Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9831576, 1 page
https://doi.org/10.1155/2023/9831576

Retraction

Retracted: Spam Identification in Cloud Computing Based on Text Filtering System

Wireless Communications and Mobile Computing

Received 27 June 2023; Accepted 27 June 2023; Published 28 June 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] R. Mu, "Spam Identification in Cloud Computing Based on Text Filtering System," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 2309934, 7 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2309934, 7 pages
https://doi.org/10.1155/2022/2309934

Research Article

Spam Identification in Cloud Computing Based on Text Filtering System

Rong Mu

Network Center & Information, Xi'an University of Science and Technology, Xi'an Shaanxi 710054, China

Correspondence should be addressed to Rong Mu; mur@xust.edu.cn

Received 28 June 2022; Revised 25 July 2022; Accepted 28 July 2022; Published 21 August 2022

Academic Editor: Mohammad Farukh Hashmi

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. Due to the uncertainty and life cycle of spam, feedback re-judgment is added to the anti-spam system, and a text filtering system based on active learning with four stages of training, filtering, feedback, and re-filtering is implemented. Compared with the original system, the filtering system with feedback can improve the filtering of keywords. In order to effectively reduce the misjudgment rate of ordinary mail and improve the accuracy of spam judgment, it is suggested to improve the use of weighted decision-making of email header information to implement effective auxiliary classification. For emails lacking content, the filtering method of title weighting is feasible and effective, which can improve the identification of spam with relatively little text content. Because the filtering method on the cloud is far more advanced than the traditional algorithm, the development of the Internet can effectively solve the infinite increase of spam. Therefore, this paper makes an in-depth study on spam identification in cloud computing based on text filtering system by summarizing and analyzing the current anti-spam technologies.

1. 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 [1–3]. 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 [4, 5]. 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 [6–8]. 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 [9]. 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.

2. Spam and Text Filtering System

2.1. 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 [10–12].

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 [13, 14]. 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. And the life cycle of spam is shown in Figure 1.

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 [15–17]. 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 [18, 19]. 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 [20, 21]. Of course, a normal email is the email that the recipient wants to receive.

2.2. 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.

2.3. 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.

2.4. Overview of Text Filtering. Text filtering is mainly divided into cooperative filtering and content-based filtering. Cooperation is also called social filtering. In content-based filtering, each user operates independently. In cooperative filtering, everyone must belong to a group, not exist independently.

Generally, people will get the information they want from the results recommended by others. Therefore, suggestions are made to other users based on the comments of users who have the same or similar interest in the corresponding text. This mode is content-independent, so it is suitable not only for text format, but also for non-text media, such as audio, images, and videos. Therefore, it can be seen that this paper is based on the text filtering system.

2.5. Cloud Computing and Spam Filtering Technology

2.5.1. 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.

Although the concept of cloud computing was put forward in 2007, its development speed is very fast, and its attention is very high. For example, Apple Computer Company, google Company, Microsoft Company, and Oracle Bone Inscriptions Company have established organizations dedicated to cloud computing. The scale and growth rate of China's cloud computing market from 2015 to 2023, and its forecast are shown in Figure 2.

2.5.2. Features of Cloud Computing. Cloud computing technology can make the best use of network resource sharing to highlight the advantages of network computing. Therefore, data can be exchanged quickly and on a large scale through the Internet. In a large-scale computer cluster, a large task can be divided into many small tasks and calculated separately. This enables users to obtain supercomputing power at a lower cost and to conduct large-scale data processing.

We need to make the best use of today's Internet to develop this technology. The present situation can already meet the software and hardware requirements of cloud computing applications. The network bandwidth that affects the data transmission rate is also constantly improving and increasing. This lays a foundation for large-scale data transmission and also provides more possibilities for cloud computing.

The core of "cloud" is resource sharing, such as storage space resources, computing resources, and information resources, which collects all resources available to users.

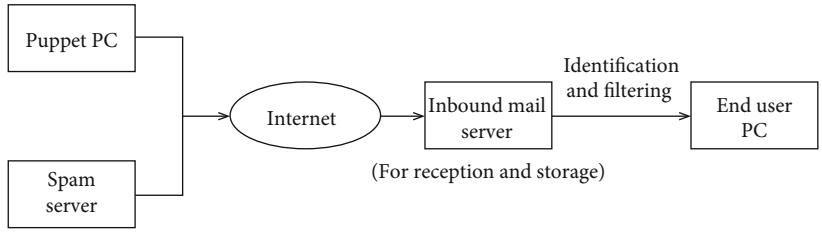

Figure 1: Life cycle of spam.

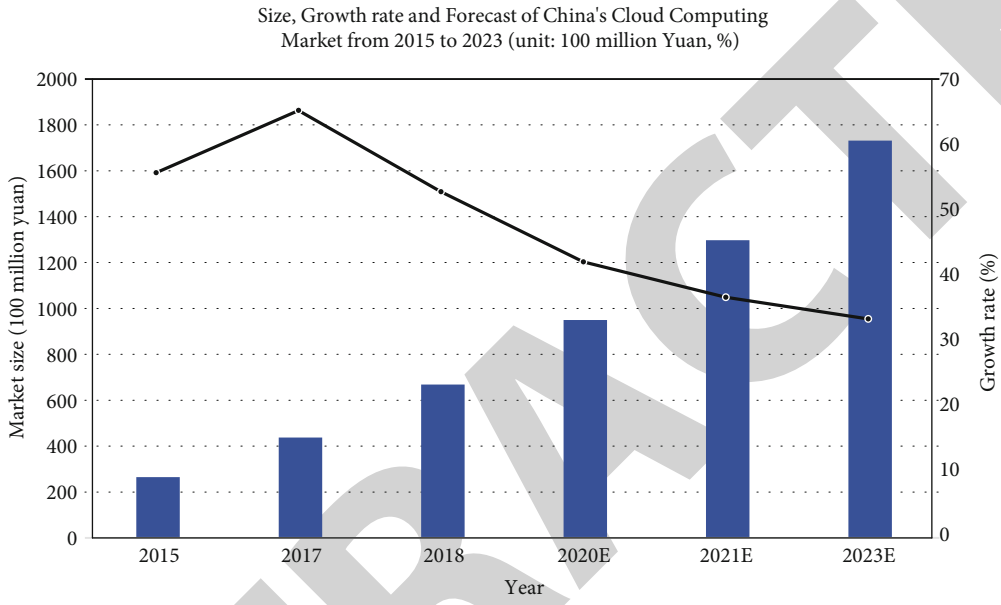

Figure 2: 2015-2023 China cloud computing market size and growth rate and forecast.

The resource is a virtual cluster of mainframe computers. Clusters of computers can run pre-programmed computer programs and intelligently perform self-learning management of large amounts of information, as well as delete unwanted information in real time. Cloud computing can pull information from many places, including a lot of information about the servers and networks of large computer companies. The biggest advantage of this information technology is that it can achieve many functions and has self-management function, that is, no manual intervention is required.

2.6. 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, so the e-mail transmission path is shown in Figure 3.

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.

2.7. 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.

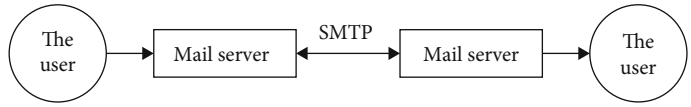

Figure 3: E-mail transmission channels.

3. Research on Spam Identification

3.1. Anti-Spam-Related Methods. The current filtering technology mainly includes the following three methods:

(1) 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

(2) 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.

(3) Statistical methods. Statistics is the method of analyzing 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

3.2. 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.

3.3. 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.

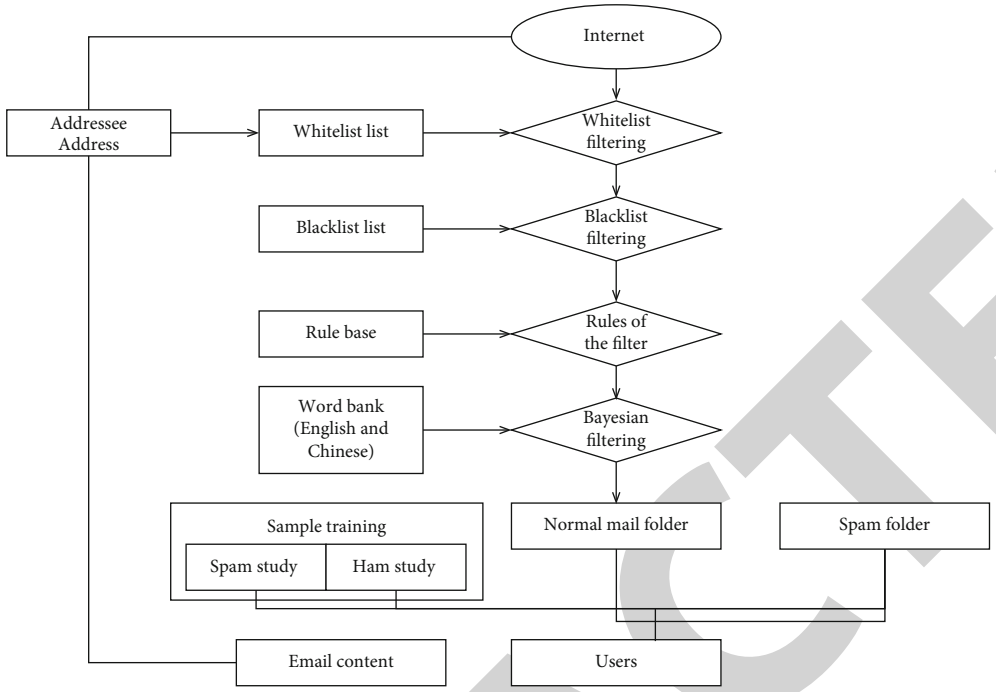

Figure 4: Structure diagram of mail filtering system.

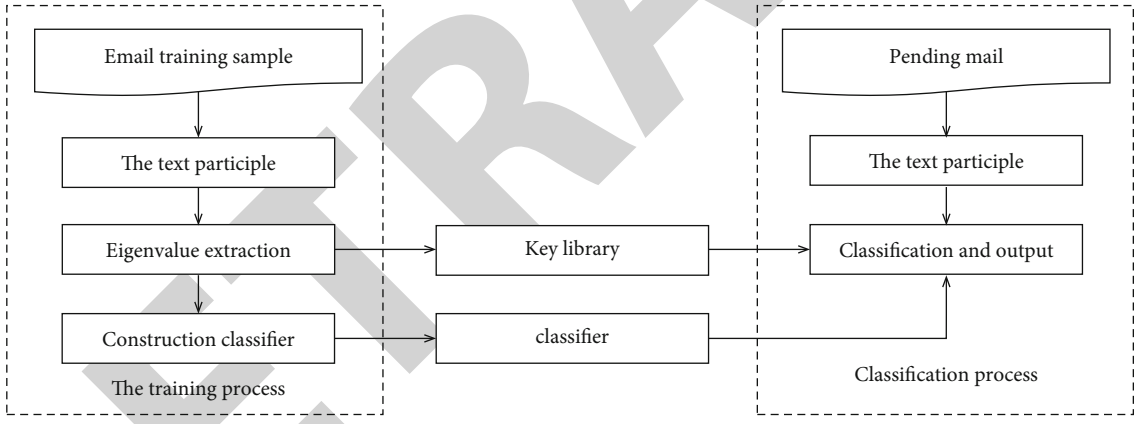

Figure 5: Frame diagram of spam filtering system.

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. The structure diagram of mail filtering system is shown in Figure 4.

3.4. Anti-Spam System Process Optimization. In the anti-spam system, whenever the cloud receives a message from a client agent, the cloud will deliver the message to the anti-spam network service, and then the server will generate a result report, which contains the information obtained after analyzing and processing the mail, and finally report the mail result. After returning to the client, the mail will be stored in the sample library according to the results.

In order to make more effective use of the unlimited storage and fast computing advantages of the cloud, a new feedback process will be proposed, and the sample library of email filters will be expanded. That is, the determined e-mail will be stored in the sample library according to the results. The example library can reevaluate the identified e-mail in the actual situation. While effectively adjusting the error, it can identify the new variant spam

Figure 6: Design of feedback operation.

feature words. The structure of the spam system is shown in Figure 5.

In the anti-spam system, the selected training samples are divided into spam samples and common samples, and the number of copies of each sample is counted, respectively. For each training example, word segmentation is performed separately. In the process of word segmentation, the frequency of characteristic words in spam sample set and common email sample set is counted, and the results are saved in the database. The feedback operation design is shown in Figure 6.

The information gain value of each word is calculated, and the synonyms of word segments are sorted according to the size of the information gain value. Then, the number of feature words is extracted according to the self-defined vector space dimension to build the feature word database. After receiving a batch of samples, the system will train the samples one by one and update the feature word frequency by calculating the probability of updating the feature word database.

4. Summary

With the integration of global information and the rapid development of the Internet in China, the effective identification of Internet e-mail as an important research topic of network security has attracted more and more attention from people in the industry and users. With the rapid development of cloud computing this year, it also has its own advantages in large-scale text processing, so it can be widely developed and applied in the field of email filtering. However, spam will be affected by the current status of e-mail, so in order to effectively solve the problem of e-mail identification, integrated identification has become a necessary method.

Firstly, this paper studies and analyzes the current spam filters. Based on the analysis of the influencing factors of email identification accuracy, an optimization system process based on cloud computing is established, and the intelligent optimization algorithm is combined to improve the identification accuracy. Therefore, this paper introduces

genetic algorithm and tabu search algorithm in anti-spam system.

The improvement of the spam filtering system includes the following: Firstly, a lot of manual and repetitive work can be reduced through autonomous learning and secondly, an adaptive algorithm with feedback is studied, the implementation process of the system framework is described, and the accuracy optimization model is established. The research results show the effectiveness of this model. With the continuous development of the Internet, the number and variation speed of spam are developing to an infinite number. Cloud computing is the development direction of spam filtering. If we can conduct experiments on emails sent from the whole network, the results will be more satisfactory.

Data Availability

The figures used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to show sincere thanks to those techniques who have contributed to this research.

References

[1] 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. 23–37, 2019.

[2] H. Herzogenrath-Amelung, "The new instantaneity: how social media are helping us privilege the (politically) correct over the true," *Media, Culture & Society*, vol. 38, no. 7, pp. 1080–1089, 2016.

[3] 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.

[4] 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.

[5] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "Security and accuracy of fingerprint-based biometrics: a review," *Symmetry*, vol. 11, no. 2, p. 141, 2019.

[6] R. R. Kobak and C. Hazan, "Attachment in marriage: effects of security and accuracy of working models," *Journal of Personality and Social Psychology*, vol. 60, no. 6, pp. 861–869, 1991.

[7] B. C. Williams, L. B. Demitrack, and B. E. Fries, "The accuracy of the national death index when personal identifiers other than social security number are used," *American Journal of Public Health*, vol. 82, no. 8, pp. 1145–1147, 1992.

[8] R. G. Saltman, "Accuracy, integrity and security in computerized vote-tallying," *Communications of the ACM*, vol. 31, no. 10, pp. 1184–1191, 1988.

[9] G. Li, B. Liu, S. J. Qin, and D. Zhou, "Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: the dynamic T-PLS approach," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2262–2271, 2011.

[10] J. A. Evans, "Electronic publication and the narrowing of science and scholarship," *Science*, vol. 321, no. 5887, pp. 395–399, 2008.

[11] S. Harnad, "Electronic scholarly publication: quo vadis?," *Serials Review*, vol. 21, no. 1, pp. 70–72, 1995.

[12] G. Taubes, "Publication by electronic mail takes physics by storm," *Science*, vol. 259, no. 5099, pp. 1246–1248, 1993.

[13] H. Cheng, D. Yang, C. Lu, Q. Qin, and D. Cadasse, "Intelligent oil production stratified water injection technology," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3954446, p. 7, 2022.

[14] 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.

[15] G. V. Cormack, "Email spam filtering: a systematic review," *Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2008.

[16] L. F. Cranor and B. A. LaMacchia, "Spam!," *Communications of the ACM*, vol. 41, no. 8, pp. 74–83, 1998.

[17] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, pp. 1–24, 2015.

[18] H. Cheng, J. Wei, and Z. Cheng, "Study on sedimentary facies and reservoir characteristics of Paleogene sandstone in Yingmaili block," *Geofluids*, vol. 2022, Article ID 1445395, 14 pages, 2022.

[19] J. Wei, H. Cheng, B. Fan, Z. Tan, L. Tao, and L. Ma, "Research and practice of" one opening-one closing" productivity testing technology for deep water high permeability gas wells in South China Sea," *Fresenius Environmental Bulletin*, vol. 29, no. 10, pp. 9438–9445, 2020.

[20] W. Zhang, Z. Cheng, H. Cheng, Q. Qin, and M. Wang, "Research of tight gas reservoir simulation technology," *IOP Conference Series: Earth and Environmental Science*, vol. 804, no. 2, article 022046, 2021.

[21] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.