structure change of the sensitive words. At last, the paper presents a three-level text information filter model.*

**Keywords:** *Information security, Sensitive words identification, Text pre-processing, Information filter*

## 1. FOREWORD

According to the statistic report on the internet development in China[1], by December, 2013, the number of web users has grown up to 618 million and the popularization rate of internet has reached 45.8% in China. Except for impelling the development of the social economy, the internet also increases the working and learning efficiency, as well as enriching people's life at spare time. Nowadays, internet has become the largest information base in the world and one of the most important channels for global information transmission. However, because of its openness, the internet is full of junk information, especially some reactionary remarks, superstition, violence and pornography which put serious threats to the network security and lead to spiritual pollution on our life and the development of the society. Such harmful information seriously damages people's physical and psychological health, particularly the growing adolescents.

There are two main causes for bad information[2]:

a) Political purpose: the reactionists publicize and broadcast reactionary information with an intention to topple the regime of the country.

b) Economic purpose: Much bad information like pornography shows up in the forms of pictures or other links to induce the web users to click and login in order to seek more economic interests.

The junk information keeps emerging despite of the repeated bans because junk information creators succeed in getting away from the monitor of the filters through many methods. For instance, use special symbols like "*""[]""、-"、"&" in the bad information to space the sensitive words; Split sensitive words; Replace sensitive words with characters or phonetic spellings.

Thus, how to filter the harmful information and prevent the bad information from spreading on the network becomes an urgent task at present. This paper puts forward the pre-processing design plan to filter the bad text information, identification and treatment methods for the features of sensitive words, as well as a three-level text information filter system.

## 2. RESEARCHES ON BAD INFORMATION FILTER

The term "information filter" occurred for the first time in 1982 at the Communications of the ACM written by Belkin NJ[2], President of ACM. Mr. Belkin NJ used the word "electronic junk" to remind the researchers: the research shall consider not only the automatic generation and spread paths of the

electronic texts, but also the way to control the receipt of the information which means "information filter".

One of the important applications of information filter is to filter the bad information on the internet. The researches on information filter concentrate on two aspects: filter bad information and filter to obtain pertinent information. The purpose of the former is to clean internet environment and guarantee the health of the network information; while the latter aims to obtain information which is closely connected with the users' demands.

Bad information filter distinguishes itself from general information filter in many aspects. Hereafter is the analysis on their features from two aspects:

(1) Orientation judgment: It is difficult to judge the orientation of general text; as a result, both kinds of texts which appeal and not appeal to the users will be obtained at the same time after information filter. However, in the process of bad information filter, the negative texts are less and difficult to be judged as the positive texts are much easier to be obtained.

(2) Expression: the expression of general information filter is stable and convenient to collect the statistics of keywords and words frequency. While the creators of bad information always change the expressions to bypass the bad information filter system consequently enhances the difficulties of bad information filter.

The information filter consists of the following steps: pre-process text (text collection), extract features, and match text with users' demands, etc. Chart 1 shows the model of information filter:

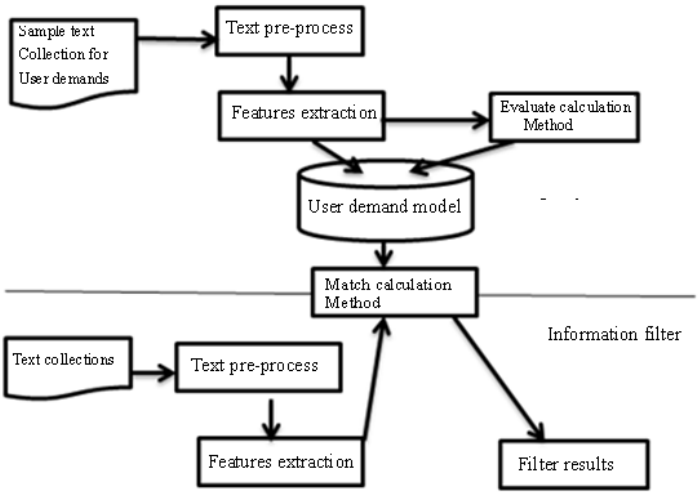


**Chart 1.** *General Model of Information Filter*

Basic process of bad information filter[3]:

(3) Collect the samples of bad text information and make pre-process;

(4) Calculate the weight of the text and set up features dictionary;

(5) Based on the calculated eigenvalue of the two kinds of texts, one meet the filtering requirements, the other doesn't, set up the threshold for the filtering module;

(6) Determine the applicable process method through judging whether the eigenvalue exceeds the threshold or not.

Although the researches on information filter techniques in China started later than other countries [4, 5], the development was quite fast. At present, the researches on the basic theories of information filter and their applications have been carried out in many research institutes and universities. With the progress of the information filter theories and techniques, a large number of information filter products and application systems were developed and achieved certain success. For instance, Xi'an Jiaotong University has developed NIFS system (Network Information Filter System); Ms. Li Dongyan of Shanxi University presented a content filter method based on rules, Netmatron and security filter system for email contents, etc. The author of this paper has done a lot of works in this field [6-8].

## 3. TEXT PRE-PROCESSING

### 3.1 Importance and Necessity of Text Pre-processing

As indicated in the basic process of information filter, the key step of information filter is to pre-process the texts. As the source of information, internet contains an abundance of semi-structured text which are made up by the information which is irrelevant to the text content such as the page headers, break, typesetting codes, tables, pictures, music and animations, script descriptions of the webpage and characters connected with other webpage, etc.

As for information processing, the main expression of the text is VSM (Vector Space Model) which uses the vector to indicate texts. In VSM, the text is represented as (W1W2W3…Wn), among which, Wi is the weight of the No.i term and n refers to the dimension of the term. Therefore, the quality of separating the words of the subsequent text and extracting the eigenvalue is affected by the reserved irrelevant information. Likewise, pre-processing quality also influences the extraction of information, feature expressions and filtering efficiency.

### 3.2 Text Pre-processing Process

As for the collected information, the main pre-processing task is to format the text. But it also includes the process of obtaining VSM, which means separating the words of the extracted text with POS tagging dictionary.

The general process for text pre-processing is shown in Chart 2:

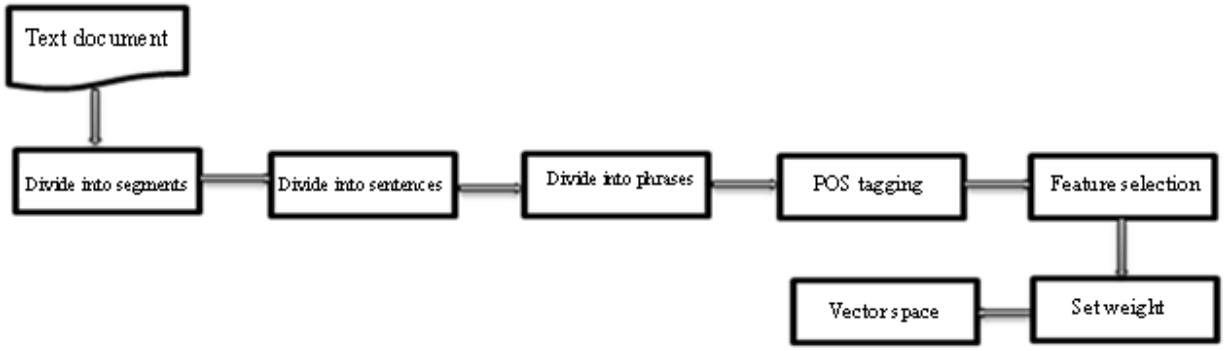


**Chart 2.** *General Process for Text Pre-processing*

(1) Delete the space at the beginning and the end, space between segments and others of the input text

(2) Identify the line break to divide segments; use segment mark to tag the segments.

(3) Identify the sentence breaking symbols like ";" "!" "?" and tag the sentences with sentence breaking symbols.

(4) Divide the sentences into phrases and make POS tags.

(5) After that, choose out feature words from the separated words by designated strategy. That means feature words identification.

(6) Set up the weight of feature words.

(7) At last, build up the VSM of the text.

### 3.3 Process Irrelevant Text Information

At present, a lot of information on the internet is written by HTML (Hyper Text Markup Language). As for the non-text information in the semi-structured text written with HTML which is convenient for reading but unable to be understood by the computer, for example, the page headers, typesetting codes, tables, pictures, music and other irrelevant information, as well as the tags on the webpage, the common treatment method is to delete them and preserve only the useful text information.

## 4. IDENTIFICATION AND PROCESSING METHODS FOR BAD TEXT INFORMATION

Although keywords filter has big limitations, it is a very good way to roughly filter the text. Particularly when realizing filtering on the conceptual network needs very large and complicated calculations, rough filtering through keywords can greatly reduce the quantity of calculations and make the filtering system work better.

### 4.1 Special Symbols are used to Separate Sensitive Words

While writing Chinese texts, common, period, semicolon and other symbols are generally used to break the sentences. Identifying whether the contents in the text are legitimate or not can be realized through checking whether the text contains special symbols like "@" "#" "&" "[]", etc. We can find out all the special symbols in the text and delete them by checking loop statement. In this way, the former sensitive words which are separated by special symbols will reunite, for instance "po*hai-" (persecute), "xiu*lian*zhe" (practitioner), etc. In the actual applications, special symbols may have other functions, like when we calculate the total price, total price=unit price*quantity. In this case, deleting the special symbols doesn't change the nature of the text.

### 4.2 Phonetic spelling is used to replace the characters of the keywords

To judge whether the information in the text is legitimate or not, the next step is to see whether the information is replaced by Phonetic spelling or other symbols. Many texts with bad information replace part of the sensitive words with Phonetic spelling, for example, "法轮功" are represented by "f*a 轮 G*o*n*g". If there are only a few such expressions, the user can input the Phonetic spelling of the collected words into the database in training phase for matching. Dictionary can be used to help find out the characters which match with the Phonetic spelling.

(1) When the phonetic spelling occurred is for one character, use dictionary to find out the corresponding character which can form a phrase with the adjacent character; make a check list for the Phonetic spelling and its corresponding characters and count how many times the Phonetic spelling occurs in the text.

(2) When the phonetic spelling occurred is for more than one character, use dictionary to directly find out the possible corresponding phrase and calculate the total occurrence frequency of the corresponding phrase so as to sort out the feature terms in the text.

### 4.3 Sensitive Characters are split into radicals and incomplete components

In order to avoid being identified by the filter software, some lawbreakers split the characters of the keywords into several radicals. For instance, split "的" two components "白勺", split "勃" into "孛力". This is a compound of the radical and simple character. Another type is to split it into a compound of two radicals, like splitting "功" into "工力", splitting "难" into "又隹".We can use the lexicon to make matching tests for the adjacent characters; then to adjust the probability and eigenvalue of a designated character in the text.

Identify bad information through checking whether there are radicals or incomplete components occurring in the text. The identifying calculation is shown as follow:

(1) First of all, find out the radicals and incomplete components of the characters in the text; then determine whether the character on its right is radical or not.

(2) If the character formed by the combination of the left radical and the right radical exists in the dictionary, use the word segmentation method to calculate the possibility of the character to form a phrase with the characters next to it. Count how many times the phrase occurs in the text while establishing the vocabulary. Otherwise, jump to step 3.

(3) If there is no corresponding character, the possibility of being a transformation of the sensitive words can be eliminated. Move to step 1 to check whether there is other radicals in the subsequent sections.

### 4.4 Function words

Delete those phrases which have no actual meaning, for instance, name of a person, name of a place and some certain phrases. Then delete the stop words which refer to those phrases frequently occur but have no actual meaning. More than a thousand of stop words were collected, hereafter in table 1 are some examples:

**Table 1.** *Stop Words Vocabulary (Partial)*

| People | Not only | Besides | Although | And its | Among |
|---|---|---|---|---|---|
| Unlike | Not only | As for | Begin | Through | Because |
| Therefore | As long as | As | According to | Rest of | You |
| In general | Apart from | On the contrary | …… | | |

### 4.5 Other Forms

At present, judgment on wrongly written characters cannot reach 100% correct. Such phenomenon also exists in bad information texts, for instance use "珐" to replace "法".

As for the judgment on homonyms and synonyms, the corresponding context of the phrase has to be considered. In addition, make matching text with the help of professional synonym dictionary, homonym dictionary, hyponymy dictionary and ambiguous words description base.

In accordance with the transformation features of the bad text information, the process of identifying bad text information is shown in chart 3 as below:

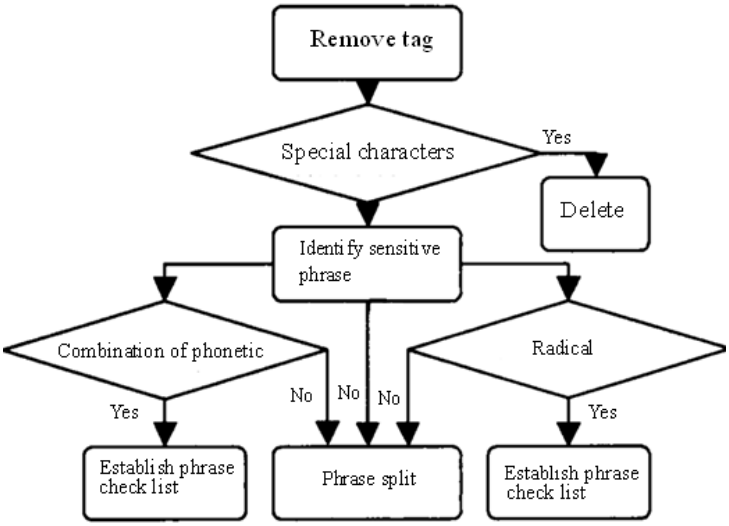


**Chart 3.** *Process for Identifying Sensitive Character (phrase)*

## 5. CONCLUSION

Information filter technique is a kind of data processing technique which has strong vitality and bright future. That's why it attracted wide attentions. Researchers regard it as a significant topic; business circle treat it as an important technology which can bring huge profits. Its applications become wider and wider. This paper discusses the text pre-processing method and three-level text information filter system according to the features of the bad text information. However, there is still much to do before solving the junk information. Thus, further studies on the bad text information filter are needed.

## ACKNOWLEDGMENTS

This work was supported by This work is supported by Shangdong Province Young and Middle-Aged Scientists Research Awards Fund(BS2013DX033), ,Science Foundation of Ministry of Education of China(14YJC860042)National Nature Science Foundation of China (61373148), Nature Science Foundation of Shandong Province (ZR2012FM038), Foundation of Shandong Jiaotong University(Z201302). The authors also gratefully acknowledged the helpful comments and suggestions of the reviewers, which has improved the presentation.

## REFERENCES

[1]. China Internet network information center (CNNIC), China Internet development statistics report [R]. Beijing: China Internet network information center, 2014

[2]. Belkin NJ , Croft WB. Information filtering and information retrieval : Two sides of the same coin. Communications of the ACM , 1992 , 35(12) : 29-38.

[3]. LEE P Y, HUI S C. An intelligent categorization engine for bilingual Web content filtering [J]. IEEE Trans on Multimedia, 2005, 7(6):1183-1190.

[4]. CongJian. Bad information filtering technology research [D]. Beijing: Beijing University of posts and telecommunications, 2012.

[5]. Huang Xuan qingsong, Xia Yingju li-de wu. The text filtering system based on vector space model [J]. Journal of software, 2003, 14 (3) : 435-442.

[6]. ZHU Zhenfang , LIU Peiyu. Feasibility research of text information filtering based on genetic algorithm , Scientific Research and Essays , 2010,5(22):3405- 34104.

[7]. ZHU Zhenfang , LIU Peiyu , WANG Jinlong , ZHENG Yan. Hybrid filtering model based on particle swarm optimization and genetic algorithm, International Journal of the Physical Sciences , 2011, 6(14):3518-3523.

[8]. ZHU Zhenfang , LIU Peiyu , LV Taoxia. Research of feature weights adjustment based on Semantic paragraphs matching , Journal of research and surveys on Innovative Computing, Information and Control , 2010,4(2):559-564 , 2010.

## AUTHOR'S BIOGRAPHY

**ZHU Zhenfang,** PhD, lecturer, he was born in 1980, Linyi City, Shandong Province. He obtained Ph.D. in management engineering and industrial engineering at the Shandong Normal University in 2012, his main research fields including the security of network information, network information filtering, information processing etc.. The authors present the lecturer at the Shandong Jiaotong University, published more than 30 papers over the year.