International Journal of Security and Its Applications
Vol. 10, No. 7 (2016) pp.363-372
http://dx.doi.org/10.14257/ijsia.2016.10.7.32

Web Spam Detection Based On Link Diversity and Content Features

Xu Gongwen[a], Li Xiaomei[b], Zhang Zhijun[a*]and Xu Li'Na[a]

[a] *School of Computer Science and Technology Shandong Jianzhu University*
[b] *Cancer Center of the Second Hospital Shandong University*
zhangzj@sdjzu.edu.cn

Abstract

In order to get a higher ranking, spam pages deceive the search engine using cheating technology, which will disturb the users to find useful information via search engine. The web spam is designed for search engines rather than for users, so it is important to make a distinction between the normal web pages and the web spam pages. The links of the normal web pages have a wide variety of sources and the content feature of the normal web pages are distributed regularly, while links source of the web spam pages is single and the content features of them are distributed disorderly. So after analyzing the link diversity and content features distribution of the web pages, a new web page ranking algorithm was proposed in this paper. In this method, the web pages ranking score is calculated by the TrustRank method combining web pages links diversity and the web pages content features. It can be shown from the experimental results that this method can effectively reduce spam pages ranking score.

Keywords*: web spam, normal web pages, link diversity, content features, TrustRank, PageRank.*

1. Introduction

Searching engines are one of the main ways for users to find useful information on Internet, and a survey in 2009 showed that [1], 68% of people regularly use search engines, 84.5% of people get new information by using the search engine. Since most users are only interesting in the search engine returns results in the first three pages [2], in order to get higher rankings in search engine results, the site managers will work through search engine optimization (SEO) technology [3] to improve their page quality. SEO technology make the web site more suitable for search engines to retrieve principle through the properly designed web content and links, so the web site can obtain a higher ranking when using the search engine to gain more profit. However, driven by commercial interests, some web sites use improper search engine optimization techniques to deceive search engines in order to obtain a higher ranking, and these pages are called web spam page [4, 5]. Web spam not only affects the user to find useful information through a search engine, but also wastes resources of search engine seriously. When search engine indexes web pages according to user's request, it has to deal with a lot of web spam pages. So for the web search engines, it is one of the top challenges to combat with web spam[6].

At present, web spam pages are divided into three types according to the cheating methods: the first one is the cheating method based web content. Content-based Web spam cheating method means to modify web content such as accumulating a large number of search keywords, adding the content to web pages maliciously to improve search engine ranking results. Referring to this type of web spam pages, most studies

[*] Corresponding Author

ISSN: 1738-9976 IJSIA

detected web spam by extracting, analyzing the content and features of normal web pages. For example, Urvoy [7] identified spam web pages according to the similarity comparison of the web pages style; Cafarella [8] analyzed the distribution of the page keywords by applying the statistical method, and then detected spam pages based on the distribution of the keywords; Ntoulas *etc*. [9] constructed the classifier by extracting, analyzing web content features to detect web spam. The second cheating method bases on website pages link structure. Spam page deceive search engine ranking algorithms by adding extra links or misleading other links to direct to it. The most common technique is the link farms, which creates a lot of web spam pages which point to the special target pages. Because the target pages have a large amount of inbound links, they can get a higher ranking through the sorting algorithm. The study of Adali [10] showed that all the web spam pages directing to a target page is the most effective link cheating method. Many studies detected link-based spam pages through web links structure graph. PageRank algorithm [11], which ranks pages by the link's contribution value between web pages, is one of the famous ranking algorithms based on web link graph.

In the PageRank algorithm, $o(p)$ represents the number of outgoing links to the web page p. The PageRank value of page p can be obtained by the following formula:

$$p(p) = \alpha \sum_{(p,q)} \frac{p(q)}{o(q)} + (1-\alpha)\frac{1}{N} \tag{1}$$

Where α is an attenuation coefficient, and (p,q) means there is a link between page q and p. N is the number of all web pages[11].

Based on trust propagation model and PageRank algorithm, Gyongyi proposed a TrustRank algorithm [12], which assigns a value to each page using the trust propagation method. The web pages are ranked according to the above trust value. There are two steps in the TrustRank algorithm: seed selection and trust propagation. An initial value for pages of seed collection is set by TrustRank algorithm, and the Trustrank value of all pages is defined as below:

$$TR = \beta \times TR \times T + (1-\beta) \times d \tag{2}$$

Where d is the static score distribution vector of seeds collection, T is the matrix calculated based on PageRank algorithm, β is an attenuation coefficient which is set to 0.85 usually.

Using the page trust value, Asano *et. al.,* [13] modified the HIST algorithm to detect the spam pages based on link cheating. Jacob [14] detected the spam page based on the network graph regularization. The third cheating method is hiding technology, which hides web content by setting key words or other webpage content to the same color of the background color, and then the page content “lost” in the same color background.

When calculating a page ranking score, the existing page ranking algorithm based on links structure only thought about two factors, the number of the pages and the quality of the pages that directing to this page. If the sources of a page’s link come from other different pages, it indicates that this page has been recognized widely. Otherwise, if a page’s link comes from a single source, this page may be a spam page, even though it has more voters. Therefore, we analyzed the variety of the pages’ link source and extracted the web page content features based on the TrustRank algorithm. The page ranking score was calculated by combining the link diversity and content features. After that, the high quality web page was given a high page rank as much as possible, leaving the web spam pages on the list.

2. Web Link and Content Features

The links of normal web page usually come from other normal pages, but spam pages usually use link farms and other cheating ways to improve the number of links. Because of the obvious difference between normal web pages' link and spam pages' link, we analyze the information of web pages' link firstly.

2.1. Link Diversity

The cheating method based on link spam usually has two kinds. The first one is that a large number of web spam pages pointed to the target page by link farm to improve the score of the target page. The second way is links exchange cheating. In this way, multiple target web spam pages are created. Besides the target web pages are linked by link farm, they also linked between each other to improve their ranking score. In order to achieve higher score for the target pages, web pages in link farm usually only point to the target page, while the normal web page rarely points to the spam pages, so the link source of the spam pages are relatively simple.

According to the above theory, now the diversity of the web pages link source is analyzed [15]. The link diversity of the web p and q is defined as follows:

$$D(p,q)=1-\frac{|U(p)\bigcap U(q)|}{|U(p)\bigcup U(q)|} \tag{3}$$

Wherein, $U(p)$ represents the k nearest neighbor set of web page p , and $U(q)$ represents the k nearest neighbor set of web page q . $|*|$ represents the number of pages in the set, and k is usually set to 3. As the pages in link farm point to each other, so the diversity of spam pages link is usually lower. The link weight between the link farm pages and target cheating pages is weakened according to the link variety, reaching the purpose of reducing the value of the spam pages. As normal web pages are pointed by other normal web pages and have higher link variety, the link weight of normal web pages is less affected by this operation.

Combining with the diversity of web links, the new weight of link between web pages p and q is calculated as the following formula.

$$D(p,q)=W(p,q)\times D(p,q) \tag{4}$$

Where $W(p,q)$ represent the original weight. The smaller the link diversity between web pages, the greater the penalty for the link weights, the smaller the contribution value of the web page.

2.2. Content Features of Web Pages

As the web spam pages aim to improve search engine ranking score, most of the web content is the accumulation of large amount of search keywords and large irrelevant content maliciously added. So we analyzed the content features of the web pages in the data set UK-2007 provided by Yahoo search engine. After analyzing the contents of the spam pages and normal web pages, some obvious content features were extracted, the number of title words and web pages compression ratio.

2.2.1. The Number of Title Words: When using search engines, we generally input key words to find what we need, so in many spam pages a large amount of keywords which are not related to web contents are put together as a page title, this way can improve matching degree, made spam pages more easily to be searched by search engine. This is the so-called keyword stuffing. We have statistically analyzed the number of title words in spam pages and normal web pages respectively, the results are shown in Figure 1 and Figure 2.

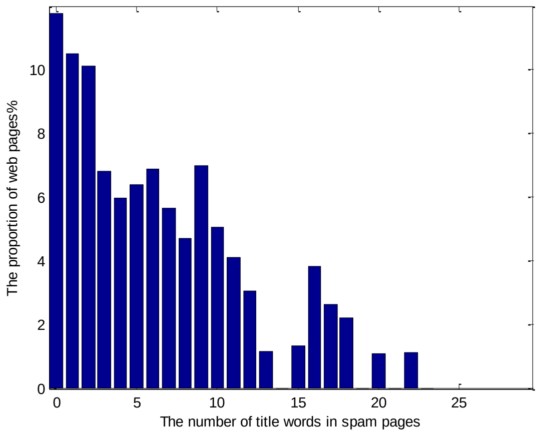

Figure 1. The Distribution of the Number of Title Words in Spam Pages

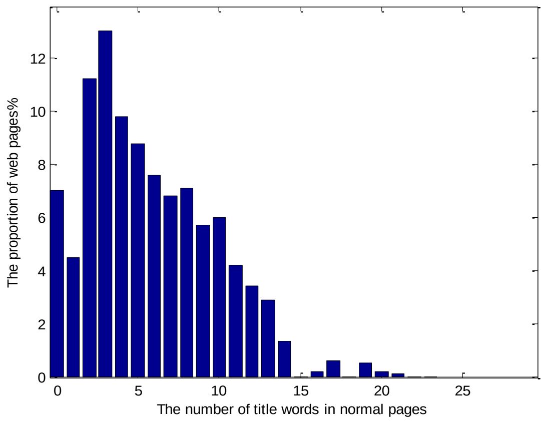

Figure 2. The Distribution of the Number of Title Words in Normal Pages

As can be seen from the Figure 1 and Figure 2, the number of normal pages is 2.92% when the number of title words is more than 15 in the data set. While the spam pages titles are filled with malicious keywords or the repeating target keywords in order to get a higher ranking, the proportion of the pages whose number of title words is more than 15 is as high as 12.08%. It can be seen from the figures that the distribution of normal web pages title length is near the normal distribution. The following formula describes the probability density function of the normal distribution.

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-u)^2}{2\sigma^2}} \tag{5}$$

Where u is the mean of the random variables of normal distribution, σ is the standard deviation of the random variables.

2.2.2. Web Pages Compression Ratio: If same query keywords appear in a web page many times, the search engine will give this page a higher ranking score. When users search information according to this keyword, this web has more chance to achieve a high ranking and be selected by users. For example, for a given query keyword, the one appearing many times in the content of the web page ranks higher than the one appearing only once. In the dataset, the distribution of compression ratio of normal web pages and spam pages are shown in Figure 3 and Figure 4.

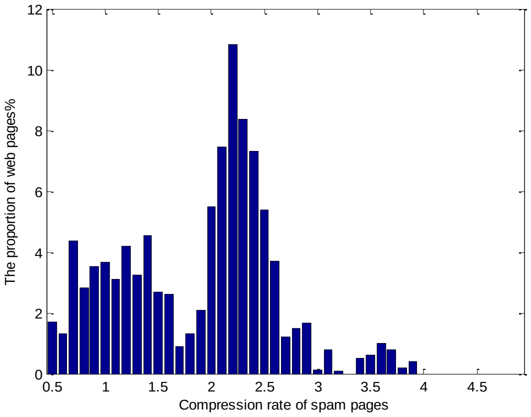

Figure 3. The Distribution of the Compression Rate about Spam Pages

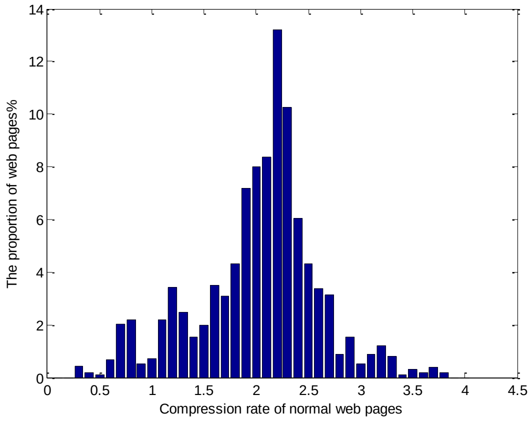

Figure 4. The Distribution of the Compression Rate about Normal Web Pages

As can be seen from Figure 3 and Figure 4, the proportion of normal web pages plummets when the compression rate is more than 2.8, the proportion is only about 6.52%. While the proportion of spam pages is 9.45% when compression rate is more than 2.8, the proportion is much higher than normal pages.

From Figure 4 we can get the information that the distribution of normal web compression rate is approximate Poisson distribution, and the probability distribution of Poisson distribution is as follows.

$$P(X=k)=\delta\frac{\varepsilon^{k}}{k!}e^{-\varepsilon},k=0,1,2,... \qquad (6)$$

Where the parameters ε is larger than 0, the value of k is the integer part of the web compression rate divided by 0.2, and the distribution of normal web compression rate can be well fitted by the parameters $\varepsilon=10,\delta=80$.

We also analyzed the distribution of web page content, such as the number of text words, the proportion of visual text, the proportion of anchor text, and the average word length and so on. After the analysis of the distribution in the normal web pages content and the spam pages content, we found that the distribution of the content in normal web pages had the certain regularity, and the distribution of the contents of the spam pages was not regular.

For different web pages content features, we use approximate distribution function to fit them, and calculate the difference between characteristic value of web content and distribution function. Because the content of normal web pages is in accordance with the law, the difference is small. While the content distribution of the spam pages is out of order, the difference is large. So we can use this difference to distinguish spam pages from the normal web pages.

2.3. Ranking combining link diversity and content features

We define the difference value based on the content features as follows.

$$g(p)=\sum_{i}^{n}\left|P_i(p)-C_i(p)\right| \tag{7}$$

Wherein, the value of $P_i(p)$ is the proportion of pages when the content feature is i, $C_i(p)$ is the approximate distribution function which fits the content feature of web pages properly, n is the number of content features we have chosen. $g(p)$ is the difference between characteristic value of web content and distribution function.

Combining web page link diversity, web content feature distribution and TrustRank algorithm, we calculated the rank score of web pages in the database. Calculation formula of TrustRank algorithm is defined as follow.

$$TR(p)=\alpha\sum_{q:(q,p)}\frac{r(p)}{o(q)}+(1-\alpha)\frac{1}{N} \tag{8}$$

Where $r(p)$ represents the number of incoming links of page p, and $o(p)$ represents the number of outgoing links of page p, α is a decay factor, and $q:(q,p)$ means there exists a link between page p and q, N is the number of the web pages.

The new calculation formula is as follow.

$$r(p)=\alpha\sum_{q:(q,p)\in\varepsilon}D(p,q)\frac{r(q)}{o(q)}+(1-\alpha)\frac{1}{N}-\theta\times g(p) \tag{9}$$

Wherein, $D(p,q)$ represents the weight between web page p and web page q combined with diversity of links. If page p is a normal web page, the weight $D(p,q)$ between web page p and web page q is large, and the value of $g(p)$ is small, so the score value of the web page p will be high. If page p is a web spam page, the weight $D(p,q)$ is small and the value of $g(p)$ is high, web page p will get a low score. As a result of the actual calculation, the value of $g(p)$ is relatively small, in order to effectively penalize the spam pages, we set a weight θ for $g(p)$ and the value of θ is set to 10 here.

3. Experiment and Results

3.1. Dataset

We used the UK-2007 as a data set, which is published by the Yahoo laboratory. In the data base, web pages are manually labeled as three categories: non-spam, spam, undecided. Among them, only the "non-spam" and "spam" are selected as the data set. A total of 5797 pages are labeled in the data set, including 321 spam pages and 5476 normal pages. Since the ratio of spam and non-spam is about 1:17, in this experiment we adjusted the proportion of positive samples and negative samples appropriately.

3.2. Measurement Standard

In order to effectively detect the experimental results, we use the precision, the recall rate and F-measure as the results testing standard.

The number of spam samples which are identified correctly is expressed by TP; the number of spam samples which are identified as non-spam is expressed by TN; the number of non-spam samples which are identified as spam is expressed by FP; the number of non-spam samples which are correctly identified is expressed by FN.

Precision is the proportion of the real spam pages in the forecast spam pages. The higher the precision is, the smaller the probability that normal web pages mistaken as spam pages is. The calculation formula of precision is as follow.

$$\mathrm{Pr}\,ecision = TP / (TP + FP) \tag{10}$$

Recall means the proportion of correctly forecast pages in the real spam pages, the calculation formula of recall is as follow.

$$\mathrm{Re}\,call = TP / (TP + TN) \tag{11}$$

F-measure is the average of the precision rate and recall rate, which integrates the precision and recall as an indicator. The calculation formula of F-measure is as follow.

$$F = 2 \times \frac{\mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{12}$$

3.3. Results

In this section, the validity of the method is verified by the test on the data set. The data set is UK-2007, which is published by the Yahoo laboratory.

We calculated the score of the web pages in four different cases. According to the value which is calculated by different algorithm, web pages are divided into 10 barrels in descending order, and the proportion of spam pages in each barrel is calculated. These four cases are set as follow. The first one, only the TrustRank value of the web page is calculated; The second one, the value of web page combining the TrustRank and link diversity value is calculated; The third one, TrustRank value of the web page is calculated combining the content feature distribution; The last one, TrustRank value of the web page is calculated combined with link diversity and the content feature distribution. The experimental results are shown in Figure 5.

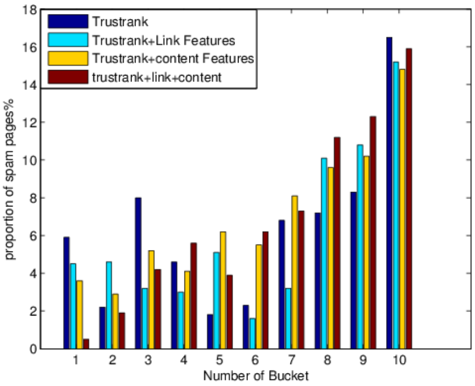

Figure 5. Relationship between Number of Bucket and Proportion of Spam Pages

As can be seen from Figure 5, in ten barrels, each barrel has a certain percentage of the spam pages. The bigger number of the barrels is, the more proportion of the spam pages is. The proportion of spam pages in the first few barrels is much smaller than the proportion of pages in the barrels behind. TrustRank algorithm only considers the link information, so there are still a lot of spam pages in the first three barrels, the proportion is 16.06%. When the TrustRank algorithm combined the link features and the content features respectively, the proportion of the spam pages was reduced to 12.38% and 11.65% in the first three barrels; the proportion of spam pages has increased in the later barrels. When the TrustRank algorithm combined with link diversity and content features, the proportion of the spam pages was reduced to 6.59% in the first three barrels, the proportion of the spam pages increased to 39.39% in the last three barrels. The method combined with link diversity and content features can reduce the ranking of spam pages effectively.

The precision and recall are also used to detect the validity of the proposed method. Each page has a value calculated by our algorithm, and the page is ranked according to the value. In order to evaluate the performance of the algorithm, a threshold is selected. If the ranking value of the web page is greater than the threshold, it is considered to be normal web page. If the value is less than the threshold, it is considered to be a spam page. The results are shown in Figure 6.

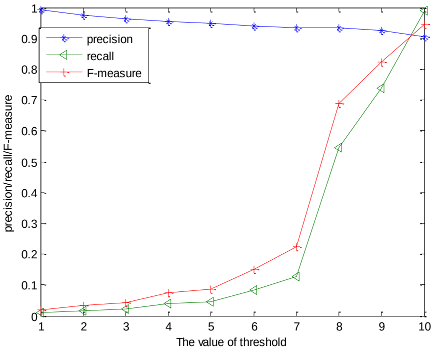

Figure 6. The Results of Measurement Standard

When the 9th barrel is selected as the threshold, the spam pages whose values are less than the boundary value of 9 barrel can be detected. The precision is 92.6%, recall is 73.9% and F-measure is 0.822.

4. Conclusion

In this paper, the diversity of web links and the distribution of the content of the web pages were analyzed. The score of the web page was calculated according to the TrustRank method combined with link diversity and the content feature distribution. The ranking score value of the spam pages is punished by the nature of link diversity and the distribution of the content feature. Experimental results showed that TrustRank calculation method combining the link diversity and the distribution of the content feature could effectively reduce the ranking score of the spam pages, and reduce spam pages ranking in the search engine results accordingly.

Acknowledgements

This work is partially supported by A Project of Shandong Province Higher Educational Science and Technology Programs (J12LN31, J13LN11, J14LN59), the Scientific Research Fund of the Second Hospital of Shandong University (S2015010004), the Development Projects of Science and Technology of Shandong Province (2012GGX27073, 2014GGX101011, 2015GGX101018, 2016GGE27402).

References

[1] CNNIC (China Internet Network Information Center).The 23rd report in development of Internet in China, **(2009)**, pp. 1-3.

[2] C. Silverstein, H. Marais, M. Henzinger and M. Moricz, "Analysis of a very large Web search engine query log", Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(1999)**, California, US.

[3] L. Becchetti, C. Castillo and D. Donato, "Web spam detection: link-based and content-based techniques", In The European Integrated Project Dynamically Evolving, Large Scale Information Systems (DELIS): proceedings of the final workshop, **(2008)** Paderborn, Germany.

[4] Z. Gyongyi and H. Molina, "Web spam taxonomy. Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the web", Chiba, Japan, **(2005)**.

[5] R. Wei, J. Zhen and L. Bao, "Study on Mining Big Users Data in the Development of Hubei Auto-Parts Enterprise", Mathematical Modelling of Engineering Problems, vol. 4, no. 2, **(2015)**.

[6] Y. Yan, "A Practice Guide of Predicting Resource Consumption in a Web Server", Review of Com puter Engineering Studies, vol. 3, no. 2, **(2015)**.

[7] T. Urvoy, E. Chauveau, P. Filoche and T. Lavergne, "Tracking Web spam with HTML style similarities", ACM Transactions on the Web, vol. 2, no. 1, **(2008)**.

[8] M. Cafarella and D. Cutting, "Building Nutch. Open source", Queue, vol. 2, no. 2, **(2004)**.

[9] Ntoulasa, M. Najork and M. Manasse, "Detecting Spam Web Pages through Content Analysis", Proceedings of the 15th International Conference on World Wide Web, ACM, Edinburgh, Scotland, UK, **(2006)**.

[10] S. Adali, T. Liu and M. Magdon-Ismail, "Optimal link bombs are uncoordinated", In First international workshop on adversarial information retrieval on the web (AIRWeb'05), Chiba, Japan, **(2005)**.

[11] KT. Oren, L. Lillian and U. Cornell, "PageRank without hyperlinks: Structural reranking using links induced by language models", Proceedings of Sigir 2010, Geneva, Switzerland, **(2010)**.

[12] Z. Gyongyi, H. Garcia-molina and J. Pedersen, "Combating Web spam with TrustRank", proceedings of the 30th VLDB Conference, ACM Press, Toronto, Canada, **(2004)**.

[13] Asano, Yasuhito, T. Yu and T. Nishizeki, "Improvements of HITS Algorithms for Spam Links", Ieice Transactions on Information & Systems, vol. 2, no. 91, **(2010)**.

[14] J. Abernethy and O. Chapelle, "Graph regularization methods for Web spam detection", Mach Learn, vol. 2, no. 81, **(2010)**.

[15] B. Yang, H. Chen and G. Zhu, "A Novel Page Ranking Algorithm Based on Analyzing the Diversity of Inbound Hyperlinks", Chinese Journal of Computers, vol. 4, **(2014)**.

[16] Y. Yan, "A Practice Guide of Predicting Resource Consumption in a Web Server", Review of Computer Engineering Studies, vol. 3, no. 2, **(2015)**.

[17] R. Wei, J. Zhen and L. Bao, "Study on Mining Big Users Data in the Development of Hubei Auto-Parts Enterprise", Mathematical Modelling of Engineering Problems, vol. 4, no. 2, **(2015)**.

Authors

Xu Gongwen, He received his Master's degree in 2005 from Shandong University. Now he is pursuing the Ph. D. degree in Shandong Normal University. He works in School of Computer Science and Technology Shandong Jianzhu University. His current research interests include spam detection, image processing.

Li Xiaomei, She received her Master's degree in 2004, Ph. D. degree in 2014 from Shandong University. She is a member of Cancer Center of the Second Hospital, Shandong University. Her research interests include medical image processing and spam detection.

Zhang Zhijun, He received her Master's degree in 2005 from Shandong University, Ph. D. degree in 2015 from Shandong Normal University. He works in School of Computer Science and Technology Shandong Jianzhu University. His current research interests include spam detection, machine learning.

Xu Li'Na, She received her Master's degree in 2006 from Shandong University. She works in School of Computer Science and Technology Shandong Jianzhu University. Her current research interests include spam detection, statistical sparse learning.