Contents List available at VOLKSON PRESS

New Materials and Intelligent Manufacturing (NMIM)

DOI : http://doi.org/10.26480/icnmim.01.2018.08.11

Journal Homepage: https://topicsonchemeng.org.my/

ISBN: 978-1-948012-12-6

RESEARCH AND APPLICATION OF INTERNET PUBLIC OPINION INFORMATION RETRIEVAL METHOD

Yanbo Wu, Zhongxi Hu, Dawei Xiang*

Department of Electronics Engineering, Hubei University of Police, Qiaokou, Wuhan, China.
Corresponding author Email: 2914516980@qq.com

ARTICLE DETAILS

Article History:

Received 26 June 2018
Accepted 2 July 2018
Available online 1 August 2018

ABSTRACT

This article briefly analyzes the structural system of the Internet public opinion system, and briefly introduces and analyzes the information retrieval technology and the Internet public opinion information retrieval technology involved in the Internet public opinion system, based on the existing Internet search engine. Combined with web crawler technology, RSS technology, web page text information extraction and other key technologies, corresponding research and design of network public opinion search system were conducted. Based on the feasibility of the system and the optimization of the system, the four key modules of the keyword detection module, the key personnel information monitoring module, the statistical report module and the synonym module are designed, and some experimental data of the system operation are graphically displayed. Initially achieved the discovery of Internet public opinion hotspot information and web information collection and storage.

KEYWORDS

Internet public opinion, Internet public opinion monitoring technology, Internet public opinion information retrieval.

1. INTRODUCTION

Internet public opinion means that in the social living space, all kinds of social public events happen, grow and change through online platforms. At the same time, the public thinks about the ideas and expressions generated by these public issues and the social managers. They hold general names for socio-political positions, beliefs, and socialist values. It is the sum of the manifestations and external information expressed by the majority of the public in regard to various phenomena and problems in social life, such as individual, personal beliefs, positions, opinions, attitudes and emotions. In recent years, the Internet is developing at a very rapid pace around the world. Under the circumstances of such an era, the speed of the occurrence and dissemination of Internet public opinion is extremely alarming. Compared with traditional media, online media is widely recognized by the world as a new emerging media industry after newspapers, radio and television, and the Internet has become a major new media industry. It has become one of the important platforms to respond to all kinds of public sentiment information in society.

According to the "Statistical Report on the Development Status of the 40th Internet Development in China" published by the China Internet Network Information Center, as of June 2017, China's Internet users have reached 751 million, accounting for about 54% of the country's total population [1]. The users also account for 20% of the total number of Internet users in the world. The popularity of the Internet exceeds the global average of 4.6 percentage points. Faced with such a large number of Internet users, the speed of the occurrence and dissemination of Internet public opinion can be said to be very rapid. At the same time, the social impact brought about by it has also become exceptionally large. It can be seen from the theoretical discussion of the Internet that General Secretary Xi has been in charge so far that the party and the state must develop a peaceful and stable social environment. This requires that national Internet workers need to have an excellent ability to deal with Internet sentiment, especially in the face of the Internet in the face-to-face era, it has the ability to efficiently communicate online, obtain public opinion, and guide the public on the Internet's hot spots.

The establishment of various types of government platforms and systems helps to collect public opinion. In recent years, on the basis of the practice of local network policy platforms in various provinces and municipalities in China, government agencies have gradually formed and summarized some ideas and methods for guiding Internet public opinion. The most important thing in the network public opinion monitoring system is the Internet public opinion information retrieval technology. Without the Internet public opinion information retrieval technology, the network public opinion monitoring system cannot obtain the source of information, and subsequent social network analysis and topic prediction technology cannot continue and so on. Below we describe the key technologies such as information retrieval, web crawler technology, RSS technology, and extraction of web page text information, which are involved in the design of "Internet Public Opinion Search System."

2. INFORMATION RETRIEVAL

Information retrieval refers to the method by which users search for information on the Internet. Information retrieval can be divided into two dimensions: From a small perspective, information retrieval refers to information query. Information query refers to the process of finding out the information that is useful to users by using certain methods from the collection of information and using different search tools according to their own needs. From a large perspective, information retrieval refers to processing, organizing and organizing all information in the Internet platform in a certain manner and storing the integrated information, and then reusing relevant information according to the specific needs of different users. Accurately find and display the process from the integrated information. Small-scale information retrieval includes the following three implications:

1) Get user's query requirements;
2) Use of information retrieval methods and techniques;
3) to achieve the user's information acquisition needs.

3. INTERNET PUBLIC INFORMATION RETRIEVAL TECHNOLOGY

The network public opinion monitoring system is a general term for the

computer detection of the network public opinion. The network public opinion monitoring system involves many subsystems, such as: network public opinion analysis system, network information acquisition system, acquisition information analysis system and search engine data management system. The core technologies involved in the network public opinion monitoring system can be divided into network public opinion information retrieval technology, social network public opinion analysis technology and topic prediction technology from a general perspective. The network public opinion information retrieval technology can be divided into network information collection technology and network text information extraction technology.

3.1 Network Information Collection Technology

There are many kinds of public opinion monitoring systems nowadays, which mainly use meta search technology and web crawler technology to design a network public information collection program. Below we introduce meta search technology and web crawler technology.

The meta search engine is an engine that helps users select from other source search engines through a unified user platform interface. Therefore, the meta search engine has the title of the mother of search engines. The general description of the metasearch program is a search program that integrates, invokes, controls, and optimizes multiple independent search engines. The search engines that use metasearch technology include: 360 search, MEZW search, search, and so on. Compared to the meta search engine, an independent search engine that can be used, we usually call it a source search engine or a search resource program such as: Google, Baidu, Sogou, Yahoo, etc. The core of meta search engine is meta search technology. Metasearch technology does not need to include all Internet information to obtain the desired query results, and the complexity is relatively low.

Web crawler refers to a program or script that can automatically grab network information from web pages according to certain rules [2]. They are commonly used by Internet search engines or other similar websites. The web crawler can automatically collect all the web page content information it can access. Through this method, the website web page content information and retrieval methods are obtained and updated. From a functional point of view, the web crawler can generally be divided into three parts: 1 information collection, 2 information processing, and 3 information storage. The web crawler can be divided into two types, the traditional crawler and the focused crawler. The workflow of the traditional crawler is to start from the URL of one or more initial web pages, to obtain the URL on the web page of the initial web site, and to continuously crawl the web site. In the web page process, the new URL extracted from the current web page is placed in a pre-designed queue of the program until the stop condition of the program is satisfied, and then the crawler program stops working. The process of focusing on reptiles is more complex than that of traditional reptiles.

Focused reptiles need to filter links that are irrelevant to topics based on certain web page analysis algorithms, retain valid links, and place valid links waiting to be crawled. The URL program is pre-designed in the queue. Then, the focused crawler will select the URL of the webpage to be crawled from the pre-designed queue of the URL program to be crawled according to certain search rules, and repeat the above process until the program stops working when a certain termination condition of the program is reached. In addition, all webpage information crawled by the crawler will be stored by the program, then filtered, analyzed, and indexed for later user query and retrieval; the analysis results obtained in this process will focus on crawlers. Follow-up crawling feedback and guidance. The differences and advantages of the meta search technology and the web crawler technology are shown in Tables 1 and 2, respectively.

Table 1: Differences between metasearch technology and web crawler technology

Meta-search technology and web crawler work difference	
Metasearch technology	Web crawler technology
By sending user-submitted search statements to multiple member search engines, multiple search engines can be accessed.	Reptiles are mainly responsible for traversing websites. Web pages crawled by reptiles will be stored by the system for certain analysis, filtering, and indexing for later querying and retrieval.

Table 2: Advantages of metasearch technology and web crawler technology

The advantages of meta search technology and web crawler technology	
Metasearch technology	Web crawler technology
It can access the web content of the deep network that the web crawler cannot climb to. Metasearch technology has the advantages of simple operation, low complexity, and large search breadth.	The crawler call is invoked on the Web server. How to use it on a regular basis, these crawlers can be used. Therefore, reptiles have the advantages of: for the web crawling can be biased for observation, easy to call.

It can access the web content of the deep network that the web crawler cannot climb to. Metasearch technology has the advantages of simple operation, low complexity, and large search breadth.

The crawler call is invoked on the Web server. How to use it on a regular basis, these crawlers can be used. Therefore, reptiles have the advantages of: for the web crawling can be partial observation, easy to call.

3.2 Network text information extraction technology

Text information extraction technology is a technology that extracts specific required information from text data information. Text data is composed of specific units, such as sentences, paragraphs, and chapters. Text information is composed of small, specific units, such as words, words, phrases, sentences, paragraphs, or a combination of these specific units. The subject noun, person's name, and location name in the extracted text information are all extracted from the text information. In addition, other types of information can also be used as textual information extraction techniques. Currently, web page text information extraction technology is mainly based on template extraction and web-based structure information.

4. NETWORK PUBLIC OPINION SEARCH SYSTEM STRUCTURE DESIGN

4.1 General Framework of Internet Public Opinion System

The overall framework of the Internet public opinion system envisaged in this article can be divided into public opinion information acquisition layers. The process of system work is illustrated in the three levels of the public information processing layer and the public information report form return layer, as shown in Figure 1.

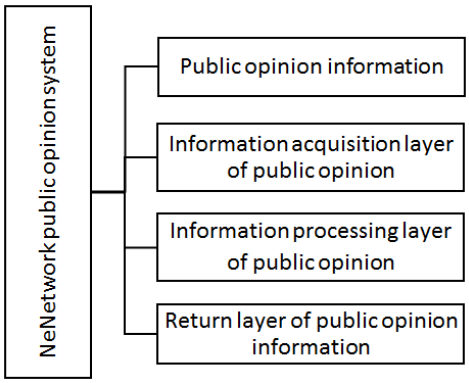

Figure 1: The overall framework of the Internet public opinion system

4.2 Details design and application of sensation system

This paper designs a keyword detection module, a key personnel information monitoring module, a statistical reporting module, and a synonym module to implement a network public opinion search system.

4.2.1 Keyword Detection Module

The keyword detection module uses the search engine technology to search for the preset keywords, and carries out URL automatic deduplication, information element collection, keyword extraction, partition storage, full-text retrieval, and other processing to achieve the monitoring of Internet sensitive information. (Including Weibo, forums, post bar, website pages, video images). The module retrieves network information from two methods: metasearch technology and web crawler technology. Through metasearch technology, the ability to search the Internet for information is maximized. The web crawler technology uses the web crawler program Nutch to conduct in-depth excavation of user-requested website page information and retrieve and match webpage keyword information [3,4].

4.2.2 Key personnel information monitoring module

Key personnel information monitoring is conducted mainly through the module program to pre-set key personnel information (such as: webmail files, microblogging, blog, QQ number, forums, paste it, QQ space, Internet IP address, etc.) and network activity automatically Detection, automatic tracking and monitoring of important targets. The system uses RSS technology (as I described in detail below) to track the updates of key social networking platforms such as QQ, Weibo, blogs, and Qzone, and to open host operating system version information and hosts [5]. Computer information such as port numbers is obtained by integrating NMAP open-source scanning tools.

The network hotspot information discovery algorithm designed by the lyric secondary clustering model analyzes the elements such as post title, author information, release date, reading amount, and response amount, and performs automatic deduplication, data classification, and information storage into the database to achieve monitoring of Weibo, blogs, forums, post bars and other posting platforms [6]. The program parses the source code of the posting page, obtains the reading amount, the reply amount of the posting, and responds to the heat information such as the sensitive content and the number of postings, performs the heat calculation, and realizes the hot spot discovery function of the forum. Improves the relevance matching of sensational webpage information, and makes the analysis data of the collected web public information closer to the facts, with a small degree of error.

4.2.3 Statistical Reporting Module

The statistical report module generates a report on the information stored in the database after processing by the search engine. The user can extract the generated report and process the information content, information upload, and other functions.

4.2.4 Synonym Module

The first step in solving this problem is to determine the scope of the topic. To give two simple examples, when people search for topics such as "terrorism incidents" and "hit rob events" or "social networks" and "social networks", although there are subtle differences between different words, they are in reality. In the context, people often do not distinguish between their specific meanings and use them in a broad sense. If they only analyze the content under the topics of “terrorism incidents” or “social networks”, they are likely to be only small samples, affecting the entire situation. The validity of the analysis. Therefore, for the topics entered by the user, it is necessary to find out other words that express an approximate meaning, that is, synonyms, so as to make the sample more general.

In order to achieve this goal, it is necessary to add a synonym table outside of the topic information table. After the user enters a specific topic, he or she will not directly query the topic information table, but will first transfer to the synonym table and get it. Groups of topics, including known topics, return a list, then inquire into the topic information table, and finally summarize the results. If there is no synonym for this topic, a list of length 1 is returned. Pseudocode implementation:

List (topic){List list=getHibernateTemplate().find;
If{ adds the specified topic to the list;}}
List topiclist= List selectTopic(topic);
For() {// delete the same field;
if(topiclist.get(i)==topiclist.get(i+1)){
Toplist.remove(i+1);i-=1;}}
For () {in order to obtain information corresponding to each topic in the topic information table;
For() {Object[][] object; sums the corresponding information and stores it in object[][];}}
For(){List sumlist;Sumlist.add(i, object[]);}

4.2.5 RSS Technology

In the era of the Internet entering the web2.0 era, RSS technology has been widely used. RSS technology refers to the way of using XML (Extensible Markup Language) to distribute content collected on a certain website to many other websites and is based on the current environment. The most common and widely used technology for XML standards. RSS technology is mainly used in the process of sharing information between websites.

```
<? xmlversion= " 1.0 " encoding= " gb2312 " ? >
<rssversion= " 2.0 " >
<channel>
    <title>Cool Blog 我的博客日志</ title>
    <link>http: //blog.sina.com.cn/eaglehaoren</link>
    <description>博客日志系统</ description> <language>zh-cn</ language>
    <generator>Dreamweaver8</ generator>
    <item>
        <title>这是某篇日志的标题
    </ title><link>http: //blog.sina.com.cn/ eaglehaoren / showlog.asp? b-id=1
    </ link>
        <category>前台技术</ category>
        <description>这篇日志的内容</ description>
        <pubDate>2017-3-20 12: 10: 25</ pubDate>
</ item>
</ channel>
</ rss>
```

In essence, RSS is a specification of process data exchange for sharing information between websites. An RSS file is a subset of standard XML data, and RSS files are often suffixed with rss and xml files. Therefore, the RSS micro-blog and post-packet parsing interfaces are implemented by extracting the node information in the RSS Feeds. This operation is also called the parsing of the XML document. For example, for example:

Among them, the specific basic information of an article is represented by <pubDate>. We define "XMLOperater" as an interface for manipulating XML documents. At the same time, we use "XMLOperater" to define two methods of loading and parsing XML documents. RSSBlog microblogging, paste bar file loading and parsing interface contains such as: posting user name, user IP address, a detailed list of the log, the title of the article, IE link, article content information, release time, comment information and a series of basic attributes Information, including the UML diagram shown in Figure 2.

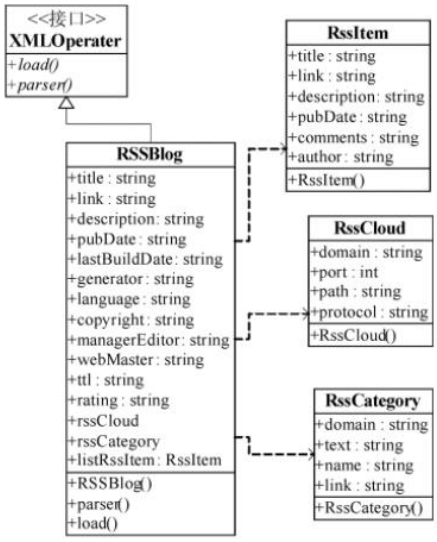

Figure 2: XMLOperater interface UML diagram

5. EXPERIMENTAL DATA

We use the commonly used search engines, RSS technologies, and online text information extraction techniques in the Internet to achieve the search function of public opinion information. We conduct experiments through Baidu and display the experimental results in the form of a table.

The sensation retrieval function of this article is based on commonly used search engines in the Internet. Therefore, our search page is also Baidu. We now search for the keyword “trade war” through Baidu, and analyze the URL characteristics of search engines as follows: https://www.baidu.com/s?tn=58025142_5_oem_dg&rtt=1&bsst=1&wd=%C3%B3%D2%D7%D5%BD&origin=ps

The keyword tag: wd= keyword URL code, Baidu uses GBK code, so in this example : ” %C3%B3%D2%D7%D5%BD” is the GBK code for the keyword "trade war".

Below we show some of the experimental data, as shown in the following table 3.

Table 3: Some experimental data sheets

Key words	URL	Time
trade war	http://news.ifeng.com/a/20180418/57667543_0.shtml	2018/04/18
	http://finance.ifeng.com/a/20180418/16127947_0.shtml	2018/04/18
	http://news.sina.com.cn/w/2018-04-19/doc-ifyuwqfa4161653.shtml	2018/04/19
	https://finance.qq.com/a/20180419/011029.htm	2018/04/19
	http://business.sohu.com/20180418/n535199363.shtml	2018/04/18
	http://finance.ifeng.com/a/20180418/16129992_0.shtml	2018/04/18
	…….	…….

6. SUMMARY

In this paper, the basic concepts of information retrieval involved in the network public opinion monitoring system, meta search technology and web crawler technology in the network public information retrieval system are introduced in detail. In addition, by using related technologies such as commonly used search engines, RSS technologies, and online text information extraction in the Internet, we have initially achieved the collection, analysis, and integration of specific information on the Internet, and used it to retrieve and analyze data on the public opinion of the Internet and to understand the public opinions. Changes provide strong technical support.

In the rapidly developing world of this network, online public opinion guidance has become an important means to maintain the harmony and stability of the social life environment. The related research on network public opinion monitoring technology has increasingly attracted the attention and favor of the majority of researchers. At the same time, because the Internet has become the idea of people. In order to build a harmonious and stable network environment and a social environment, information gathering places are based on the Internet public opinion monitoring technology and have very important practical significance.

However, due to the fact that China’s computer development is not a small gap compared to other developed countries in the world, the Internet public opinion is a relatively new research direction in the computer field and an effective and comprehensive Internet public opinion is implemented in the national Internet environment. There are still many problems in information monitoring and guidance that need to be researched and practiced by researchers. Therefore, we hope that relevant researchers can obtain certain research ideas and methods in the preliminary summary of this article and the application examples proposed.

ACKNOWLEDGEMENT

It is a project supported by Hubei Provincial Humanities and Social Science Key Research Base Hubei Police Academy Social Security Governance Research Center Project(Funded) Project, Hubei Provincial Department of Education Research Project(B2017 240), Key Scientific Research Projects of Hubei Police Academy(2015 ZD009).

REFERENCES

[1] Achina Internet Network Information Center 2017. The 40th Statistical Report on Internet Development in China [R]. Beijing: CNNIC.

[2] Qreview of focused reptile technology - Primrose column - Blog Channel - PROG3.COM

[3] Khare, B., Cutting, D., Sitaker, K., Rifkin, A. 2004. Nuthc: A Flexible and Scalable Open-Source Web Search Engine [R]. CommerceNet Labs Technical Report.

[4] Jianfeng, C. 2009. Nutch's study of Chinese issues [J]. Research and Development, 61.

[5] Dongping, D., Dongbiao, G. 2011. Implementing RSS Subscription Service Using C# [J]. Computer and Modernization, (3), 140-142.

[6] Wei, E., Xin, X. 2009. Cluster-based hot discovery and analysis of Internet public opinion [J]. Intelligence Analysis and Research, 3.