Research on the Hotspot Information Push System for the Online Journal Based on Open-Source Framework

Jiya Jiang, Tong Liu, Yanqing Shi and Changhua Lu

Abstract This paper analyzes the technology of the open-source framework as HttpClient, HTMLParser, and IKAnalyzer, and then gives a system for the individual needs of researchers. The system can collect the online journal information automatically, analyze the hotspots and then push the hotspots to the researchers.

Keywords Online journal · Hotspot analysis · Information push · Open-source framework

1 Introduction

At present, all kinds of scientific papers increase at a rate of more than two million articles each year [1]. Finding and utilization of the massive data become the common concern of the researchers. There are three questions in the use of the journal articles: Firstly, for the copyright reasons, most of the journals that appear in the digital publisher's Web site have a few months lag, but these journals can update the information of the latest articles on their own official Web site usually; Secondly, the digital publishers have a large scale of the digital publication and offer a variety of convenient query for the researchers, but require the user to take the initiative to search, and lack of personalized hot push function. Thirdly, some research institutes do not buy data resources, which bring more inconvenience to the journal articles query and utilization.

J. Jiang (✉) · T. Liu · Y. Shi · C. Lu
Beijing Science and Technology Information Institute, Beijing 100044, China
e-mail: jiya_jiang@sina.com

T. Liu
e-mail: Liu_tongmiss@163.net

T. Liu · Y. Shi · C. Lu
Beijing Academy of Science and Technology, Beijing 100089, China

S. Patnaik and X. Li (eds.), *Proceedings of International Conference on Soft Computing Techniques and Engineering Application*, Advances in Intelligent Systems and Computing 250, DOI: 10.1007/978-81-322-1695-7_10,

For the above phenomenon, this paper proposes a tracking and hot push system based on open-source framework. According to the individual needs of researchers, the system crawls the latest online journal in a targeted manner regularly and automatically, and pushes the hot topics of concern to the user automatically. Thus, the researchers can use the information conveniently.

2 Key Technologies

In this paper, the main idea is as following: firstly, determine the collection Web site; secondly, crawl the latest information from concern online journals; thirdly, generate the knowledge database and analyze the hot spot; finally, push the analysis for the scientific and technological workers. All of the technology in this paper utilizes the java-based open-source framework, just call the simple interface to complete the complex data acquisition and data analysis for researchers. The open-source frameworks used in this paper include HttpClient [2], HTMLParser [3], and IKAnalyzer [4].

2.1 HttpClient

The JDK java net package provides HttpURL Connection technology, many of the early applications adopt the jar package for data acquisition, but for most applications, the functions provided by the JDK library itself are not enough rich and flexible. In recent years, the developers are keen to the HttpClient technology to achieve data acquisition. HttpClient is a subproject under the Apache Jakarta Common, can be used to provide efficient, latest, feature-rich support for the HTTP protocol client programming toolkit, and support for the latest version of the HTTP protocol and recommendations.

The HTTP protocol is the most used in the Internet latest years, the most important protocol, more and more Java applications need to be directly through the HTTP protocol to access the network resource. HttpClient has been applied in many projects, and two other open-source projects, such as Apache Jakarta famous Cactus and HTMLUnit, use the HttpClient. The latest version of HttpClient is HttpClient 4.2 (GA).

2.2 HTMLParser

HTMLParser is a pure java html-parsing library, and it does not depend on other java library files, mainly for the modification or withdrawal of HTML, and is currently the most widely used html-parsing and analysis tools, and its latest

version is HTMLParser 2.0. The HTMLParser has two main functions for information extraction and conversion. Information extraction features include five subfunctions: text information extraction, such as HTML, effective information search; link extraction, used to link text with the link to the page automatically label; resource extraction, such as some of the pictures, the sound of the resources at its disposal; link Checker is used to examine the HTML link is valid; and monitor the content of the page. Information conversion function consists of five subfunctions: the link rewrite, used to modify all hyperlinks in the page; web content copy for the web content saved to the local; contents of the test can be used to filter some words on the page; HTML information cleaning, HTML format; and into XML format data.

2.3 IKAnalyzer

Most open-source software is from abroad, so the Chinese word is segmented as single word, this way is ineffective. IKAnalyzer is an open source, light weight java language development-based Chinese word segmentation tool kit. This open-source project is developed by Lin Liangyi et al. all of whom are Chinese. It is widely used as the Lucene Word Breaker for Chinese word. With Lucene version updates and constantly updated, it has been updated to IKAnalyzer 2012 version. Initially, it is based on the open-source project Lucene of the main application, combined with a dictionary of words and grammar analysis algorithm of Chinese word segmentation component. From version 3.0, IK develops from the common word components for Java, independent of the Lucene project, while providing a Lucene default optimization to achieve.

IKAnalyzer provides a unique forward iteration, the most fine-grained segmentation algorithm, with 600,000 words/s high-speed processing capability. And the use of multiprocessor analyzes the submode support: English alphabet, digital, Chinese vocabulary, etc.

3 The Realization of the System Framework

The system is mainly composed of two parts: One is the information crawl, and another is hot spot analysis. According to the interest of the researchers, information crawl set up the source sites, analyze of Web page structure, and design crawl mode; Then, use the HTMLParser and HttpClient to get the information from those Web sites and store these information into the database; By the analyses of these papers in the database, the hot spot is analyzed; At last, the hot spots are pushed to the researchers. The information system workflow is shown in Fig. 1.

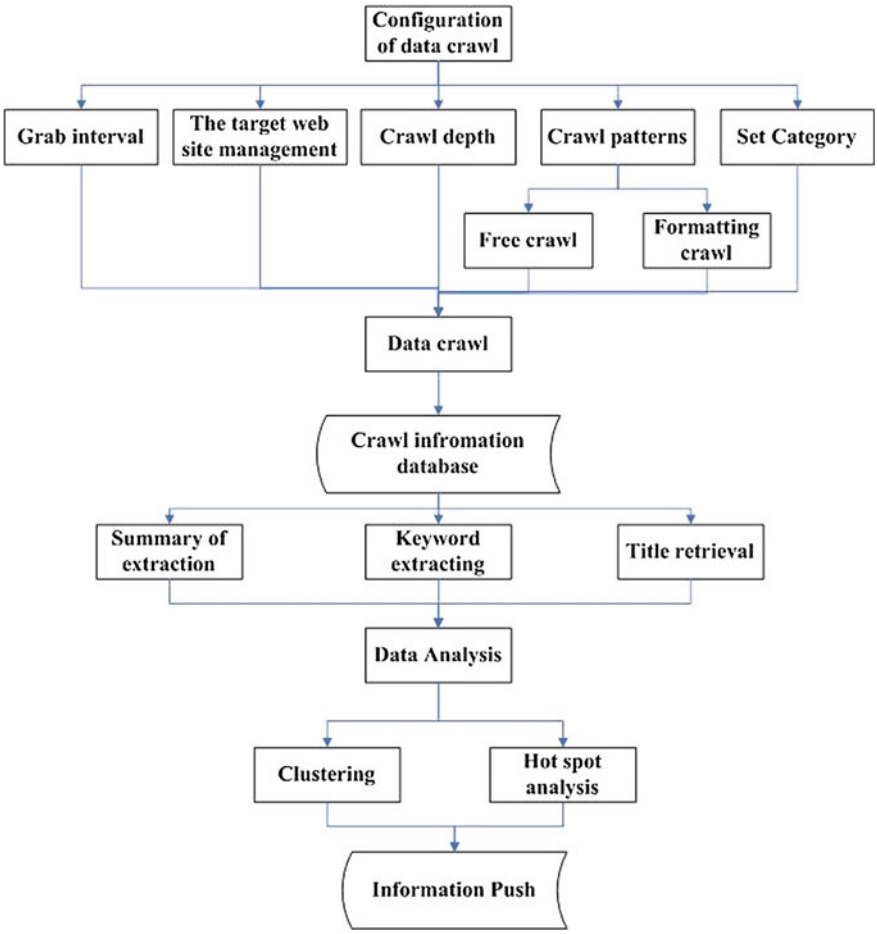

Fig. 1 The program of the system

4 Examples

For example, a research worker sets a Web site as his focus journal and selects concern "semantic" hot, and then, the system will be collected regularly for catalog of periodicals published in the journal's Web site. In the online journal, the system acquisition in recent years' paper information 442, the data are stored in the database. After the hot spot analysis, the high-frequency words are gotten as follows: semantic, cloud, mining, extraction, patterns, knowledge base, identity, search engine, OPAC, labels, a retrieval system, certification, CSSCI, acquisition, folksonomy, biomedical, k-means, public opinion, theme indexing, and crawling. The system will push the paper list about "semantic" to the researcher. Online publication in Springer Link.

5 Summary

This paper completed automatic acquisition and hot papers analyses, based entirely on open-source architecture to achieve the secondary development just need to make appropriate adjustments to the open source code. The system can realize information collection, information filtering, analysis of hot words, and information push.

Acknowledgments This work was financially supported by the program of Large-scale Network Authentication Center affiliated to Beijing Municipal Institute of Science and Technology Information (No. PXM2012_ 178214_000005), the program of Innovation Group for Internet Real-name System (No. IG 201003C2) and the program of Beijing Talent Training Plan (No. 2012D 0020 2200 0002). Thanks a lot for them.

In addition, I want to thank all our colleagues, both past and present, for their assistance during the progression of this research.

References

1. Apache Software Foundation: The Apache HttpComponents project. Apache Foundation, March 30, 2013. http://jakarta.apache.org/commons/httpclient/. Accessed 10 Feb 2013
2. Baidu Encyclopedia: The introduction of Htmlparser. Savagert. http://baike.baidu.com/view/1174491.htm. Accessed 10 Feb 2013
3. CodePlexProject Hosting for Open Source Software: Htmlparser. http://htmlparser.codeplex.com/. Accessed 20 Mar 2013
4. Lin, L.: The Chinese word V2012 user manual for IKAnalyzer. 30 Mar 2012