Advances in Social Science, Education and Humanities Research, volume 347
3rd International Conference on Economic Development and Education Management (ICEDEM 2019)

Analysis of Bad Website Filtering System on Intelligent Terminal under Cloud Computing

Yan Li
College of Intelligent Science and Information Engineering
Xi'an Peihua University
Xi'an, China

Kun Chen
College of Intelligent Science and Information Engineering
Xi'an Peihua University
Xi'an, China

***Abstract*—With the rapid development of information technology and intelligent terminal, it is more convenient for people to use intelligent terminal to surf the Internet. More and more users choose to use intelligent terminal to surf the Internet. This paper studies how to implement an efficient and accurate method for filtering undesirable information on intelligent terminals, realizing automatic screening of undesirable website functions, and ensuring smart phones to surf the Internet in a civilized network. At the same time, the overall design of the system is given in combination with cloud computing platform, which is of great significance.**

Keywords—cloud platform; website filtering; system design

I. INTRODUCTION

In recent years, the Internet has been integrated into people's life. More and more people choose the information they need from the Internet, which has greatly promoted the booming development of the Internet. At the same time, people are no longer relying only on personal computers to obtain information on the network. Mobile devices represented by intelligent terminals have become the priority choice for people to connect to the Internet anytime and anywhere due to their advantages of convenient carrying and fast use. According to a recent statistical report on the development of the Internet in China, the total number of Internet users in China had reached 772 million by the end of 2017, among which the number of mobile Internet users reached 753 million, accounting for 97.5%.

It can be seen from this that the Internet in China is developing rapidly. All kinds of information on the Internet enrich people's life, and people are willing to share knowledge on the Internet, thus attracting more people to join the Internet family. And with the rapid development of intelligent terminal, its characteristics of convenient to connect to the Internet to attract people, and the traditional Internet company also gradually in the intelligent terminal to provide their own business, people also can be done through intelligent terminal by connecting to the network at any time and place for news, watching video, online shopping, and search information, etc. The Internet can meet many people's needs, but it also brings some disturbing things, such as advertising harassment, trojans, Internet fraud, and bad information dissemination and so on. Among them, bad information refers to all kinds of information that violates relevant laws of the People's Republic of China and social morality. However, bad website filtering is a process of selecting a large number of websites and their contents to meet the objective needs.

Subsidy Project at School Level of xi'an Peihua University.

At the international level, many countries have also taken a series of measures to crack down on bad information to prevent its negative impact on adolescents. The Japanese government has held relevant meetings, introduced the Youth Network Regulation Act and the Bad Website Countermeasure Act, which stipulate that mobile phone service providers are obliged to provide filtering services to help teenagers stay away from bad websites, while filtering services are simply black-and-white lists. South Korea is considering mandatory installation of filtering applications on teenagers' smartphones, and also requires mobile network providers to provide filtering services to block pornography and other undesirable information. The European and American countries also take similar measures or introduce manual auditing mechanism, but the main work is still to protect the PC-side, mobile-side products are still the primary stage of black-and-white list filtering, for the growing number of websites can not be intelligently identified, it is not suitable for promotion and use.

II. CLOUD BAD WEBSITE FILTERING RELATED TECHNOLOGIES

At present, the main filtering technologies for bad websites at home and abroad fall into two categories: static filtering technology and dynamic network content analysis identification filtering technology.

Static filtering technology mainly includes url-based filtering and keyword-based filtering. Static filtering is commonly used and has been commercialized. It mainly determines whether a website is bad or not by comparing databases. Dynamic network content analysis, identification and filtering technology includes text analysis technology and image analysis technology, which can effectively identify and filter undesirable content. Text content filtering technology is most often used to categorize text content, by detecting the categories that text belongs to, And filter it as a judgment. The whole process of filtering can be generally divided into two stages, the first stage is over the establishment of filtering rules, the text content on the web page is trained before filtering, and the rule base of filtering is established in advance.

The second stage is to test the content. At this time, judge according to the rule base trained in the first stage. So throughout It is very important to select and improve the algorithm in the process of filtering. At present, the techniques of text classification and filtering include naive bayes and god

through the network, KNN and other algorithms, many researchers have proposed improved SVM. Filtering system in text classification, feature selection is also very important. The purpose of feature selection is to reduce the dimension of feature vectors, eliminate some original input contents, and retain a minimum subset of features so as to achieve the best classification performance. Feature selection can not only improve text scores class accuracy, but also can improve the speed of classification, so for text classification filtering, choose the appropriate classification method is also ten part important.

At present, skin color detection, face detection and other methods have been widely used in bad information image detection, and in practical applications has made a lot of progress. Skin color detection technology is mainly used in face detection, body recognition and other bad image detection problems. For bad images, The most intuitive feeling is that there is a lot of exposed skin, which is also an important feature in bad images. At present the main all the bad image detection technologies required will adopt skin color detection technology as a criterion for judging. By recognizing the approximate position of the human body in the image and referring to other features, the position and contour of the face can be determined. Therefore, the accuracy of skin color recognition will affect the identification of pornographic images.

III. OVERALL DESIGN OF THE SYSTEM

This system realizes real-time detection and filtering of bad websites based on cloud computing platform, and judges whether the websites visited by users on mobile end are bad websites. It first adopts two kinds of mature filtering technologies, URL filtering and keyword filtering, to judge whether the websites visited by users are bad websites. And focus on the analysis of the pictures. The functions of the system are displayed on the mobile phone, while the analysis and filtering of data are carried out on the cloud platform.

The main functions of this system are as follows:

(1) Analysis and detection of text and text on Web pages browsed by users. If it is bad text, you need to prompt not to

Good type.

(2) Analyse the pictures on the web pages that users browse. If they are bad pictures, they need to prompt the bad types.

(3) If users browse websites with both bad text and bad pictures, they will combine the two to analyze bad websites.

Type, need to prompt bad type.

(4) Record the URLs of bad websites as the result of analysis into the blacklist database so as to visit similar websites next time.

When the station is in operation, it filters directly without any further analysis and judgment.

This system is mainly divided into two modules: mobile terminal and cloud. Users can set custom filtering rules on the phone, Contains url black and white list and custom text keywords two items. The cloud is based on Spark platform for data processing and analysis, and USES K nearest neighbor algorithm for web page. Bad information can be classified and processed, including bad text and bad pictures. First samples will be collected in the network text and image sample input to Spark platform based on KNN classifier, classification model, and generate the filter rule base, so that subsequent bad information filtering, after the treatment of test data to classify, if the web page information both text and image information, will be in after handling these two kinds of information, the two types of information to the decision tree model is analyzed, finally it is concluded that the site is bad websites. The overall architecture of the system is shown in Fig. 1.

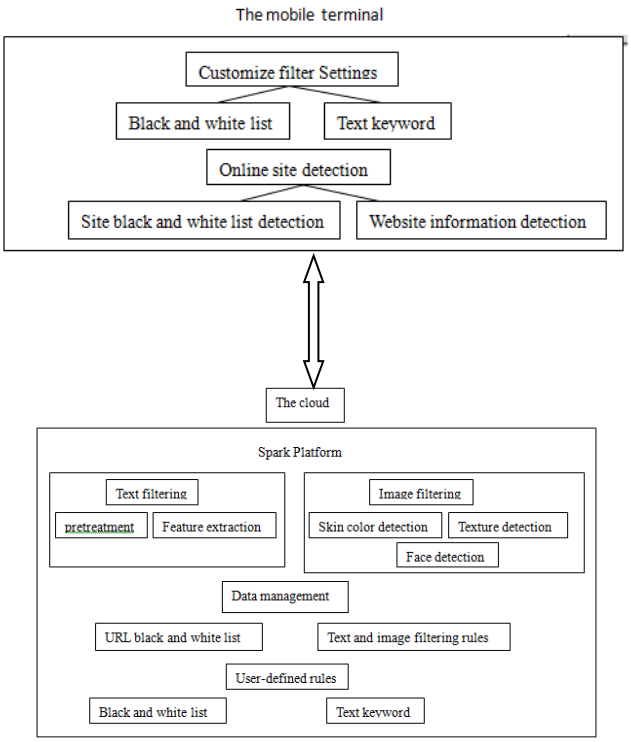

Fig. 1 System Overall Design Architecture

A. Client Function Module

The main operation of the bad website filtering system based on cloud computing is to display on the mobile client. When the user chooses to turn on the intelligent filtering mode, the user can realize the custom filtering when browsing the web page by setting the filtering rules on the mobile client. The mobile client also provides access. Through the way of inputting the website, the website is checked.

a）User-defined filtering rules

The function of this module is mainly to filter the content of the website and the web page according to the users' preferences. Some known websites can be blocked by setting up a black-and-white list of websites, such as websites such as advertising and gambling. Another way is to set keywords to block bad information on the web page, and by setting customs. Keyword, when the user browses the web page and detects the occurrence of these texts, can be reminded to let the user choose whether to continue reading.

b）Online Detection of Websites

This module is used to provide the function of checking unknown websites. When users do not know whether the websites they want to visit are bad websites, users can get the test results by manually entering the URL of the websites on the mobile end. There are two kinds of detection methods. The first type of detection method is black-and-white list detection, which is based on static URL detection method, and then the URL is transmitted to the cloud, which returns the detection results by comparing the black-and-white list in the database. The second kind of detection method detects the content of the website, that is to say, it detects and analyses the content of the input web site. Because it needs to extract the features of the content, this function takes more time than the first kind of detection method.

B. Cloud Function Modules

Cloud computing-based intelligent terminal bad website filtering system core functions are implemented in the cloud, so it will be in the cloud.

The cloud filters bad text and bad pictures. In order to filter bad websites at the right time, first of all, we will adopt a preliminary filtering method, that is, more mature filtering technology, URL filtering based on blacklist and fixed keyword filtering. These do not need to analyze and process the content, only need to compare and identify the data in the database. If the preliminary filtering fails to meet the filtering criteria, the text and pictures of the web page will be filtered. Before such filtering, in addition to pretreatment of these contents, a classifier for bad content information needs to be established.

IV. CONCLUSION

This paper mainly intelligently analyses bad websites according to bad content to prevent mobile terminal users from browsing bad websites. With cloud computing as the core technology support, intelligent analysis is carried out on the websites browsed by users on intelligent terminals to prevent users from accessing unsafe websites with bad content.

The tremendous advantages brought by cloud computing, especially the emergence of Spark platform, make more and more enterprises deploy their systems on cloud computing platform. On the smartphone side, the filtering of bad websites in China uses more traditional filtering methods. Through the analysis of the demand for bad website filtering of smart terminals, This paper presents the overall design of the detection and filtering system of bad websites based on Spark cloud computing platform combined with cloud computing platform. Needs of the platform are analyzed and the system architecture is designed. Then the functions of intelligent terminal and the function modules of cloud are introduced respectively, which lays the foundation for the next system development.

REFERENCES

[1] Santos C, Souto E, Santos E M D.ANDImage: An adaptive architecture for nude detection in image[C], information Systems and Technologies. IEEE, 2015.

[2] Zhou K, Zhuo L, Geng Z, et al. Convolutional Neural Networks Based Pornographic Image Classification[C],IEEE Second International Conference on Multimedia Big Data. IEEE, 2016.

[3] Arya Surendran, Samuel Stephen. Detection of obscene images and ejection of external devices[C], ICECA.IEEE,2017.

[4] Shang E X, Zhang H G. Image spam classification based on convolutional neural network[C], International Conference on Machine Learning and Cybernetics. IEEE, 2017.

[5] Geng Z, Zhuo L, Zhang J, et al. A comparative study of local feature extraction algorithms for Web pornographic image recognition[C],IEEE International Conference on Progress in Informatics and Computing. IEEE, 2016.

[6] Yang Lei, Cao Cuiling, Sun Jianguo, et al. Research on the improved Naive Bayesian algorithm in spam filtering [J]. Journal of Communications,2017.

[7] Huang Cheng. Research and application of a high-speed URL filtering algorithm[J].Modern computer,2016.

[8] Liu Yanbing, Shao Yanbing, Wang Yong, etc. A multi-pattern string matching algorithm for large-scale URL filtering [J]. Journal of Computer Science,2014.

[9] Zhou Qiaojie, Ni Hongjun. A Semantic-based Spam Message Filtering Algorithms [J]. Laboratory Research and Exploration,2016.

[10] Praseetha V M, Vadivel S. Face Extraction Using Skin Color and PCA Face Recognition in a Mobile Cloudlet Environment[C],IEEE International Conference on Mobile Cloud Computing, Services, and Engineering. IEEE, 2016.