Efficient Retrieval Method of Malicious Information in Multimedia Big Data Network Based on Human-Computer Interaction

Jia-ju Gong[1], Wen-da Xie[2], and Jing-hua Wang[3](✉)

[1] Jiangmen Polytechnic, Electronic and Information Technology, Jiangmen 529030, China
gongjiaju5866@163.com
[2] Guangdong Polytechinc of Science and Technology, Computer Engineering Technical College, Guangdong 519090, Zhuhai, China
[3] College of Electronic and Optical Engineering & College of Microelectronics, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
xiewenda4376@163.com

Abstract. In order to solve the phenomenon that the malicious information retrieval in the traditional method is not comprehensive and the precision is not high, an efficient retrieval method of malicious information in multimedia big data network based on human-computer interaction is proposed and designed. On the basis of analyzing the principle of information retrieval, the human-computer interaction form is used to divide the metrics of malicious information. On this basis, the human-computer interaction retrieval logic is established, and the malicious information clustering method is used to realize the multimedia big data network. Efficient retrieval of malicious information. Through the method of experimental argumentation analysis, the validity of the human-computer interaction retrieval method is determined. The results show that the method has a high recall rate of 28.31% compared with the traditional method, and the retrieval accuracy is extremely high. In this paper, the method of network malicious information retrieval is effective and suitable for popularization.

Keywords: Human interaction · Big data network · Malicious information · Search processing · Information clustering

1 Introduction

Human-computer interaction is the study of people, computers and their mutual influence of technology. Man-machine interaction refers to the communication between the user and the computer system. “Conversation” here is defined as a communication, that is, an exchange of information, and a two-way exchange of information, which can be input by a person to a computer, or feedback by the computer to the user. This exchange of information takes the form of various symbols and actions, such as keystrokes, mouse movements, and displaying symbols or graphics on the screen. Systems can be a variety of machines, or they can be computerized systems and software. A human-computer interface is usually the part that is visible to the user. The

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 273–282, 2021.
https://doi.org/10.1007/978-3-030-67871-5_25

user communicates with the system through the man-machine interaction interface and carries on the operation. The principles of interaction design need to include an understanding of the user, the user environment, and the design strategy. Different product designs require specific coping principles and strategies, but the principles of design are the same for all products. Although application principles are not a specific method to solve problems, they are a general principle that can be applied to the whole product design system to interact with users, and can be used as a part of the design strategy, as a guide to the design of interactive toy products. Based on user psychology and cognitive science, the following basic principles are proposed to guide interaction design:

(1) Usability principle

The principle of ease of use requires that interactive toy products be available to most people and have a wide range of applications. Users of different ages, cultures and regions can experience the convenience, efficiency and fun of products in the process of human-computer interaction. A good interactive product can ensure the user in the use of the process of availability and convenience, users in the experience and use of the product will not be unable to understand the manual or complex operation and lose interest.

(2) Adaptability principle

The user is the main body of human-computer interaction and should be in the control position. Therefore, the hardware and software design of the product should adapt to the user's needs in many aspects.

(3) Principle of participability

The main body of human-computer interaction is human and computer. The principle of participability emphasizes the spiritual experience of users when using the product. The product should not only meet the functional requirements of users, but also be easy, effective, interesting, fulfilling and satisfying in the process of interaction between users and the product.

Big data networks can handle multiple large-scale information at the same time, and when it comes to flexibility, it can copy data to multiple locations, and the size of the processed information will become larger and larger. But in fact, the most important attribute of a big data network is not its scale, but the ability to split big data into many small data. It can spread the resources of a task to multiple locations for parallel processing. The efficiency of the data processing process [1]. At this stage, the concept of big data network has become synonymous with cluster environment. According to the characteristics of different applications, the cluster requirements that meet the application are established, and the data load inside the functional partition is realized, and the quantitative relationship between each data load is correctly processed.

Human-computer interaction is a study of the interaction between the system and the user. The system mentioned here can be a variety of machines, or it can be a computerized system and software. The human-computer interaction interface is usually visible to the user part. The user communicates with the system through the human-computer interaction interface, and performs operations to realize the situational description of information communication, and interacts with the real environment to meet the functional requirements of the person.

Reference [2] proposed a network malicious information retrieval method based on big data network, which can obtain implicit feedback information from the network by observing the actions selected by users when browsing web pages, and establish user interest update model. Vector is used to describe the web documents that users browse, and the corresponding weight is given to each browsing behavior. Malicious information is extracted and implicit feedback information retrieval is carried out, which can effectively improve the detection time. But the recall rate is low and the detection efficiency is poor. Literature [3] proposed a Bayesian network malicious information retrieval method, through the Bayesian network technology to build a Bayesian network retrieval model, detailed analysis of its working principle, and the malicious information retrieval method is optimized, this method can improve the detection efficiency, but the detection accuracy is low.

Aiming at the above problems, this paper proposes and designs an efficient retrieval method of malicious information in multimedia big data network based on human-computer interaction, and validates the effectiveness of the method in the simulation platform. The results show that the human-computer interaction retrieval method can Improve the recall rate of malicious information, and thus improve the retrieval accuracy of malicious information, which has higher superiority than traditional methods.

2 Design of Efficient Retrieval Method for Malicious Information

2.1 Fragmentation of Malicious Information Metrics

The collection of malicious information in big data networks is the basis of retrieval. It has important significance [4, 5]. This paper uses metrics to collect and classify malicious information. Firstly, the Kullback-Liebler algorithm is used to cluster the malicious information in the big data network, namely:

$$KL = \sum_{w_i \in d} \frac{n(w_i, d)}{|d|} \log \frac{|d|}{nc/|c|} \tag{1}$$

Among them, KL represents the semantic distance between standard information and malicious information, obviously there is a situation of $KL = 0$; $n(w_i, d)$ is the number of times the malicious information w_i concept appears in the big data network d; $|d|$ represents the total number of malicious information; c represents the cluster of the algorithm Coefficient, this calculation does not do orientation analysis.

With the development of network technology, more and more big data networks adopt an open information distribution structure, which provides a development platform for the efficient retrieval process of computers [6, 7]. Therefore, this paper combines the human-computer interaction mode to simultaneously apply this efficient retrieval method. Applied to the human-computer interaction interface, the technician can feel the interference process of malicious information through the computer, and then accurately classify the retrieval criteria of the malicious information.

According to the classification form of the retrieval database, the browsing mode and the search engine are queried. During the query process, the document information in the entire big data network is sorted according to the probability of generating the malicious information model [8, 9]. Assume that in the big data network model based on human-computer interaction, each document information represents a polynomial of each term t in the information D vocabulary, then each document information is sorted according to its production probability when performing information retrieval. The probability of generating document information is calculated by the following formula:

$$sim(Q, D) = \frac{P(D|Q)}{KL/n(t|Q)} \tag{2}$$

Where $sim(Q, D)$ represents the frequency of occurrence of document information D and $Q.P(D|Q)$ represents the equivalent feature of all document data corresponding to the statistical item, and does not affect the sorting, and can be ignored in the actual calculation process. As can be seen from the formula, the hypothetical terms in each malicious information model are independent of each other. $n(t|Q)$ is the test probability of selecting document information, which can be used for the setting process of weights such as malicious document information retrieval parameters. All $P(D)$ values are the same in the calculation process of this paper, so it can be ignored in calculation.

2.2 Human-Computer Interaction Retrieval Logic Establishment

In the malicious information metric defined above, the spatial knowledge of all semantic concepts is concentrated on the level of words and sentences [10–14]. Firstly, a word-to-concept knowledge comparison table should be established based on the HNC concept node table to map and demap the malicious information features in the big data network. This paper introduces the concept of human-computer interaction and establishes the key information in the big data network information knowledge base. And its characteristics are shown in Table 1.

Table 1. Big Data Network Information Characteristics

Features	Type	Describe
frequency	Relevance	Frequency of Key Features of Malicious Information Query in Title and Text
BM25	Relevance	Query based on BM25 formula, including the relevance score of title, text and words of network information
N-gram BM25	Relevance	Query based on BM25 formula, including the relevance score of the title, text and meta-index items of network information
Edit distance	Relevance	Relevance Score of Title, Text and Editing Distance of Query Network Information
Number of incoming links	document information	Number of inbound links on Web pages
PageRank	document information	PageRank-based Web page importance is related to the number and quality of web pages'inbound links
Clicks	document information	Retrieve the number of clicks per page
BroseRank	document information	The importance of Web pages based on BroseRank score is related to the click probability of users when they browse the Web pages.
Malicious Information Assessment Value	document information	More than 80% of malicious information in web pages
Web Quality Assessment Value	document information	The Possibility of Web Pages as Low Quality Pages

Analysis Table 1 shows that the biggest advantage of introducing human-computer interaction retrieval logic is that there is no semantic ambiguity in the information concept. In the multimedia big data network, an information concept symbol corresponds to a certain semantic, which can fundamentally solve the inaccurate retrieval of malicious information. The problem. At the same time, according to the identification method of the information concept shown in Table 1, the establishment process of the retrieval logic can be realized by the clustering calculation of the concept. details as follows:

If an information concept is successfully identified, then the information retrieval process is considered as a problem of judging which document has malicious information, and only the reference value of the test data is considered, and the threshold clustering calculation is performed, and the interactive object of the malicious information is determined. You can get spatial information about any malicious information.

Assuming that the spatial coordinate of a malicious information is $N(x, y, z)$, the probability of generating the information $sim(Q, D)$ is introduced. According to the increasing arrangement of the distance values, the establishment basis of the retrieval logic is determined on the function curve, and then the spatial coordinates of any point on the curve are calculated. The spatial threshold of the malicious information is obtained, and the clustering result is automatically formed according to the threshold.

Finally, the matching process of the retrieval logic is completed through the query and clustering based on human-computer interaction. The function expression of the malicious information retrieval logic based on human-computer interaction calculated in this paper is as follows:

$$\begin{cases} N_x = \dfrac{c(sim(Q,D)+1)}{N} \\ N_y = \dfrac{1 - c \cdot sim(Q,D)}{N} \\ N_z = 1 - \left(N_x / N_y\right) \end{cases} \tag{3}$$

Where N_x, N_y and N_z represent the retrieval logic of the spatial abscissa, ordinate, and height coordinates of malicious information, respectively; M represents the semantic implied hint of malicious information in a big data network.

The human-computer interaction retrieval logic is established by referring to the HNC node concept, which provides a basis for realizing the retrieval and calculation of malicious information.

2.3 Efficient Retrieval of Malicious Information

In order to efficiently and accurately classify malicious information of big data networks, it is necessary to further search for malicious information based on the criteria of classification and the similarity threshold. Generally speaking, in a random text message in a big data network, the emergence distance of malicious information will continuously jump with the increase of density. These jumping points are the thresholds of malicious information we are looking for, and the distance between jump points is in ascending order, the corresponding threshold of malicious information can be determined by fitting the form of the incremental function. The incremental sequence of malicious information is shown in Fig. 1.

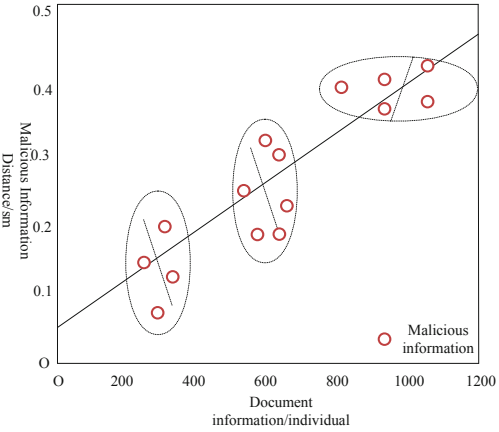

Fig. 1. Schematic diagram of the incremental sequence of malicious information

The curve of Fig. 1 is fitted by the idea of least squares method, and the problem of malicious information retrieval is transformed into the problem of finding the $f(x)$ value of the curve, so that the height fitting of the above malicious information N_z is minimized.

Suppose ξ_i is a malicious information point on $N_z|N_z \in f(x)$, where $f(x) = x_1 + x_2 + x_3 + \cdots + x_m$, then only the coefficient $x_1 + x_2 + x_3 + \cdots + x_m$ is determined, so that ξ_i is minimized, and the height fitting of N_z can be minimized.

Bring the expression of $f(x)$ into the calculation formula of ξ_i, and let the partial derivative of ξ_i to m_j be equal to 0, then get $m+1$ equations and solve m_j from the equations.

According to the actual fitting effect, it is found that it is most suitable to fit the curve with the third-order polynomial. Combine the fitting equation of the curve to solve the second derivative, and make $f(x)' = 0$, then the x value is the curve inflection point. One inflection point is the curve coefficient, which is the minimum threshold of x corresponding to the N_z minimum value.

The minimum threshold obtained above includes the noise of malicious information. In order to improve the retrieval precision, the malicious information is also subjected to noise statistics, and the interference noise is eliminated to generate an absolute model of malicious information retrieval.

The noise reduction process is calculated as follows:

$$p = \frac{1 + \xi_i}{\log d_{\max}} \tag{4}$$

Where, p represents the malicious information retrieval coefficient without noise interference; $d_{\max}$ represents the maximum value of the malicious information cluster.

After the above definition, the efficient retrieval method based on human-computer interaction is used to retrieve and deduct malicious information in a random text of a big data network. By demarcating the metrics of malicious information, the method of establishing human-computer interaction logic is determined, and the retrieval coefficient is determined. The search coefficient is denoised to ensure that the method has an efficient and accurate retrieval advantage. In the process of retrieval, considering the existence of a single malicious information node, the method uses a set value to determine any information point. When the minimum value of the semantic distance ξ_i between the document information is greater than the retrieval coefficient, the information is determined to be an isolated point, which is consistent with the retrieval logic, and realizes efficient and accurate retrieval of malicious information in the multimedia big data network.

3 Simulation

In order to evaluate the efficient retrieval method based on human-computer interaction, the malicious information retrieval test was carried out on the method, and the malicious information retrieval test system based on Jelinek-Mercer smooth model and Bayesian model was obtained from the big data network. After proper processing and

modification, the experiment is compared with the traditional malicious information retrieval method, and the malicious information retrieval effects of the two methods are counted.

3.1 Experimental Test Set

During the experiment, all the test data are the Chinese information retrieval test data set provided by TREC6. The test set includes a total of 164,811 big data network articles, all of which are from well-known big data networks, of which 1/10 are malicious information. The probability of 0.5% maliciously interferes with the multimedia big data network. All of these malicious information coexist in both Chinese and English. By changing the amount of malicious information and the length of interference, the comprehensiveness and accuracy of the search are analyzed.

3.2 Comprehensive Comparison of Malicious Information Retrieval

In order to analyze the effect of malicious information retrieval, taking recall rate as the experimental index, the traditional method and the method in this paper are used to compare the recall rate of malicious information retrieval of the two methods, and the results are shown in Fig. 2.

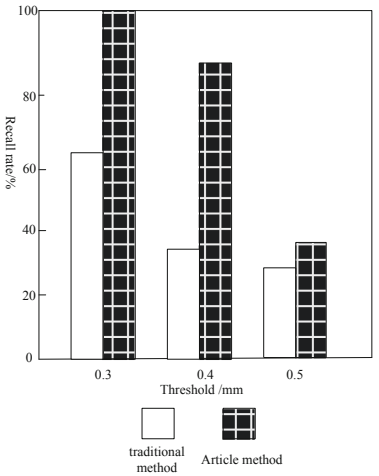

Fig. 2. Comprehensive comparison of malicious information retrieval results

The comparison results show that both methods can retrieve malicious information in the multimedia big data network, but the recall rate is different. When the threshold value is 0.3, the recall rate of traditional method is 65%, and that of this method is 100%. When the threshold is 0.4, the recall rate of traditional method is 35%, and that of this method is 88%. When the threshold value is 0.5, the recall rate of traditional method is 30%, and that of this method is 38%. Analysis of the reasons for this shows

that with the increase of the threshold, the search process requirements for information matching are also improved, the number of malicious information that meets the requirements is reduced, and the recall rate is naturally reduced. When the threshold is between 0.3 and 0.4, the method has a better effect on the expansion of malicious search terms, and the correlation is greater. The result is that the recall rate is higher than the traditional method. Therefore, in order to ensure the comprehensiveness of malicious information retrieval, it is necessary to give a suitable threshold value for retrieval quantity control before retrieval.

3.3 Malicious Information Retrieval Accuracy Comparison

In order to analyze the effect of malicious information retrieval, taking the accuracy of malicious information retrieval as the experimental index, the traditional method and the method in this paper are used to compare the malicious information retrieval of the two methods, and the results are shown in Fig. 3.

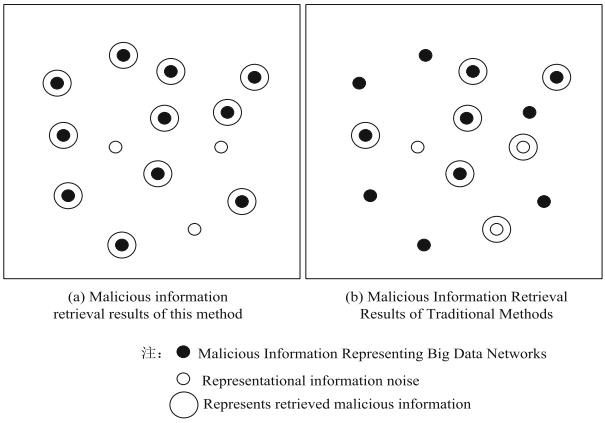

Fig. 3. Comparison of malicious information retrieval accuracy

Through the analysis of the experimental results in Fig. 3, both methods can achieve the retrieval of malicious information in the multimedia big data network, but the retrieval accuracy is different. The analysis shows that this method can accurately retrieve and mark malicious information. However, the traditional methods have some defects in the retrieval of malicious information. It can be seen from the figure that the traditional method does not retrieve edge information in the big data network, so there must be accurate retrieval of malicious information. There is also the phenomenon of false information retrieval noise, which affects the final retrieval result of malicious information. Through the above analysis, the validity of the human-computer interaction retrieval method can be explained. When searching, the malicious information in the big data network can be accurately found and marked, and the validity and relevance of the query index words are guaranteed. Precision, superior to traditional methods

4 Conclusion

This paper has carried out a comprehensive and systematic research on the malicious information retrieval method based on human-computer interaction, but there are still many shortcomings. In the future research, we will further consider the in-depth development of human-computer interaction, with different modes. Human-computer interaction behavior improves the retrieval performance of malicious information. Also consider the semantic relationship between human-computer interaction and big data network, and use these relationships to enhance the semantic representation of malicious information, and then use semantic representation relationship to comprehensively analyze and compare malicious information retrieval methods, and further improve the efficiency of retrieval methods. Improve the information environment of big data networks.

References

1. Xu, L., Weiqun, W., Zengguang, H., et al.: Human-machine interactive control method for rehabilitation robots. Chin. Sci. Inf. Sci. **25**(1), 101–103 (2018)
2. Juanjuan, Q.: Simulation of implicit feedback information retrieval for users of big data network. Comput. Simul. **36**(9), 430–433+468 (2019)
3. Wei, Z., Hongxu, H., Jing, W.: Application of Bayesian network for information retrieval. Inf. Sci. **36**(6), 136–141 (2018)
4. Yao, Z., Bo, L., Xiansheng, H., et al.: Introduction to the special topic of multimedia data processing and analysis. J. Softw. **74**(4), 59–63 (2018)
5. Qi, Z.: Research situation analysis of multimedia classrooms in china based on big Data. Econ. Res. Guide **26**(18), 165–166 (2017)
6. Agosti, M., Crestani, F., Gradenigo, G.: Towards data modelling in information retrieval. Compr. Remote Sens. **15**(6), 143–162 (2018)
7. Berit, E.A., Janne, B.C., Lars, T.: Hospital nurses' information retrieval behaviours in relation to evidence based nursing: a literature review. Health Inf. Libr. J. **5**(1), 3–23 (2018)
8. He, J., Liming, N., Zeyi, S., Zhilei, R., Weiqiang, K., Tao, Z., Xiapu, L.: ROSF: leveraging information retrieval and supervised learning for recommending code snippets. IEEE Trans. Serv. Comput. **12**(1), 34–46 (2019)
9. Lu, Mengye, Liu, Shuai: Nucleosome positioning based on generalized relative entropy. Soft. Comput. **23**(19), 9175–9188 (2018). https://doi.org/10.1007/s00500-018-3602-2
10. Jihong, L., Weiguang, Z.: Study on the innovation of network information retrieval for university library in big data era. Agric. Network Inf. **1**(4), 30–32 (2017)
11. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
12. Li, J., Li, N., Afsari, K., et al.: Integration of Building Information Modeling and Web Service Application Programming Interface for assessing building surroundings in early design stages. Build. Environ. **153**(1), 91–100 (2019)
13. Stanton, L.: Eighth Circuit Says VoIP Is Information Service, Preempts Minnesota PUC. Telecommun. Rep. **84**(17), 1,35–36 (2018)
14. Tang, Z., Srivastava, G., Liu, S.: Swarm intelligence and ant colony optimization in accounting model choices. J. Intell. Fuzzy Syst. **38**(2),1–9 (2019)