Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 1, No. 3, 2022

Design of Bad Information Filtering System for Web Pages

Ronghua Lu

Jingdezhen Ceramic University, Jiangxi, 3334403, China.

Abstract: The openness and globalization of the Internet make the information spread on the network mixed, and all kinds of advertisements, spam, reactionary, obscene, violence and other bad information seriously interfere with the normal use of the Internet by Internet users. This paper studies the workflow of the web page bad information filtering system, analyzes the functional modules of the filtering system, and designs the network data processing, text data processing and adaptive processing modules in detail.

Keywords: Web page information filtering; Network data processing; Text data processing; Chinese word segmentation technology; Text representation technology; Adaptive processing.

1. Web page bad information filtering system

Web page bad information filtering system is content-based filtering, filtering object is Chinese text-oriented web pages, filtering purpose is topic detection. Running on the LAN gateway, the system can monitor incoming web pages on the LAN, discover and filter the topic information specified by the network administrator. For example, when enterprises, schools and Internet cafes manage the bad information in the internal LAN, the system can shield information such as violence, obscenity and reaction.

The topics filtered by the system are initially generated by users, and new filtering topics can be regenerated as needed during filtering implementation. In the experiment, we selected the knowledge of sexual medicine as the filtering theme. For the filtered documents, a feedback mechanism is provided, and the finally determined group is imported into the template training library, and the user's intention is tracked and learned through genetic algorithm to generate more accurate user templates.

The web page bad information filtering system should deal with the following transactions in each stage: 1) information acquisition stage: network information filtering is based on packet capture. Information acquisition requires capturing HTTP packets in the network and parsing the packets into a text format for processing according to IP, TCP, and HTTP protocols. 2) Information representation stage: process the text obtained from the information acquisition stage, extract the keywords that can represent the characteristics of the document and calculate the weights. Since there is no obvious separator in Chinese documents, word segmentation should be carried out before feature extraction.3) Matching stage: the text representation of the unknown document is matched with the known user template (knowledge pattern). The similarity of the vector space model is used to calculate the correlation between the unknown document and the actual demand. After reaching a certain threshold, the unknown document is identified. 4) Information classification stage: Through system identification and user feedback on identification results, the documents are input into the corresponding document set to facilitate the reconstruction of user templates. 5) Knowledge mode: that is, the establishment and update of user templates. The system established a learning evolution mechanism, according to the user feedback, the user template to learn genetic algorithm, to improve the adaptive ability of the system.

2. 2.Design of Bad Information Filtering System for Web Pages

The performance of the information filtering system is mainly reflected in the accuracy of filtering. The main reason why the filtering system is unreliable is the accuracy of user templates. However, the rapid changes of knowledge and information and the instability of user requirements determine that user templates need to be updated constantly. The bad information filtering system of the web page has fully considered the improvement requirements of the system in the design, and will continuously optimize the user template in operation to improve the accuracy of filtering. The overall design of the system is shown in Figure 1.

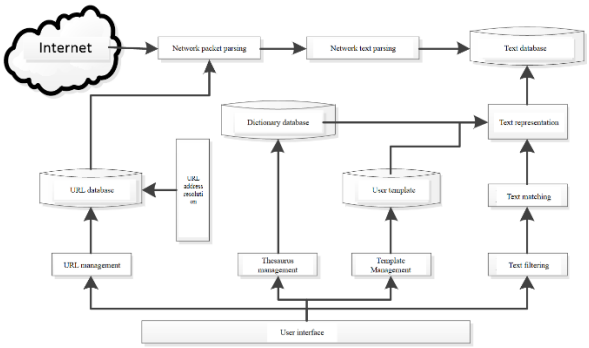

Figure 1. Design of Web page bad information filtering system

First, build the basic "dictionary database" and "user template". Dictionary database includes word segmentation database and stop word database. Stop words are generally selected as words that appear frequently in all documents, but contribute little to the topic of the document's content. The establishment of initial user template is to manually select several documents that can represent this category, and generate the vector space expression of this category document through training. Then test a batch of related and unrelated documents to determine the threshold of similarity that can match the category. The functions of each module of

the system are as follows:

1) Network packet parsing: capture Ethernet frames in the LAN, filter out TCP packets containing HTTP text information, and screen and filter pages with the same link address as those in the URL Database. In the filtering mode, a socket is set up to intercept the packet and forward the packet so that the message "This message is intercepted" is displayed on the terminal user interface. This module is also responsible for message parsing and web page reorganization of the extracted HTTP packets, and the complete web page is passed to the "Web page text parsing" module.

2) Web page text parsing: convert web pages into plain text format, mark the text in specific labels, and store it in the "text database".

3) Text representation: refer to the keyword set in the "user template", use the improved algorithm of web page text representation, perform word segmentation, and calculate the weight, and finally express the document as a vector space model.

4) Text matching: calculate the similarity between the vector expression obtained in the "text representation" and the vector expression in the "user template", find out the documents with a certain matching relationship (that is, the web pages to be filtered), and mark the evaluation results of the system.

5) Text filtering: submit users' filtering requests and deliver the evaluation results of the system.

6)Template management: According to the learning mechanism of user templates customized by users, a certain number of pages from the "URL database" will be added to the training set, and the genetic algorithm will be used for self-learning and inductive update to improve the "user template", so as to ensure the filtering performance of the system.

7)Thesaurus management: complete the dictionary improvement function. The user interface provides a management interface. You can preview the word segmentation results of a document, view the reasonable and unreasonable words in the dictionary database and the stop word database, optimize the "dictionary database", and the word segmentation results will be more accurate.

8) URL management: complete user feedback of system evaluation results and get the final results. The user interface provides an interface to view and browse the page of the existing system evaluation, give a manual evaluation, and record it in the URL Database. In addition, in view of the network information change fast, poor stability, short web page survival characteristics, the "URL database" data set time limit, from the time of entry into the database, beyond the time limit of the data will automatically be removed from the database.

The "text database" stores the text data and related attributes of all intercepted web pages, and the "URL database" records the URL link addresses and related attributes of the pages that are rated to be filtered.

3. Analysis of Key Technologies

The web page bad information filtering system is divided into three functional modules: network data processing, text data processing and adaptive processing.

Network data processing: WinPcap, a network packet capture library under Windows platform, is used to capture network data by setting the network card to a hybrid mode; Network data analysis is another key technology of network data processing. In the process of analysis, network protocol formats are Ethernet data frame, IP protocol, TCP protocol and HTTP message. HTML page reorganization is also a very important technology in web data.

Text data processing involves two important techniques. The first is Chinese word segmentation technology: for the processing of Chinese word segmentation, the web page bad information filtering system adopts some Chinese word segmentation algorithms combined with dictionary database. First, the forward maximum matching method and the reverse maximum matching method are used for word segmentation, and the results are compared. If there are different, the results are selected by the maximum probability method. The word segmentation process is shown in Figure 2:

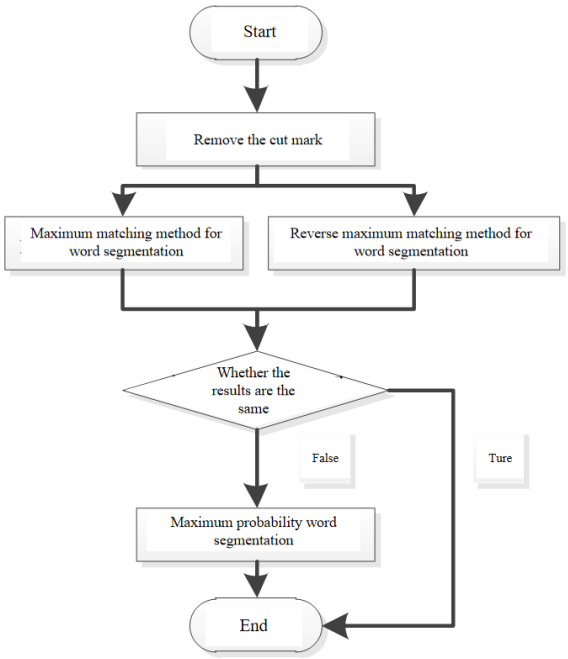

Figure 2. Word Segmentation Process

The second is text representation technology: firstly, extract feature items from documents by referring to user templates. In this paper, an improved algorithm for web page text representation is proposed, which extracts feature items from web pages. This method is accurate and fast, and can ensure the real-time performance of web page filtering. Then calculate the weight of the keyword, and use the vector space model to represent the document. Finally, the Angle between the user template vector and the document vector is calculated to determine the similarity. The similarity is compared with a certain threshold to determine the ownership of the document. The establishment of the initial user template is also a key technology in text representation. The filtering system uses the vector space model to represent the initial user template by segmentation of the training document and counting the word frequency.

The adaptive processing part is achieved through user feedback and machine learning. The web page bad information filtering system designed a human-computer interaction interface for various information feedback, such as word segmentation preview, keyword preview, sensitive data (that is, the system for the first time judged to be filtered by the web page) preview, and providesan interactive interface for users to maintain dictionary databases, stop words databases, sensitive data, etc. The core problem to be solved by any filtering system is the improvement of accuracy and intelligence, and the filtering system is no exception. The

filtering system uses the evolutionary mechanism of genetic algorithm to track and learn user interests from user templates. With the accuracy of user templates, the amount of user feedback will continue to decrease, the system will become more intelligent, and the adaptive function will become stronger and stronger.

4. Experimental analysis

We selected 48 Web documents from the documents downloaded from the Internet for testing, including 24 related documents, 4 of which are from 1KB to 12KB in size; There are 24 irrelevant documents, 4 of which are from 1KB to 12KB in size. According to the traditional method, the Web document is first transformed into ordinary text format, and the word segmentation is carried out. The matching feature words are found from the word segmentation results, and then the weight is calculated. Finally, the similarity between the Web document and the user template is calculated to obtain the document category attribution.

By identifying the label meaning of the feature words, the improved algorithm can increase the semantic understanding of the document and more accurately grasp the weight of the feature words in the document. According to the improved method, the feature word is directly taken from the user template, its position and occurrence times in the document are found, the weight calculation formula is brought in, and then the similarity between the user template and the feature word is calculated to determine its category. The precision is improved.

In this experiment, the number of relevant documents returned by the traditional algorithm is 16, and the precision ratio is 0.333; The number of related documents returned by the improved algorithm is 20, and the precision ratio is 0.417.

Acknowledgments

This work was supported by the Science and Technology Project of Jiangxi Provincial Department of Education, and the project number is GJJ170797.

References

[1] Liu Wei. Research on Content based Bad Web Page Information Filtering Method [D]. Jilin University. 2013.

[2] Sun Kai. Research on bad content web page filtering technology [D]. Liaoning University of Petrochemical Technology. 2012.

[3] Qiu Siheng. Research and Design of Wireless Internet Bad Information Filtering System [D]. Beijing University of Posts and Telecommunications. 2009.

[4] Zeng Zhiyuan, Zhang Li. An Improved Algorithm for Web Page Text Representation Based on Vector Space Model [J]. Computer Engineering. 2006 (03).

[5] Tang Jiangang, Xiong Guoping. Research and Practice of Adaptive Bad Web Page Filtering Mode [J]. Computer Engineering and Design. 2008 (20).