A Method of Opinion Tendency Analyzing Based on Regular Expression

WANG Zhijuan
College of Information Engineering
Minzu University of China
Beijing, China
e-mail: wangzj_muc@126.com

WANG Zhijuan
Minority Languages Branch
National Language Resource Monitoring & Research Center
Beijing, China
e-mail: wangzj_muc@126.com

***Abstract*—Negative Internet information is harmful for social stability and national unity. Opinion tendency analyzing can find the negative Internet information. Here, a method based on regular expression is introduces that needn't complex technologies about semantics. This method includes: building negative information bank, designing regular expression and the realization of program. The result gotten from this method verified it works perfect on judging the opinion of the web pages.**

Keywords-opinion tendency, negative opinion, regular expression, web page

I. INTRODUCTION

Internet is like a double-edged sword. When it provides valuable information, it brings much harmful information also. Therefore, it is necessary to analysis opinion tendency of Internet information, find the web pages that their opinion are negative, and to prevent these negative information from giving people, especially young people, impassive impact on their lives and affecting stability of social.

This paper introduces several popular methods to analysis opinion tendency of firstly. Then, the analyzing method based on regular expression is introduced, including its concept design and realization. Finally, the test result verified that this method works perfect on analyzing opinion tendency of Internet information.

II. THE METHODS OF OPINION TENDENCY ANALYZING

There are mainly four methods which used to analyzing opinion tendency. Their overviews will be introduced in following.

A. Artificial Analyzing

Artificial analyzing needs full-time staffs check the web pages and found the web pages those contain the negative information. Semantic meaning shall be considered in this method. Although it has high accuracy， its efficiency is too low to be used in analyzing massive information.

B. Analyzing based on simple key words

In this method, some key words are selected to identify negative opinion of web page. These key words are used to find whether there is some information matching the key words or not automatically. This method does not judge the real meaning according to the context in sentence, paragraph or even the whole document. Its advantages lie in that its principle is simple and it is easy to realize. But it can't reflect the weight of these key words in negative opinion and does not consider the semantic meaning. So its accuracy is low.

C. Analyzing based on syntax rules

This method needs the support of Natural Language Understanding (NLU), Natural Language Processing (NLP), Artificial Intelligence, and Data Mining. According context of web pages, it analyzes and understands the opinion of text, and find whether negative information is contained or not. Word segmentation, pos tagging, semantic frame as well as algorithms including Vector Space Model method (VSM), Latent Semantic Indexing method (LSI), Neural Network, Rules Algorithm, Rough Set method and so on are need to be considered in this method. Its theoretical basis is clear and theory accuracy is high. But it is difficult to realize thus its real accuracy is low.

D. Analyzing based on regular expression

Regular expression is good at describing and matching strings. According to the features of negative information such as syntax, fixed usage, special symbol and so on, the regular expressions can be designed. Using the regular expression models that reflect negative opinion, the web pages that contain negative information can be found easily. So, this method easy to realize and has better accuracy.

Among four methods, the method based on regular expression is actually realizable and has better accuracy expect artificial analyzing. Therefore, the method of using regular expression to analyze the opinion tendency on web pages is introduced in this paper.

III. THE CONCEPT DESIGN OF OPINION TENDENCY ANALYZING BASED ON REGUALR EXPRESSION

The method to analyze opinion tendency of web pages is designed as what is shown in Figure 1.

Firstly, an information bank that reflects negative opinion is build according to corpus. Secondly, certain regular expressions for specific negative information and rules defined by users are designed. Thirdly, convert the format of web pages downed from Internet from HTML to XML. Finally, the xml files are read and judged according to regular expression to check whether there are strings to match the regular expressions or not . If yes, the web page

Published by Atlantis Press, Paris, France.

is considered to be a page that contains negative information and its opinion is negative.

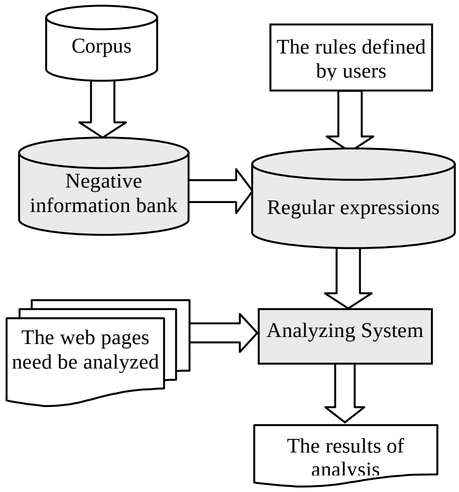
Corpus
The rules defined by users
Negative information bank
Regular expressions
The web pages need be analyzed
Analyzing System
The results of analysis

Figure 1. The structure of opinion tendency analyzing system based on regular expression

A. Build information bank that reflects negative opinion

In order to analyze the opinion tendency of web pages, an information bank that reflects negative opinion should be constructed. For identifying the features of different opinion tendency, the information bank should be designed in different category as shown in Table Ⅰ.

TABLE I. THE CATEGORIES OF NEGATIVE INFORMATION BANK

Num	Category	For example
1	The name of certain people	Li honhzhi, Bin Laden
2	The name of certain organization	Falungong
3	Certain event	9 • 11
4	Others	Violent incident, anti-government, genocide…

As shown in Table Ⅱ,for same words, different opinion can be expressed using different syntax, fixed usage, and special symbol.

TABLE II. THE OPINION TENDENCY FOR SAME INFORMATION

Num	Category	Negative Opinion	Positive Opinion
1	The name of certain people. For example AB	May be AB	Most be Professor AB, Dr. AB, Mr. AB
2	The name of certain organization. For example XYZ	Most be "XYZ"	May be XYZ
3	…	…	…

In order to escaping monitoring, some information that reflects negative opinion often appears in some special variations as shown in Table Ⅲ. We must consider these variations of negative information when certain regular expression is designed.

TABLE III. THE VARIATION EXAMPLE OF NEGATIVE INFORMATION

<table>
<tr><th>Negative information</th><th>Variation</th></tr>
<tr><td>People's name
Such as AB</td><td>A*B, A^_B,
^A^B, …</td></tr>
<tr><td>The name of certain organization
Such as XYZ</td><td>X|Y|Z, X*Y-Z,
X~Y~Z~, …</td></tr>
</table>

B. Design regular expression

Regular expression consists of a series of ordinary characters (such as 'a' to 'z' and so on) and some special characters (such as '*', '|', '?' and so on). It can find opinion tendency by several certain strings with syntax, fixed usage, and special symbol. For example, (abc|def).*xyz means a string begin with abc (or) def and end with xyz. So it can describe some semantics.

Here, regular expressions are used to identify the opinion tendency of web pages. According the front analysis, the regular expression should be designed as following:

(1) Using fixed usage

Some opinion can be expressed by fixed usage. For example, honorific usually is express opinion. If name of negative people with honorific, the opinion is negative. If the names of state leaders without any address, the opinion is negative too.

(2) Using special symbol

Some punctuation symbols also express opinion tendency. For example, if the name of organization is XYZ. "XYZ" often used in negative position.

(3) Using special variations

Special variations often used to escaping monitoring. So if the web page contains the special variations, its opinion usually is negative.

Some example of regular expression are shown in Table Ⅳ.

TABLE IV. THE EXAMPLE OF REGULAR EXPRESSION MODELS

<table>
<tr><th></th><th>Category</th><th>Negative Opinion</th><th>Positive Opinion</th></tr>
<tr><td>fixed usage</td><td>The name of negative people (such as AB)</td><td>((AB)+(\W|\w)*?(sage))|((Mr. AB)+)|((Dr. AB)+)</td><td>AB</td></tr>
<tr><td></td><td>The names of state leaders (such as CD)</td><td>CD</td><td>((CD)+(\W|\w)*?(sage))|((Mr. CD)+)|((Dr. CD)+)</td></tr>
<tr><td>special symbol</td><td>Organization's name (such as XYZ)</td><td>((?<!")(XYZ)(?<!"))+)|(((?<!")(XYZ)(?<!"))+)</td><td>XYZ</td></tr>
</table>

C. The program flow chart

The program flow chart of system is shown Figure 2. Firstly, set the regular expressions, introduced in part B, used in the system. Then, read the web pages and use the one of the regular expression to check the text. If matched, the web page contains the certain negative information and check next text. If not match, select next match regular expression to check again until all web pages are judged or all regular expressions are used. This process continues until all web pages are red and judged.

Published by Atlantis Press, Paris, France.

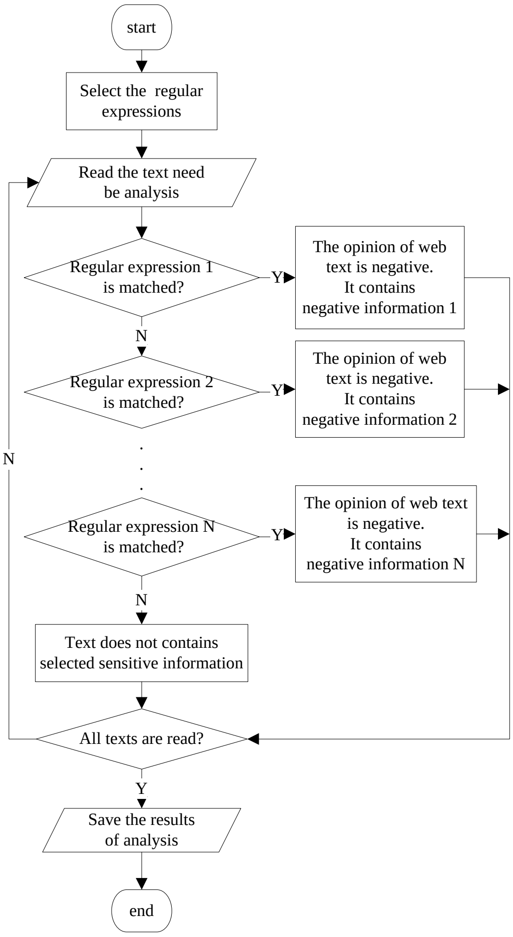
start
Select the regular expressions
Read the text need be analysis
Regular expression 1 is matched?
Y
The opinion of web text is negative. It contains negative information 1
N
Regular expression 2 is matched?
Y
The opinion of web text is negative. It contains negative information 2
.
.
.
Regular expression N is matched?
Y
The opinion of web text is negative. It contains negative information N
N
Text does not contains selected sensitive information
N
All texts are read?
Y
Save the results of analysis
end

Figure 2. The program flow chart of system

IV. REALIZATION AND TEST

A. Key codes

The key codes of analyzing system based on regular expression are following.

```
import java.util.regex.Matcher;
import java.util.regex.Pattern;
public class Match1
{
   public static void main(String[] args)
   {
      String str = "the string need to be matched" ;
                                   // the string need to be matched
      String pat = "regular expression model" ;
                                   // regular expression model
      Pattern p = Pattern.compile(pat) ;
      Matcher m = p.matcher(str) ;
     if(m.matches())
      {
         System.out.println("there is( or are) negative
 information!" + "\n the content :： " + str);
        }
      else
         {
          System.out.println("there is not negative
information!");
         }
      }
}
```

B. Test

Figure 3 is the analyzing result of Tibetan Internet information. Here, regular expression is about Tibet Independence. And 50 web pages are used, 20 web pages are positive and 30 web pages are negative. The result verifies that 26 web pages within the 30 web pages are found as negative web pages. The test verified that this method works well on judging the opinion tendency of web pages.

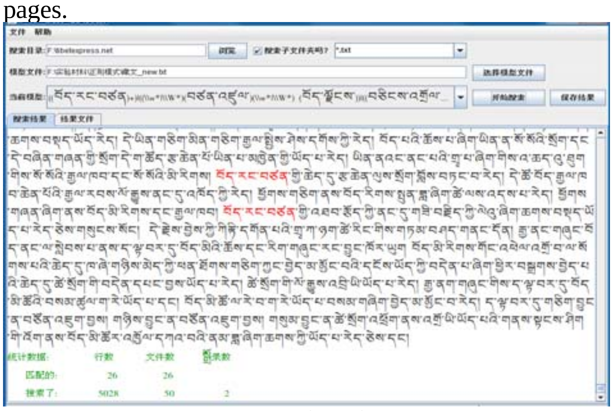

Figure 3. The interface of analyzing test

V. CONCLUSION

Network information security is important for the youths and society stability. So it is necessary to analysis the opinion of web pages. The analysis method based on regular expression has simple principle and is actually realizable. The test result verified that this method works perfect on judging the opinion tendency of web page.

VI. ACKNOWLEDGMENT

This work is supported by National social science fund (No: 11CY016), National Science and Technology Supporting Item (No: 2009BAH41B04), the State Nationalities Affairs Commission fund (No: 10ZY07) and open project of National Language Resource Monitoring & Research Center (No: NMLR201104)

REFERENCES

[1] William W Cohen.Leanring Rules that Classify Email.AAAI Spring Simposium on Machine learning in information Access.1996，96(5):18-25.

[2] Robert Cooley.Pang-NingTan.Jaideep Srivastava.WebSIFT:The web Site Information Filter System.In Proceedings of the workshop On webb Usage Analysis and Use rProfiling(WebKKD99)， Sna Diego，1999:45-57.

[3] TrinataPhyllou Evangelos.A L Soyster and S R T Kumara.Generating Logiacl exPressions from Positive and negative exmaples via a Brnaeh-and-bound approaeh.ComPuters and Opeartions Reseaerh，1994，21(2):185-19P.

Published by Atlantis Press, Paris, France.

[4] DeshPande A S.Trianatphyllu Evnagelos.A greedy randomized Adaptive search Procedure(GRASP) for inference logical clauses from Examples in polynomial time and some extensions.Mathematical and ComPuter Modelling ， 1998 ， 27(1):75-99.

[5] Salvador Nietosanehez.TriantaPhyllu Evangelos.Donald Kraft.A Feature mining based apporach for the classifiaction of text documents Into disjoint classes Inoformation.Porcessing and Hamagement，2002，38(4):583-604.

[6] SU Gui-yang ， L I Jian-hua ， MA Ying-hua ， L Isheng-hong.Improving the precision of the keyword matching pornographic text filtering method using a hybrid model[J] . Zhe-jiang Univ SCI 2004， 5 (9):1106 - 1113.

[7] Ricardo B. Y，Gonzalo N. New and faster filters or multiple approximate string matching Random Structures and Algorithms，2002 ，20 (1) :23-24.

[8] Wai H. Ho ， Paul A. Watters. Statistical and Structural Approaches to Filtering Internet Pornography[J] . 2004 IEEE International Conference on Systems ，Man and Cybernetics ，4792 - 4798.

[9] Pei Chunbaodan, Zeng Basang.The analysis and resolution of tibten processing in JAVAprogramming[J]. Tibet s Science and Technology，2008(10):68-70.(in Chinese)