Research and Application of Content Collection Control in Enhanced Search Engine

Hong Lin
Network and Information Center
North China Electric Power University
Beijing, China
e-mail: linh@ncepu.edu.cn

Yajuan Sun
Network and Information Center
North China Electric Power University
Beijing, China
e-mail: syj@ncepu.edu.cn

Baohui Wang
Software School
Beihang University
Beijing, China
e-mail: wangbh@ncepu.edu.cn

Zhirou Zhang
Network and Information Center
North China Electric Power University
Beijing, China
e-mail: zhangzr@ncepu.edu.cn

***Abstract*—collecting the content users need safely, accurately and high-efficiently, is a key technique in Enhanced Search Engine. In this paper, for enhanced search engine in complex environment, how to make full use of resource, and improve search quality and response speed, with reasonable control strategy are researched. A content collection control model for enhanced search engine in environment of Three Network Fusion is proposed, and a content collection prototype system of enhanced search engine is used to verify.**

Keywords- Enhanced Search Engine; Streaming Media Transmission; Content Collection Control; Topic Collection Strategy

I. QUALITY CONTROL OF STREAMING MEDIA TRANSMISSION

"Key technology research and demonstration for the enhanced search engine" is a subject for National Science and Technology Support Program, which number is 2011BAH11B01. The topic is to study the key technology as real-time multimedia information collection, security filters, audio and video index, results focus, collaborative recommendation, collaborative optimization enhanced search engine, in order to achieve multimedia information fusion, focusing, filtering, recommendation, supporting the three-screen integration services for mobile phone, computer and TV, supporting safe and reliable search service.

The key to improve the quality of streaming media transmission is good QoS of system. The methods to Improve QoS of streaming media include congestion control, error control, and cache mechanism.

Control systems are established with UDP respectively, based on RTCP (RTP Control Protocol)/Mean RTP (Real-time Transport Protocol) Protocol Group, whose work flow is following: Firstly, client sends audio-video playing request to server with RTCP protocol and builds client cache; after received the request, server compress the audio-video files that are collected or stored in server with relative parameters negotiated in QoS module, or implement rate shaping with feedback information in the course of transmission to have these files adapt to transmit in networks. Then the server transmits the compressed RTP data packet to the client and the client receive them into cache, and then recombine and play them.

Multiple protocols need to be supported in content collection of streaming media. Streaming media transmission and download control are difficult in technique. In the environment of Three Network Fusion, multimedia is main collection content for enhanced search engine, so having good collection control for streaming media is basic need for the system control.

II. GLOBAL PROCESS WORKFLOW ANALYSIS OF ENHANCED SEARCH ENGINE

Enhanced search engine cover all functions of regular Internet search engines and have higher requirements in search environment (complex network environment in Three Network Fusion), search content(including document and multimedia), real-time performance (real-time TV program, news, and so on), and have more users covering not only traditional internet users but also mobile users and TV users.

Enhanced Search Engine's process Workflow is divided into 3 steps: content crawling, content arrangement, and content display. They also can be divided into the following steps: content crawling, content index, content enquiry, enquired content caching, enquired content transmission and display in detail [1]. Enhanced Search Engine's process Workflow sees also Fig. 1

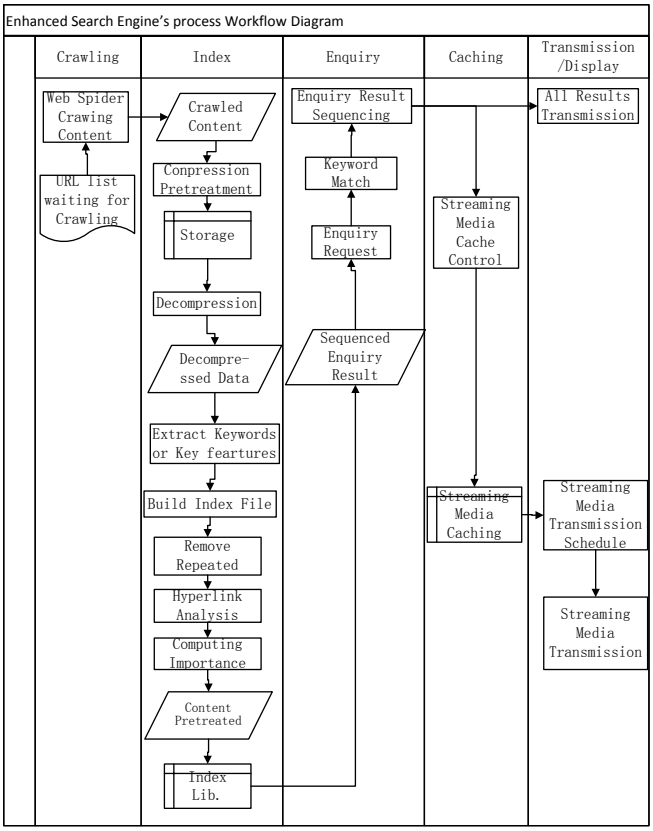

Figure 1. Enhanced Search Engine's process Workflow Diagram

Content collection control is used in content crawling mainly. Crawling module takes different collection strategy to optimize download with the preset or feedback control information. The feedback information includes sequencing, streaming media transmission, and analysis result of user's click behavior.

III. CONTENT COLLECTION CONTROL MODEL

The content collection control model is divided into 2 parts: one is artificial control that is manual controlling collection with real-time information on the information display of console; the other is automatic control that dynamically and instantaneously collects information with all kinds of sensors and detectors to send to all kinds of controllers (valves). These controllers regulate data in data flow pipelines [2]. The Content Collection Control Model of Enhanced Search Engine sees Fig 2.

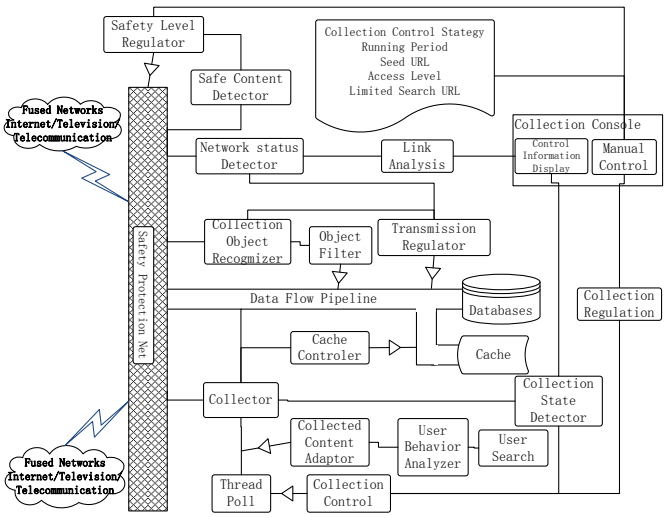

Figure 2. Content Collection Control Model of Enhanced Search Engine

Functions of parts in control model are following:

Safety Level Regulator: It regulates safety level with safety level set on the console and content safety state detected by safety content detector. The higher safety level is, the more connections are filtered out.

Safety Content Detector: It detects whether the content from networks meet the requirement of content collection safety or not, with predefined safety filter rules.

Network Status Detector: It detects network distance and network status and send the detected results to link analyzer and transmission controller to control transmission. For the content whose quantity of bytes are little (such as texts, pictures) special transmission control isn't needed, but for streaming media whose quantity of bytes are very large special transmission control is needed to guarantee the efficiency of content collection.

Collection Object Recognizer: It recognizes different collection object, especially large-size multimedia and streaming media files, with different collection strategy or algorithm, and filter out the needless content.

Cache Controller: It takes specific cache method to control cache with different cache strategy.

Collector: It collects the content and sends feedback information, such as collecting status and so on, to console and collection controller. Collection controller dispatch collection tasks with collection status control thread pool.

User Behavior Analyzer: It analyzes users' search behavior and meet their individual search requirement with different collection methods in collection content adaption.

Collection Console: It has two parts which are information display and manual control. Operators can predefine running period, limited collection URL, and so on, and manual regulate collection content setting, such as safety level, configure files, and so on, with real-time running status of collection.

IV. CONTENT COLLECTION CONTROL SYSTEM ARCHITECTURE

The logical architecture of Enhanced Search Engine includes: Terminal Showing Layer, Distribution Access Layer, Business Service Layer, Running Management Layer, Search Engine Kernel Layer, and Data Storage Layer. As a part of the search engine, content collection control subsystem which belong to kernel layer of search engine, is in charge of controlling content collection, whose control object is content collector and control information comes from feedback information of enquiry part, streaming media transmission and user behavior analysis results. Safety control and performance control of content collection lead the kernel layer of search engine, so collection strategy need take safety and performance into account [3]. Content collection control logical architecture of enhanced search engine see Fig. 3.

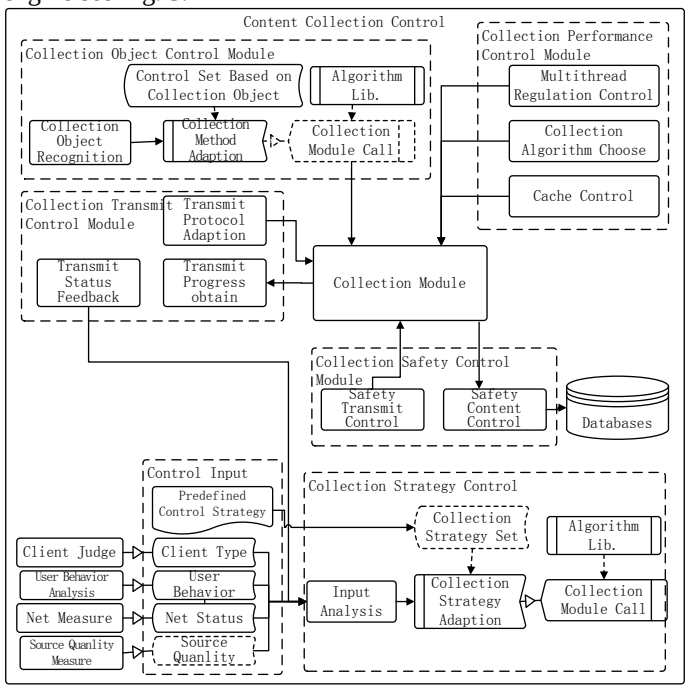

Figure 3. Content Collection Control Logical Architecture of Enhanced Search Engine

In the perspective of logical architecture, the content collection control part of Enhanced Search Engine is divided into 5 parts: collection object control module, collection transmission control module, collection strategy control module, collection safety control module and collection performance control module. There are open-loop control and close-loop control in these parts which means collection behavior can be interfered with human or regulated automatically.

Feedback information from network distance measure module, streaming media transmission control module, cache control module, and user behavior analysis module, is input to content collection control module dynamically, and the control module regulate collection behavior with these information to form a whole close-loop content collection control process.

V. CONTENT COLLECTION CONTROL SYSTEM DESIGN

A. Design of Control Module Based on Collection Strategy

Content Collection Control System is subsystem through the search engine, whose 5 parts are not independent but interconnected to control collection from different hands. Detailed design of control module based on collection strategy is illustrated as an example, whose detailed architecture see Fig 4.

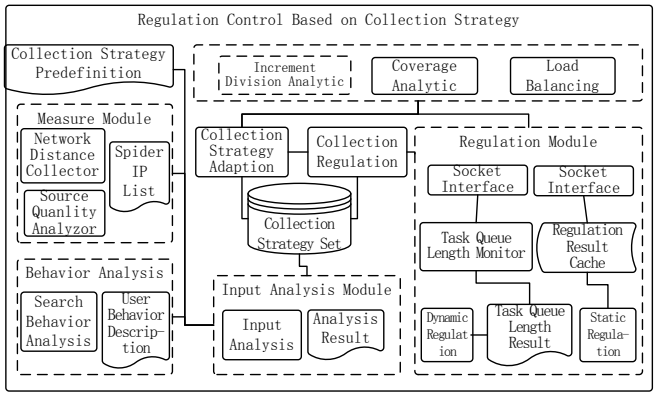

Figure 4. Architecture Scheme of Regulation Control Module Based on Collection Strategy

Control module based on collection strategy firstly reads static control information from predefined strategy control description file and gets dynamic control input from measure module and search behavior analysis module, which include network distance and source quality. Then these are analyzed by input analysis module to get collection strategies. With collection strategy adaption, a proper strategy is obtained, which is sent to collection module by collection regulation. Task regulation is in charge of regulating collection tasks.

B. Implement of Control Based on Subject Collection Strategy

The Content Collection Based on Subject is collecting relative content in a specific area in the networks. In the perspective of implement, Content Collection Based on Subject includes text classification, cluster, data mining and some relative techniques, which captures the relative subject content with subject analysis to improve search accuracy, reduce search engine's occupancy for network resource, shorten the update period of web database, and meet users' special requirements [4]. The course of Content Collection Based on Subject in Enhanced Search Engine sees Fig 5.

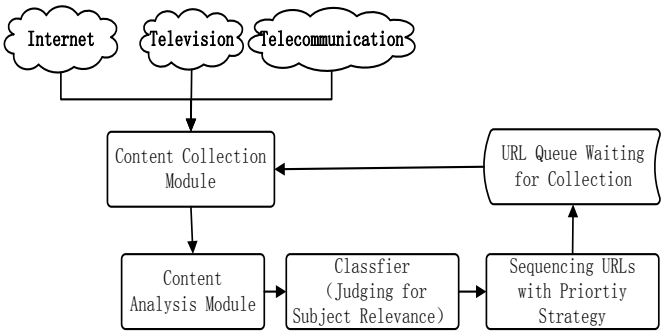

Figure 5. Content Collection Based on Subject

When a proper strategy been obtained with strategy adaption, relative content collection module is called with the strategy to send collected content to content analysis module. Analysis module analyze and obtain the content relate to the subject with classifier and sequence URL of these content with specific priority order to form a URL queue waiting for collection. Content collection module collect content with the queue [5].

Classifier is mainly used to judge relevance of collected content and determine whether expansion collection is needed or not [6]. We use Web-oriented Content Classification Subject crawler program- Focused Crawler which is proposed by Chakrabarti etc.in 1999, which uses membership function of hierarchy text classifier as web relevance function and forecast the subject of crawled web with father link and brother link comprehensively. With the information of brother link, Focused Crawler uses HITS algorithm with subject weighting to distinguish web efficiently.

VI. SYSTEM VERIFICATION

Content Collection Control prototype system of Enhanced Search Engine meet expected requirement, in which automatic recognition of collection object, automatic adaption of collection strategy and safety control all can meet the requirement of system design. Now the system runs good in test circumstance. Console can show actual running status of the system, set control information and regulate safety level manually. With the regulation of safety level, the number of unsafe connection filled out is increased.

REFERENCE

[1] Hua Jiang.Research and Design of Topic-specific Search Engine Based on Lucene.[D]. Shanghai:East China Normal University,2007

[2] Wenguo Wei, Guiguo Xie. Design and implementation of adaptive best-first Web spider [J].Chinese Journal of Computer Application 2007,27(11):2857-2859

[3] Haixia Lin, Fuyong Yuan,Jinsen Chen,Junfeng Liu. Improved Algorithm about Topic Web Crawler Search Strategy.[J]Chinese Journal of Computer Engineering and Applications,2007,43(10):174-176

[4] Lihong Luo, Zhi Chen. Search Spider of Vertical Search Engine Based on Semantic Nnalysis.[J] Chinese Journal of Computer Engineergin and Design,2008,29(18):4662-4665

[5] Guangli Li, Juefu Liu. Research and Realization of a Spider Model Facing URL. [J] Chinese Journal of East China Jiaotong University, 2007,24(1):67-69

[6] Haitao Chen. The Research on Content Searching, Locating, and Downloading Technologies in Peer-to-Peer Networks [D]. National University of Defense Technology, 2005