**A SYNOPSIS ON**

Spam identification in cloud computing

based on text filtring system



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**MASTER OF COMPUTER APPLICATIONS**

**Submitted by:**

**Student Name -SANJAY GUPTA University Roll No**

**- 23391100**

***Under the Guidance of***

**Dr. NEELAM SINGH**



**Department of Computer Applications**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

Introduction

E-mail is one of the most popular applications on the Internet. E-mail is gradually replacing the traditional way of communication. Its simplicity and instantaneity make it the main way of electronic communication in modern society, so sending e-mail is a very important means of communication in modern people’s life, study, and work. However, all kinds of spam will bring many problems to people . Therefore, effective email filtering is an important topic of network information security research. Although the utilization rate of e-mail is increasing rapidly, and it has become one of the important ways of rapid communication between mainstream social groups, however, various types of spam are spreading . Therefore, in order to ensure the normal use and security rights of users, it is necessary to ensure the accuracy and security of the email filtering system and to build and maintain an orderly, healthy, and harmonious network environment on the Internet . Therefore, anti-spam technology has become a hot research topic of many scholars.

Spam is sent through the Internet. Spam can be created and modified very quickly, so all relevant data must be kept up to date . Therefore, spam treatment needs an integrated and more flexible system platform, that is, anti-spam system. Because of the uncertainty of anti-spam system, the process of feedback and re-judgment is added to it, thus realizing the process of text filtering system based on active learning.

Objective

With the rapid increase of spam on the Internet and the diversification of its forms, how to quickly and effectively identify a large number of spam on the Internet has become an urgent topic. Cloud computing has obvious advantages in storage and processing, so it can effectively calculate a large amount of mail data.

1. Spam and Text Filtering System

Definition of Spam

Generally speaking, any email that is forcibly sent to the user’s mailbox without the user’s permission is spam. Spam will force users to receive it, and it cannot be blocked or rejected, and it does not indicate the sender’s identity, address, title, and other information. It is mostly sent in the form of advertisements, electronic publications, various forms of promotional emails, and publicity emails .There are many people who send spam in various ways on the Internet. After the mail is sent by the Web server, it will arrive at the recipient’s web mail server, where it can be saved and forwarded . However, the identification and filtering of spam filters may lead to false detection, so some spam may also be misjudged. To judge whether it is real spam, further investigation is needed.

The common way to distinguish ordinary e-mail from spam is to analyze the content of e-mail and use the rule set created by human beings or machine learning methods to judge and distinguish. However, it is difficult to find the difference between ordinary e-mail and spam only by analyzing the e-mail text to determine whether it is spam or not. Because there are many kinds of human languages, the understanding and acceptance of information includes not only text, but also graphics and associations related to the text. Therefore, it is difficult to establish a universal and efficient text filtering model to analyze whether e-mail is spam or not . Moreover, the method of artificially creating rule sets is not popular, because everyone’s experience of e-mail is different. Therefore, to distinguish and identify spam quickly and effectively, other more effective methods must be used. Judging from the above definition of spam, the difference between ordinary email and spam lies in whether email is the email that users want to receive . Of course, a normal email is the email that the recipient wants to receive.

2.Ways to Send Spam

At present, most Internet users use simple anti-spam measures. Because the free e-mail address is a very simple e-mail address for users, they often receive spam in most cases. Now, most spam will be sent in the form of relay stations. This is the server used by remote computers to send spam.

When sending e-mail, the e-mail transfer protocol is used, but because of its user authentication vulnerability, that is, allowing users to forward e-mail indefinitely or send e-mail anonymously, spam can easily abuse any e-mail address. At present, most upgraded mail servers support turning off unlimited forwarding. However, in many cases, due to the negligence of the system administrator, this vulnerability can be fixed immediately.

3. The Harm of Spam

Sometimes, junk will seriously affect our work and life. Too much junk mail may make it difficult for people to judge which emails are useful to themselves among all the emails in the mailbox. Spam usually contains advertisements or fraudulent websites that deceive us to make us trust. Inadvertent linking at this time may lead to personal information leakage or even greater losses.

As we all know, China is one of the countries with the slowest Internet access in the world, and the Internet is only used for research, work, and entertainment. So if large-scale spam is sent to our computer, the result will be unimaginable. This will not only slow down Internet access, but also cause a huge waste of limited network resources.

4.Cloud Computing and Spam Filtering Technology

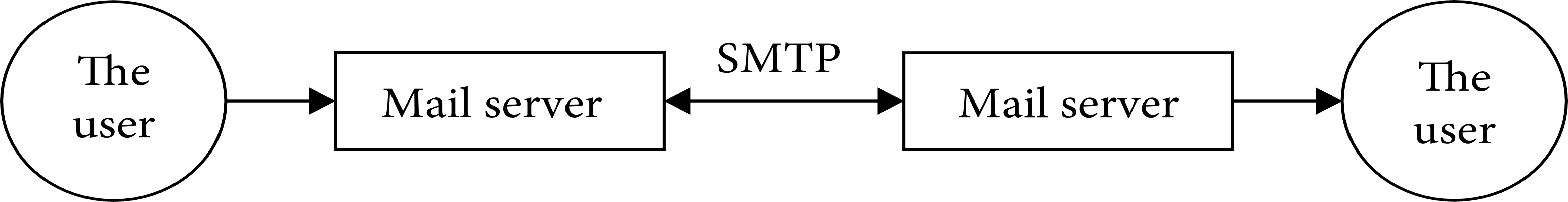
Cloud Computing

Cloud computing is a kind of distributed computing, that is, a huge data computing and processing program is decomposed into many small programs through the network “cloud,” then these small programs are processed and analyzed by a system composed of multiple servers, and then the obtained results are returned to users.

Cloud computing has the advantages of high resource sharing, low cost, and significantly improved computing speed. Using this technology on the cloud platform can quickly realize the unimaginable large-scale data processing.

Mail Filtering System

E-mail in the mail filtering system is the same as ordinary mail, and the sender specifies the recipient’s name and unique network e-mail address. The data arrives at the server containing the recipient’s domain name and is sent to the mail recipient.



To prevent the intrusion of spam and virus files, a filtering system must be installed in the mail system. This needs to be achieved by rejecting known spam to filter the host information of spam and other means. For virus files, you can use various methods to achieve various functions. In reality, there are three kinds of spam filtering technologies to realize and form a hierarchical spam filtering system, which are black and white list filtering, rule filtering, and Bayesian filtering.

5. Spam Filtering Technology

The same content in the spam will be sent to tens of thousands of recipients, that is, multiple copies will be sent simultaneously through the Internet. The problem of spam should be comprehensively dealt with and managed by combining technical means and legal means. Decentralized spam filters are faced with some shortcomings, such as incomplete collected data sets and untimely updating of algorithms and rules. Centralized spam filters are facing the challenges of high-capacity storage, high-density computing, and user privacy protection.

There are two kinds of spam filtering technologies: One is to discover when it exists, and the other is to reject it fundamentally. On-the-fly rejection means that if the system finds that the user is using self-developed software to send e-mail. In addition, the transmission was very special in that the system detected a large number of e-mail messages sent by the user in a very short period of time, and the number and speed were far beyond the normal range. That would cause the system to block such transmissions. This has proven to be very effective by using many service providers. It is not hard to find that the main research content of the staff in this field is spam filtering technology.

6.Anti-Spam-Related Methods

The current filtering technology mainly includes the following three methods: Rule-based approach. This method filters the header information, e-mail, and body text information rules of e-mail based on user-defined rules. Because the filtering rules are defined by users themselves, this method is more flexible, but the operation of this method is a bit complicated, so the quality requirements of users are relatively high Black and white list method. This method can be understood as follows: All e-mails sent by white-listed senders are considered legitimate, while e-mails sent by black-listed senders are considered spam. This is a very simple method, and it is one of the methods used for email and SMS filtering tasks. Black list and white list filtering technology need to develop and maintain this list. This list, whether it is black list or white list, can be sender, email server address of domain name, specific IP address and email address, etc.Statistical methods. Statistics is the method of ana-yzing known email information and making statistics. This method is similar to the text classification method in that it uses this information to classify e-mails. At present, most large-scale mail systems in China cannot detect spam within a certain period of time and solve it effectively, which takes up the memory space and bandwidth of the mail server, so that users are often troubled by spam and take up their energy and time. As well as posing a risk to network security, the vulnerability is abused by foreign spam, so it can cause considerable losses in terms of lost users. Therefore, developing a new generation of spam filtering system with globalization, high efficiency, and high identification reliability has become a very urgent issue.

7.Research on the Method of Intelligent Identification

The method of realizing optimization by simulating the known evolutionary methods in nature is called intelligent optimization algorithm. It mainly depends on a calculation method to popularize a series of phenomena and processes in the biological world, nature, and the objective world. Compared with the traditional algorithm, it has the characteristics of looser requirements for objective function and constraint function. The algorithm always has a good solution and can even be terminated at any time. This method has no high requirements for the objective function and will not be limited by the solution because of irregularity or discontinuity. At present, the commonly used intelligent identification methods include genetic algorithm, particle swarm optimization, ant colony algorithm, simulated annealing algorithm, and tabu search algorithm. This paper focuses on genetic algorithm and tabu search algorithm.

Genetic algorithm mainly simulates the reproduction process of natural animals and plants such as heredity, chromosome crossing, multiband reproduction, and individual variation and finally obtains the best population by the method of survival of the fittest. Select the most suitable individual as the optimal solution, construct an adaptive function according to the optimization objective of the problem, take the corresponding chromosome as the initial population, and evaluate it. Then, individuals are selected and propagated, so as to carry out genetic cross-mutation and multigeneration propagation. Finally, take exceeding algebra or being consistent with the result as the end condition, and then select the most suitable individual for optimization as the solution.

Tabu search algorithm is a combination optimization algorithm, which simulates the mental process of human beings through the introduction of flexible storage structure and tabu rules, and makes use of the contempt principle to avoid some taboo good states, so as to achieve global optimization. In order to avoid detours, it will search and record the best points that have been reached.

8.Workflow of Internet Spam Filtering System

As a new resource utilization method, cloud computing supports most users to get the required resources through the network in an on-demand, simple, and extensible way. Mail disposal has the characteristics of large computing scale, wide sharing range, and quick response. Nowadays, the cost of storage equipment is low, and the network transmission speed is greatly improved compared with before. Therefore, the advantages of cloud computing should be fully utilized in the processing of e-mail, without considering too many size restrictions, too much calculation of sample base and system, etc. E-mail filtering system on cloud should monitor new spam and variant spam more accurately based on the quality and speed of intercepted e-mail. Therefore, the mail filtering system must constantly update the sample library.

The spam filtering process on cloud platform can reduce many manual and repetitive work steps. By constructing adaptive algorithms with feedback learning, the problem of rapid increase and variation of spam can be solved. Valid special characters include $, -,1-9, comma, and #. Spammers use keywords to confuse the spam filter, for example, replace 0 with O to improve the difficulty of filtering.

Traditional e-mail filtering workflows, by default, search for specific attributes and default matches, judging by logic, such as deleting, bouncing, rejecting, and forwarding the contents of e-mail text. Through the design of the spam filtering system model and the use of the current mainstream filtering technology, we can see the advantages of multilevel message filtering from the above analysis. The system can keep the normal sending and receiving of mails, effectively filter junk mails, and minimize redundant operations, thus improving the automatic learning ability of the system. Filtering false detection or missing emails can reduce the possibility of false detection or false detection, fully collect and utilize user feedback to improve system performance, and make manual changes to improve the success of spam filtering.

References

A. Bernárdez Rodal, G. Padilla Castillo, and R. P. Sosa Sánchez, “From action art to Artivism on Instagram: relocation and instantaneity for a new geography of protest,” *Catalan journal of communication & cultural studies*, vol. 11, no. 1, pp. ,

M. Léouffre, F. Quaine, and C. Serviere, “Testing of instantaneity hypothesis for blind source separation of extensor indicis and extensor digiti minimi surface electromyograms,” *Journal of Electromyography and Kinesiology*, vol. 23, no. 4, pp. 908–915, 2013.

T. H. Silva, A. C. Viana, F. Benevenuto et al., “Urban computing leveraging location-based social network data,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–39, 2020.

View at: [Publisher Site](https://doi.org/10.1145/3301284) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Urban computing leveraging location-based social network data&author=T. H. Silva&author=A. C. Viana&author=F. Benevenuto&publication_year=2020)

G. Taubes, “Publication by electronic mail takes physics by storm,” *Science*, vol. 259, no. 5099, pp. 1246–1248, 1993.,M. Viceconti, S. Olsen, L. P. Nolte, and K. Burton, “Extracting clinically relevant data from finite element simulations,” *Clinical biomechanics*, vol. 20, no. 5, pp. 451–454, 2005.

View at: [Publisher Site](https://doi.org/10.1016/j.clinbiomech.2005.01.010) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Extracting clinically relevant data from finite element simulations&author=M. Viceconti&author=S. Olsen&author=L. P. Nolte&author=K. Burton&publication_year=2005)