The Chinese Spam Keyword Filtering based-on the Maximal Independent Set

HaiLong Wang[1] FanJun Meng[1] HaiPeng Jia[2] JinHong Cheng[3] Jiong Xie[3]

[1]Inner Mongolia Normal University,ZhaoWuDa Road No.81,Hohhot,China
[2] Air Defence Forces Academy
[3] Inner Mongolia electric power information and Communication Center
lzjtuwhl@163.com

Abstract. This paper presents a Chinese e-mail keyword filtering algorithm based on the maximal independent set, build a string-matching relation matrix helps us to improve the performance of maximal independent set for matching relation matrix. In addition, we developed a judgmental criterion according to the algorithm. We also design a behavior recognition technology, which can detect and reject the email which receiving.

Keywords. Maximal independent set; semi-diagonal line; matching relation matrix

1 Introduction

E-mail has become the important method for network communication as it is widely used among the Internet users and is regarded as one of the most commonly used network applications. However, with the development of Internet, junk e-mails (spam) bothering most people do not only bring discontent to users, but also cause some web security issues and economic losses.

Negative effects derived from spam, which bring great economic losses and result in large amounts of data and information blockage, have become a worldwide problem. Numerous experts and scholars put forward a lot of targeted prevention methods that ease the problem to some extent. Data Mining is a relatively popular technique for the filtering of e-mail contents and theme keywords, which detect spam keywords through keyword classification and statistical algorithm. Bayes filter is an effective method. The characteristics of Bayes filter are adaptation and self-learning. Bayes filter has the advantages of high detection accuracy[1]. Other widely used detecting approaches include detection based on memorial information, detection based on description of event features, and filtering based on spam feature analysis and regular expression matches.

2 the Keywords Filtering Algorithm Based on the Maximal Independent Set

2.1 Matching relation matrix of string

Given any two strings S and T, the maximum matching problem of them is equivalent to the maximal independent set of matching relation matrix[2].Define $S = a_1 a_2 \cdots a_m, T = b_1 b_2 \cdots b_n$ .Note that $< n >= \{1,\ 2, \cdots, n\}, < m >= \{1, 2, \cdots, m\}$, thus $\{(i, j) \mid i \in < m > ,\ j \in < n > ,\ a_i = b_j\}$ is called matching relation set of S and T , written $M(S,T)$,here we assume that $n \geq m$ generally[2]. C matching relation matrix, can be defined as follow:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m1} & \cdots & c_{mn} \end{bmatrix} \quad \textbf{(1)}$$

Where $c_{ij} = \begin{cases} 1 & ,if \quad a_i = b_j \\ 0 & ,if \quad a_i \neq b_j \end{cases}$

We only discuss the situation that weight $c_{ij} = 1$ in this paper.

Definition This thesis defines that each node in the independent set only exist a corresponding node at the bottom right located in the different row or different column, called quasi diagonal[3].

The set of nodes which value are 1 ($c_{ij} = 1$) over a quasi diagonal of matching relation matrix is called an independent set. The keyword matching problem can be transformed into sloving maximal independent set of matching relation matrix, and searching all independent sets in a given matching relation matrix. So the longest set is the answer[4]. In particular, we can search points in matching relation matrix that of which value is 1 to determine whether they are completely match in the process. However, the idea discuss above will make the problem more complicated, we find that finding the maximal independent sets can be regarded as searching for a road in the matching relation matrix of which value is 1. Each node in the independent set only exist a corresponding node at the bottom right located in the different row or different column, called quasi diagonal, search the bottom right according to the quasi diagonal.

The relationship between output result with original input string will satisfy the following relationship:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{2}$$

2.2 The algorithm of maximal independent set of matching relation matrix

In order to search the bottom right corresponding with the quasi diagonal, we propose an improved algorithm of maximal independent set of matching relation matrix as follows:

Assume string α and β are independent sets obtained in searching, we set the length of them as |α| and |β|.

The procedure of algorithm describes as follows:

Step 1. When search the column j, if |α| ≥ |β|, and the abscissa of the last character of α i1 < i2(the abscissa of the last character of β), then stop the operation of β.

Step 2. When search the row i, if |α| ≥ |β|, and the ordinate of the last character of α j1 < j2 (the ordinate of the last character of β), then stop the operation of β.

Step 3. When searching matching relation matrix is completed, the length of α is equal to the original string, then keywords are found[5].

In this algorithm, it may generate multiple result of β, because we just find the length of α with the length of original string, so we can get multiple independent sets in the end. The finding procedure will be present as a pseudo code in following:

```
search
  d[]=0; col=0; η[]=0; k=0;
  for j∈col,…,n do
    for i∈k,…, m do
    if equal(S[i],T[j])
      { d[]++; col++; k++;  continue; }
    if(j<n&&k>=m)
      k=0;
  End of For
    η[] = d[]/m;
  End
```

Where d[] indicates the array of length of each independent set during the storage and calculation, η[] presents the matching accuracy of each independent set, col is the next matching start position of target string T[j], This setting design of col can largely reduce the matching time complexity. K denotes the identity of search the original string. When the first searching trip of the original string is finished and the target string is not completed, and then search from the first character of the original string again, the search process will stop until it finds all the matching string[6].

2.3 Judgement Criterion

The Matching accuracy $\eta = \frac{N\alpha}{N_c}$, where N_α is the length of original keywords string, $N_c = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}$ is the length of all quasi diagonal(Cij only equal 0 or 1). If η<1，indicates the string α and the detecting target string does not match exactly, then the system outputs the results: the mail is secure; if η>1,meaningless; if η＝1, then the system shows the string α and the detecting target string match exactly, keywords hidden in the string β is found, then the system give a warning, and continue the next steps.

3 Conclusions

This paper presents a set of mail keywords filtering methods to find the maximal independent set. We design and implement an adopted algorithm which can effectively solve the problems including keywords split and combine, as well as inserting special symbols. We also design a behavior recognition technology, which can detect and reject the email which receiving. The experimental result shows that both space and time complexity are far less than O(mn), the efficiency is also satisfactory.

References

1. Ze Li; Haiying Shen . : SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box . INFOCOM, 2011 Proceedings IEEE . Page(s): 1835 – 1843(2011)
2. Huilin-Yuan; Dingwei-Wang. : The New Approach of Marking Activity-Loops Based on the String Reachable Matrix. Communications and Mobile Computing, 2009. CMC '09. WRI International Conference on . Page(s): 569 – 572(2009)
3. Aun, M.T.B.; Bok-Min Goi; Kim, V.T.H. : Cloud enabled spam filtering services: Challenges and opportunities. Sustainable Utilization and Development in Engineering and Technology (STUDENT), 2011 IEEE Conference on . Page(s): 63 – 68(2011)
4. Luo, Qin; Liu, Bin; Yan, Junhua; He, Zhongyue. : Design and Implement a Rule-Based Spam Filtering System Using Neural Network . Computational and Information Sciences (ICCIS), 2011 International Conference on . Page(s): 398 – 401(2011)
5. Li Aiwu; Liu Hongying . : Utilizing improved Bayesian algorithm to identify blog comment spam . Robotics and Applications (ISRA), 2012 IEEE Symposium on . Page(s): 423 – 426(2012)
6. Ji-Cherng Lin; Huang, T.C.: An efficient fault-containing self-stabilizing algorithm for finding a maximal independent set . Parallel and Distributed Systems, IEEE Transactions on . Page(s): 742 – 754(2003)