A Spam Detection System Based on Web Mining

Long Wang[1, a], Hua Pang[2, b]

[1]College of Information, Liaoning University, Shenyang 110036, China

[2]College of Education Technology, Shenyang Normal University, Shenyang 110034, China

[a]email_wl@163.com, [b]panghua@sohu.com

Abstract

In this paper, a new spam detection system based on web minning is presented. The web mining technology is simply introduced. The work process of this system is depicted in detail and the implementation scheme is given. The experiment shows that the system can effectively prevent the spam emails, and the scheme can make all kinds of users be convenient to use the email system.

Keywords

Web Mining, Spam, Spam Detection, KNN.

1. Introduction

The negative aspects of email are well known. Spam consumes infrastructure resources — especially bandwidth — and, like arbitrary emails from co-workers, is a huge waste of time. Unlike indiscriminate emails from colleagues, however, spam has its sinister side. V-spam and phishing are two good examples of areas where spam has metamorphosed from being purely time wasting to become inherently dangerous. So the study of spam filter technology has very important significance [1].

Spam filter technology mainly includes filter based on blacklist, filter based on email header check, filter based on email contents check, and so on. The servers and clients of the spam filter system accomplish the multi-filter by using the methods mentioned above. Presently, the filter based on email contents check is the main technology used in filtering spam.

2. Web Mining Technology

Web mining [2] is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types: Web usage mining, Web content mining and Web structure mining.

2.1 Web Usage Mining

Web Usage Mining [3] is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

Web usage mining itself can be classified further depending on the kind of usage data considered:

(1)Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.

(2)Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

(3)Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

2.2 Web Structure Mining

Web structure mining [4] is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

(1)Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

(2)Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

2.3 Web Content Mining

Web content mining [5] is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, Met Crawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.

3. Spam Detection System

3.1 System Model

According to the function requirements of the spam detection system as well as the organization relations, the cooperation relations and the business relations of the system. The function model can be distributed to different host computers. The system architecture, the relationship among the subsystems, and the relationship among the models of the subsystems are shown in Fig.1 in detail.

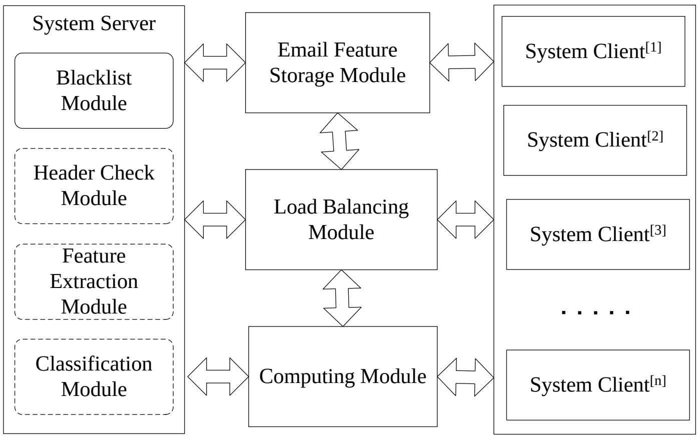

Fig.1. System Model

3.2 Detection Method

The method is the KNN Method [6], If the feature representation of the new mail is d , the sample space S has two sorts: spam m_1 and non-Spam m_2 . The similarity is:

$$Sim(d,s_i)=\frac{\sum_{j=1}^{n} d.t_j \times s_i.t_j}{\sqrt{(\sum_{j=1}^{n}(d.t_j)^2)\times(\sum_{j=1}^{n}(s_i.t_j)^2)}} \tag{1}$$

Where n is the amount of features, $s_i \in S$, $d.t_j$ is the weight of the features j in d , $s_i.t_j$ is the weight of the features j in s_i 。 Composed the neighbor set with k resources which have the bigger similarity, the probability P_1 and P_2 are:

$$P_l = \sum_{S_i \in K} Sim(d,s_i)y(s_i,m_l) - b_l \quad l = 1,2 \tag{2}$$

Where $y(s_i,m_l)\in\{0,1\}$, if $s_i \in m_l$, $y(s_i,m_l)=1$, else $y(s_i,m_l)=0$; b_l is the threshold of m_l .
If $P_1 > P_2$, then the mail is a spam, otherwise the student is a non-spam.

4. Experiment

Select 5000 emails, among which there are 2861 spam emails and 2139 non-spam emails. Extract randomly 50% emails as learning set in spam set and non-spam set. The remaining emails are divided into five parts used as 10 test sets. Then apply the data set to test the spam detection system, The result of the experiment is shown in table 1.

Table 1. Experiment Result

Test Set	Precision (%)	Time cost (sec.)
1	91.23	175
2	92.32	186
3	90.56	169
4	91.41	172
5	90.22	171
6	91.13	168
7	92.01	181
8	90.89	172
9	91.03	169
10	90.25	177
avg	91.10	174

It is shown in table1 that the precision of the method is 91.1%, and these meet the demands of spam detection.

5. Conclusions

A new spam detection system based on web mining is presented in this paper. The experiment shows that the system can effectively prevent the spam emails, and the scheme can make all kinds of users be convenient to use the email system.

References

[1] Spam Statistics 2004 [EB/OL]: http://www.spamfilterreview.com, 2004.

[2] Kolari, P., Joshi, A.: Web mining: research and practice. Computing in Science and Engineering, vol.6, pp. 49-53. (2004)

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, et al. Web usage mining: discovery and applications of usage patterns from Web data [J]. Appear in SIGKDD Explorations, 2000(2), 01:12-23

[4] Eirinaki, M., Vazirgiannis, M.: Web Mining for Web Personalization, ACM Transactions on Internet Technology, Vol.3, No.1, February 2003.

[5] Srivastava, J., Desikan, P., Kumar, V.: Web Mining: Accomplishments and Future directions. In: National Science Foundation Workshop on Next Generation Data Mining. Baltimore, Maryland (2002).

[6] Bijalwan V, et al. KNN based machine learning approach for text and document [J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.