The design and implementation of opinion extraction system based on Distributed network

Shi Zhenquan[1,2,a], Chen Shiping[1]

[1]Business School, University of Shanghai for Science and Technology, Shanghai 200093, China

[2]Nantong University, Nantong, Jiangsu 226019, China

[a]email: szq@ntu.edu.cn

Keywords: opinion extraction system; .Net Remoting; Reptiles; Lucene; In-memory database

Abstract. With the rapid growth of Internet users, how to quickly find the hot spots and sensitive topics of forums, microblogging and other social networking sites, to achieve forecast and early warning network of public opinion, it's a huge technical challenge to the researchers. In this paper, Redis memory database has been used to do the re-processing of the URL. Through .Net Remoting, the technology of remote communication mechanism, and the index and retrieval framework provided by Lucene, the Web-based distributed network public opinion extraction system has been achieved. The experiments show that the system can effectively improve the quality and efficiency of the Web for network public opinion extraction, and the implementation of the system provides a better solution for network monitoring public opinion.

Introduction

With the rapid growth of Internet users, the network has become the main channel for people to obtain and publish information, and the impact of network information on public opinion is growing. Faced with a huge amount of information of the ever-changing Internet world , how to collect and analyze the information of network public opinion quickly, find hot spots and sensitive topics, realize tracking analysis of topics and forecasting network public opinion warning, timely and effectively response to the dynamic changes in the network of public opinion, support and guarantee the correct guidance of public opinion, it's a huge technical challenge for the researchers, but also opens up a broad research space. In this context, this paper presents a Web-based distributed network public opinion extraction program, through .Net Remoting technology to build distributed network-based public opinion extraction system, using keywords matching manner to grab information of forums, microblogging and other social networking sites, using Redis memory database to improve the efficiency of the extraction system, in order to effectively improve the ability to control and monitor the efficiency of the network information. We use the Lucene to search, index, analysis and query full-text.

System Analysis and Design

The overall framework of the whole system is divided into three parts: the task allocation server (server-side), distributed web crawler end (client) and measure public opinion analysis monitoring end (application side). The overall architecture is shown in Figure 1.

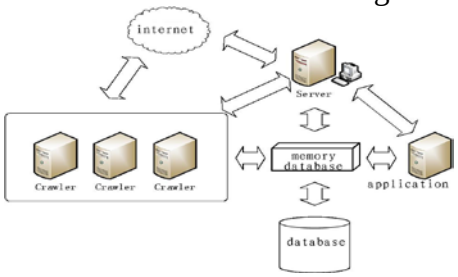

Fig 1 design of system ensemble structure

The main server is responsible for communications infrastructure set up, as a distributed

system .Net Remoting is the major technical support. The end of the URL is also responsible for seed analysis, and setting crawl depth, achieving continuous new URL address as a task, and then allocating these bands reptiles crawl URL to each node, so as to collaborate their work.

Client server broadcasts news according to the server side, extracts the mission by themselves to the server side, and then works for web crawling. And relying on Redis memory database technology, it reasonably deposits total database system deployment. The main flow of clients reptiles are as follows:

(1) According to the needs, predefine seed URL;

(2) Insert seed URL at the tail of the queue to be crawled;

(3) Reptiles extract URL from the URL (FIFO) then analyzed by DNS so as to obtain the desired IP address;

(4) Web downloader downloads web content by the use of web page downloading IP and relative paths;

(5) Store the local downloaded web page to the database, which will be prepared to generate the index based on these data;

(6) To avoid duplication crawl, quickly compare these in memory database, do the re-treatment to those URL which have been crawled;

(7) Store the URL to be crawled in the queue;

(8) Web Downloader extracts URL which is to be crawled in the queue to download the work, as long as the URL to be crawled queue is not empty, the cycle of operation, repeat steps from 3 to 7.

Application side is main to establish index data, supporting the search function, and finally to present to the user in form of public opinion briefs.

Function Realization

The entire system architecture is built on Visual Studio 2010 and SQL2008 platform, using C # language crawlers, re-processing of the URL by using Redis memory database. Via remote communication mechanism .Net Remoting technology and Lucene, it provides indexing and retrieval framework to achieve based forum for distributed, microblogging and other social networking sites Web network public opinion extraction system. It effectively improves the extraction efficiency, and improves the system control network information.

Realize Main System Framework with .Net Remoting Technology：

(1) The server-side register channel

The system for the deployment of reptiles' clients uses TcpChannel types of channels, HttpChannel type of channel is used for public opinion analysis and monitoring applications for end .Only the channel can achieve cross-domain communications.

(2) Publish a remote object

Remote object obtained by a client is not the actual object. This is just a quote of the actual object. Remoting object passed needs to inherit MarshalByRefObject. The following is the definition of the system mainly remote object class:

Public class Queen: MarshalByRefObject, Interface_Common.IQueen.

If the remote object needs to use an object, such as a class or structure, then you need to mark the class or structure serialization attributes. That is, adding attribute of the Serializable before the definition of the class.

(3) Registered remote object

After the channel is registered, the first to register the remote object, in order to further activate the remote object. Because there are multiple active modes, Singleton activation mode has been chosen in this system, so that it is easy to maintain URL task queue.

(4) Cancellation of the channel

Once register Remoting channel successfully, they will automatically open channel monitoring. If you want to turn off Remoting service, you need to close the channel monitor and logout the channel, otherwise it will cause that this channel can not accept client requests, while the channel

still exists. When registering this channel again, it will throw an unusual message. And when network has problems, it can also deal with the possible problems by cancellating the operation of the channel.

According to public opinion monitoring processes, as shown in Figure 2, the system is divided into three following modules: information collection module (reptiles information extraction), task processing module (task assignment, URL to weight), indexing module, retrieval module, and public opinion analysis module.

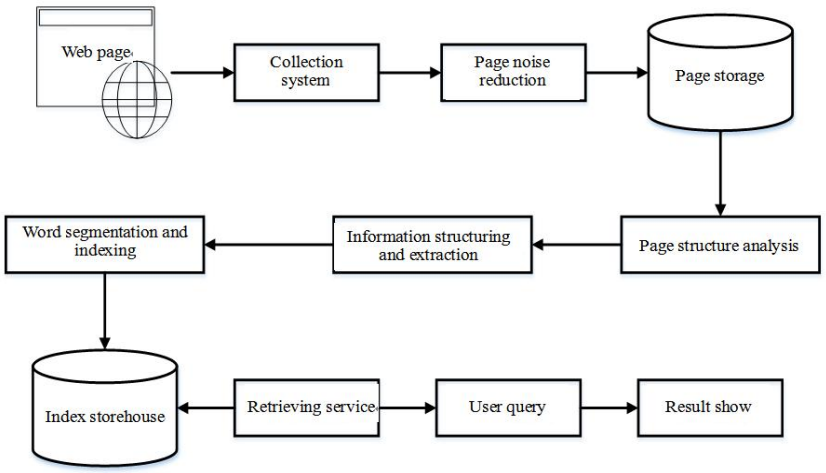

Fig 2 monitor procedure

System Task acquisition is to get the HTML code of the page through HttpWebRequest methods provided by C # , and then remove the noise in the HTML code to extract the information and determine whether the next client need further extraction work. If the page needs to crawl, it should be added to the list of pages to crawl URL.

If the URL task queue is not empty, then the server broadcasts a message to inform clients assigned to crawl. Mechanism of broadcast events is subscribed through time to achieve this process which is based on the client machine event subscription service operation mechanism.

Performance Detection and Test Results

(1) Redis memory database performance test

In this system, task processing module is using Redis memory database technology to achieve URL reduplication. Now take a test for Redis memory database performance. With 10 million records as basis, and under the different situations of the amount of concurrent clients, the testing analysis, respectively, for the three operating statements for performance is shown in Figure 3. The execution speed of the statement under Redis is increasing with the number of concurrent, and its speed gradually goes stable. When the number of concurrent reached 100, its speed is decreased as to the number of concurrent 50, Therefore, on the concurrent services at this end of the hardware platform, clients achieve maximum throughput is 50.

(2) Performance Test distributed crawler

The system performance testing uses a control factor reptile law. Under the conditions of machine performance and network conditions, the desired amount of information extracted separately for 300, 900 and 1800 of three experiments, are deployed in a stand-alone, 2, 4 and 8, client information to complete the set of different experiments carried out to fetch, fetching the key contrast total time consumed by the experimental data results, which are shown in Figure 4.

It can be clearly found from the figure that when the number of extracted messages is certain, the time consumed for reptiles crawl decreases with the increasing number of clients deployment, which fully shows that the system can effectively improve the deployment of distributed crawler caught take performance.

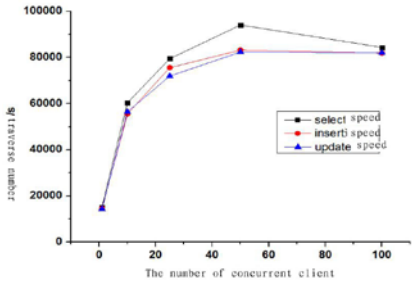

Fig 3 Redis test line chart

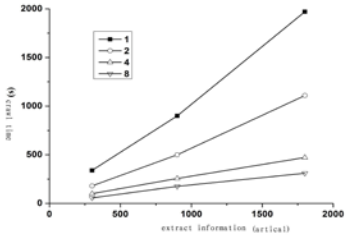

Fig 4 Performance test result

(3) Reptiles crawl clients achieve

Client first registeres the channel in order to connect with the server. Once registered, it will prompt the communication status of the client and the server. Once the server worked, it will carry out the task of processing operations, including the allocation of tasks to the client. When the client gets the task, it begins to do web crawling, and will display through a GridView to present the result in real time to crawl through the progress of the current page progress bar.

Conclusion

In this paper, a distributed monitoring system based on public opinion conducted research, analyzes the problem of main communication architecture to solve the data, commands, remote delivery with the use of .Net Remoting technology systems, make the operation of the data efficient and reliable by using the system in a variety of Redis database , and be easy to solve the problem of data integrity and consistency, and effectively improve the quality of URL reduplication. In addition, it highlights the important role of Lucene, reptiles, and other related technologies. The experiments show that the system can effectively improve the quality and efficiency of the Web network public opinion extraction. The realization of this system provides a better solution to improve the quality of Web network public opinion information analysis which is widely used in national security, national economy and information services.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61171132), Applying Study Foundation of Nantong (No. BK2012001), Nantong technology platform projects (CP2013001).

References

[1] Shkapenyuk V, Suel T. Design and implementation of a high-performance distributed web crawler[C]//Data Engineering, 2002. Proceedings. 18th International Conference on. IEEE, 2002: 357-368.

[2] Arasu A, Cho J, Garcia-Molina H, et al. Searching the web[J]. ACM Transactions on Internet Technology (TOIT), 2001, 1(1): 2-43.

[3] Thelwall M. A web crawler design for data mining[J]. Journal of Information Science, 2001, 27(5): 319-325.

[4] Zhao Ke, Lu Peng, Li Yongqiang. Search engine is designed and implemented based on the Lucene [J]. Computer Engineering, 2011,16: 39-41.

[5] Zangdong Song, Vincent Garonne, Sun gongxing. The performance analysis for a large scale distributed application system [J]. Computer Engineering, 2012,24: 37-41.