Python Network Source Automatic Evaluation System

Guogang Zhang[1, a], Xiaoping Li [2,b]
[1] Beijing Institute of Technology, Beijing, China
[2] Beijing Institute of Technology, Beijing, China
[a] email, [b] email

Keywords: Evaluation System, Natural Language Processing, Web Crawler

Abstract. With the rapid development of Internet and computer technology, the network information is exploding combined with the network capture technology and text analysis technology to achieve the evaluation of content resources has become hot research fields. The use of this method is significant to the resources evaluation. The topic of this paper comes from the project of the Ministry of science and technology project "Content Bank Evaluation System", this paper will make a detailed analysis of the evaluation mode based on the network information, and make the research and design of the network data capture and text analysis technology.

Introduction

The objective evaluation of network source is the key of network source management. As the construction of basic resource bank, it provides the orientation of construction of network source and it also reflects the preferences and demands of network searchers. We can optimize the network search engine, more importantly we can evaluate the network source in a more objective way to avoid one sidedness of subjective factors through the evaluation and arrangement of web resources. We can get a more comprehensive evaluation of network source information by analyzing a large data mining and data evaluation.

It can collect network information timely and analyze mass Internet surfer behavior of audience and the evaluation opinions combining web information crawling technology and semantic analysis technology, then it can evaluate a content resources more accurate, provide the basis for selection of the audience and provide personalized push selection. Through the extraction of key words, emotional analysis and analyzing the key words. We can know the latest development in the field of scientific research and the latest direction of the development of advanced technology.

Basic principle

Web crawler. Web crawler is also called web spider. It is a program to automatically download web pages according to rules and it is an important part of the search engine grab system. [1] The information on the Internet is scattered in the hundreds of thousands of web pages, and these pages are stored by millions of servers in every corner of the earth. Users who browse in the web world usually only get information via hyperlinks and they travel through the web pages. [2] The crawler can collect the information of multiple sites and do the further analysis and mining through the online (web page is downloaded) or offline (after the page is stored).

Design of Crawler Based on Scrapy. Scrapy is a Python language based on Framework Crawler, It uses Twisted to handle network communications and extract structured data, the framework is clear and it contains a variety of middleware interface. It can meet the various needs flexibly, such as data mining, information processing, or historical archives. Users only need to customize several modules to easily implement a crawler, it is very convenient to grab web content and a variety of pictures. Although Scrapy

was originally designed for scraping screen (more accurate to say that it is scrapying WEB), it can also be used to extract data using API, Scrapy principle is shown in Figure 2.1.

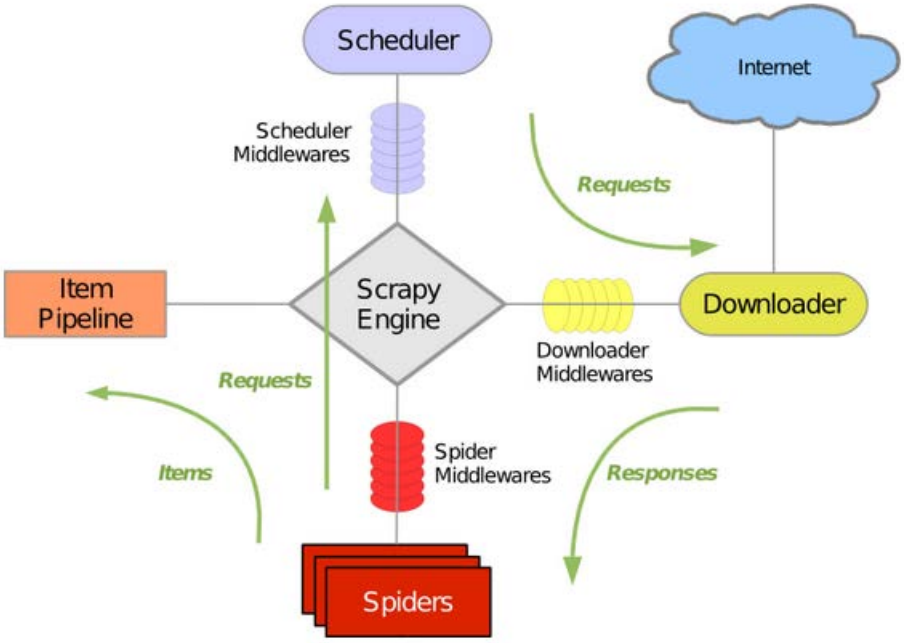

Figure 2.1 The principle of Scrapy

Natural Language Processing. Natural Language Processing(NLP) is a theory and method to achieve the communication between human and computer in natural language (such as Chinese, English etc.) [3], it is an important direction in the field of science and artificial intelligence. The earliest natural language understanding of the research work is Machine Translation. [4] In 1949, Americans Weaver first puts forward the design scheme of Machine. [5] In 1960s, foreigners did extensive research work and spent huge fees on machine translation, but they made a little progress because of underestimating the complexity of natural language and immature of the theory and technology of language processing[6]. The main method is to store words and phrases of two languages, and do the translation according to Dictionary of translation and then just adjust the same order of the language technically. However, the language translation in our daily life is far from simple, it often needs to refer to the meaning of a sentence before and after[7].

Program design and experimental results

Program design. The teaching resources evaluation system based on crawler has the perfect module design and functions of each module are independent of each other[8]. The data can be more fully utilized through the data correlation and the reasonable control process and the evaluation results are more objective and accurate, the concrete process as shown in Figure 3.1.Color figures are welcome for the online version of the journal. Generally, these figures will be reduced to black and white for the print version. The author should indicate on the checklist if he wishes to have them printed in full color and make the necessary payments in advance.

Experimental results. Scrapy based web crawler and agent based mobile network data acquisition technology has successfully achieved the demand data through the final test. Put these data in the Mongodb database, and use of MongoVUE management tools to do the data visualization management.

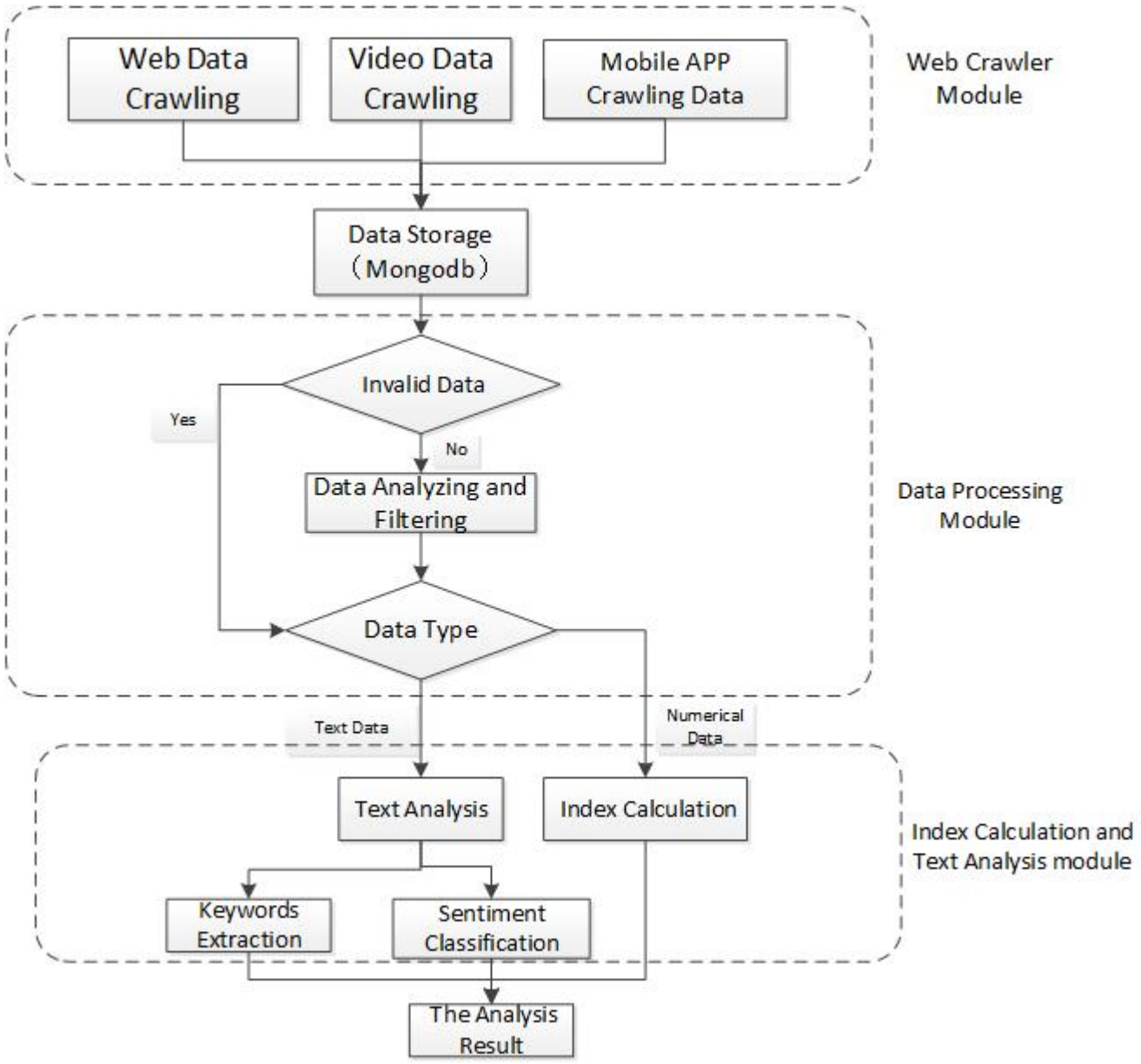

Fig. 3.1 The process of evaluation system of teaching resources

_id	__biz	mid	sn	main	msgext
55adb962ba50...	MjM5NTA3Mj...	209303753	4f7f8299c11bf3...	{8 Keys}	Array[2]
55adb990ba50...	MjM5NTA3Mj...	209303753	eca5eabffaaac...	{8 Keys}	Array[3]
55adb9cfba50c...	MjM5NTA3Mj...	209291869	7295c01bfe8a7...	{8 Keys}	Array[2]
55adb9fdba50...	MjM5NTA3Mj...	209291869	71aa1ceb5370...	{8 Keys}	Array[3]
55adba32ba50...	MjM5NTA3Mj...	209272746	8a670177eb0df...	{8 Keys}	Array[2]
55adba60ba50...	MjM5NTA3Mj...	209272746	696cd8089a4b...	{8 Keys}	Array[3]
55adbbb1ba50...	MjM5NTA3Mj...	209210767	d62211a3b52c...	{8 Keys}	Array[1]
55adbc08ba50...	MTc0NjI0NTk4...	210098052	5f6d6ccfc8caf7...	{8 Keys}	Array[5]
55adbc6cba50...	MTc0NjI0NTk4...	210088887	172f1f3243f43c...	{8 Keys}	Array[5]
55adc029ba50...	MTc0NjI0NTk4...	209911681	beba0baa643e...	{8 Keys}	Array[5]

Fig. 3.2 Data crawl results

Semantic analysis keyword extraction results. The experiment selects a text in the current education field as the analysis data, as shown in the figure 3.3, there are many HTML tag information in the process of text data capture because of the different structure of the website. We use two methods of TF-IDF and

TextRank to analyze and extract 20 key words under the condition of no information filtering or simple filtering.

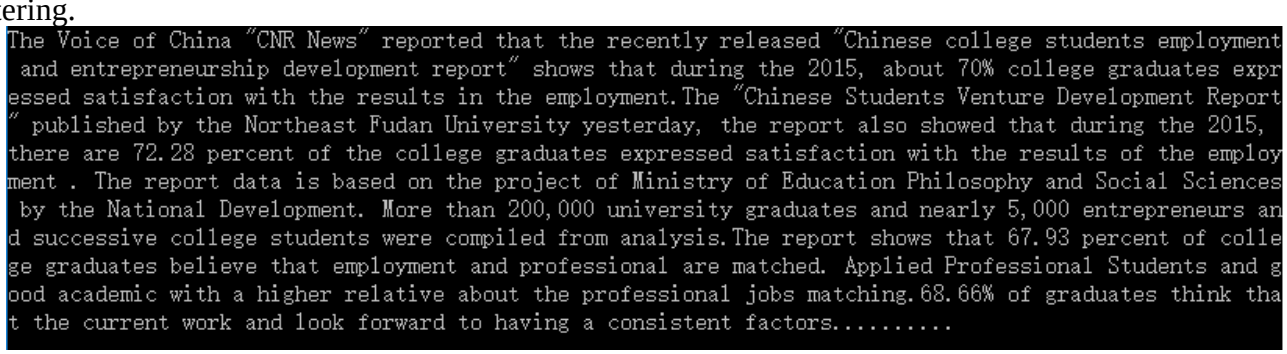

Figure 3.3 Data text filtering

Table 3.1 Comparison of textrank and TF-IDF methods before filtration

Methods	Key Words
textrank	Graduate Employment Report College Students Colleges and Universities Jobs Select Match of major Development Unit Fresh Factor Nationwide Display Higher than Profession China Occupation Financial
TF-IDF	Graduate 2015 Employment Colleges and Universities Venture Match of major College Students Report Fresh College Students Jobs Prospects style tr td table id left Display Expressed Satisfaction

Conclusions.

Web crawling technology and text analysis technology is of great value in the field of scientific research. crawling technology can quickly gather more extensive internet resources, using text analysis technology combing the extraction of the characteristics of internet resources and recommend related courses. It can achieve the effective and high speed retrieval of resources combined with the search engine technology and classification technology.

References

[1] Nripendra Dwivedi, Lata Joshi, Neeraj Gupta.Statistical Analysis of Search Engines(Google,Yahoo and Altavista) for their search result[C]. Proceedings of 2011 4th IEEE international Conference on Computer Science and Information Technology (ICCSIT 2011),2011,9.

[2] Dean J, Ghemawat S. MapReduce: simplified data processing on large cluster[J].Communications of the ACM, 2008,51(1):107-113. [2] Yu Zhenbin. Natural language understanding research [J]. Journal of east China normal university, 2005.

[3] G. Henkelman, G.Johannesson and H. Jónsson, in: Theoretical Methods in Condencsed Phase Chemistry, edited by S.D. Schwartz, volume 5 of Progress in Theoretical Chemistry and Physics, chapter, 10, Kluwer Academic Publishers (2000).

[4] Wang Meng, Natural language processing technology and its application [J]. Journal of education practice and understanding of mathematics, 2015, (20) : 151-156.

[5] Yang Haodong Jiang Ling. Domestic natural language processing research hot spot analysis [J]. Journal of library intelligence, 2011 (10) : 112-117.

[6] Bing Liu.Web Data Mining[M]. Beijing: Tsinghua University Press,2009:3-8.

[7] Gregory H. Siber,Kathleen McCoy. Efficient text summarization using lexical chains[C]. Proceedings of the ACM Conference on Intelligent User Interfaces(UIU'2000),2000:9-12.

[8] Turney,Peter D.,& Littman, Michael L. Measuring praise and criticism:Interence of semantic orientation from association[C]. ACM Transactions on Information System,2003,21(4),315-346.