Intelligent Retrieval of Tourism Information and Design of Personal Tailor

Hailan Gu, Yanbin Long, Siman Bao, Xiaoyue Chu, Yansheng Li

College of Computer and Software Engineering, University of Science and Technology Liaoning, Liaoning, China.

Abstract

With the development of network technology, in order to effectively extract relevant information, Web crawler technology came into being. It is a technology to obtain data in the network and store target information through the website, which can effectively provide data support for the private customized tourism system. The application of Web crawler technology in the private customized tourism system reflects its advantages of wide access, high efficiency and accuracy.

Keywords

Web Crawler; Tourism; Personal Tailor.

1. Introduction

With the emergence of massive information and the continuous progress of network technology, China's tourism industry is also in a period of rapid development, and China's tourism data information has also begun to produce explosive growth. How to effectively extract and use this information to meet the needs of users has become a difficult problem that many tourism websites must break through.

At present, there is no website at home and abroad that can assist users to make a complete and detailed plan according to certain conditions and tendencies of users. Therefore, in order to realize the private subscription, it is necessary to collect a large amount of relevant data and information, so crawler algorithm is very necessary to realize the private subscription function of the website.

2. Personal dingzhi

The private customization function can provide users with relevant information and scheme reference, and assist in making travel plans. At the same time, it can intelligently customize a complete and detailed plan according to certain conditions and tendencies, including destination, schedule, travel mode, accommodation, etc., giving users a higher degree of freedom, so that users can realize more diversified travel modes.

3. Web crawler

Web crawler is a program that automatically extracts web pages. It downloads web pages for search engines from the World Wide Web and is an important part of search engines. Web crawler can take the place of people to collect and arrange data information automatically on the Internet. If we only rely on manpower to collect information, it will not only be inefficient and cumbersome, but also increase the cost of collection. Using web crawlers can obtain more accurate information and improve search efficiency. Therefore, it is necessary to study web crawlers and put forward more effective optimization measures.

3.1 Web crawler workflow

First, select some carefully selected seed websites;

Put these URLs into the URL queue for crawling;

The URL to be grabbed is taken out from the URL queue to be grabbed, DNS is analyzed, the ip of the host is obtained, the webpage corresponding to the URL is downloaded and stored in the downloaded webpage library. In addition, put these URLs into the crawled URL queue.

Analyze the URLs in the crawled URL queue, analyze other URLs, put the URLs into the URL queue to be crawled, and enter the next cycle, using multithreading to effectively manage the network communication interaction.

The algorithm workflow diagram is shown in Figure 1:

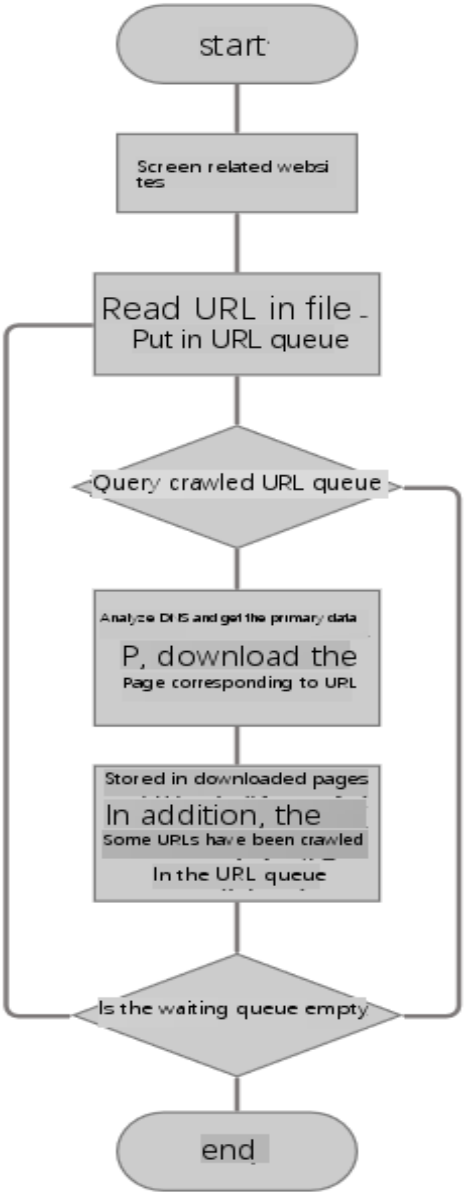

Figure 1. Flow chart of crawler algorithm

3.2 Topic relevance judgment of web crawler

At first, the system composition of topic searcher was considered to filter pages. Unlike ordinary crawlers that deal with all page links, the topic relevance between this page and restricted area will be analyzed first, and will be dealt with only when the topic relevance meets the requirements, because if the page is more relevant to this field, the links it contains are more likely to be related to this field. This improves the accuracy of crawling. Although a few pages will be missed, the overall effect is satisfactory. Therefore, topic relevance analysis is the key to the design of topic crawler.

3.2.1. The specific steps to judge the topic relevance

1. The feature items of title and text are selected by matching with topic set after word segmentation, and the title vector and text vector with the same dimension as topic vector are obtained by word frequency calculation.

2. Calculate the relevance B of topic and title by vector space model.

3. Calculate the correlation degree c between topic and text through the vector model of space vector.

4. Relevance between the theme and the whole webpage: A = 4× B+C.

5. After detailed calculation, the threshold of relevance is set to 2, and the relevance between the webpage and the topic is A>2, then the webpage is considered to be related to the topic.

3.2.2. Implementation steps and algorithm description of judging correlation algorithm

1. The feature items of title and text are selected by matching with topic set after word segmentation, and the title vector and text vector with the same dimension as topic vector are obtained by word frequency calculation.

2. Calculate the relevance B of topic and title by vector space model.

3. Calculate the correlation degree c between topic and text through the vector model of space vector.

4. Relevance between the theme and the whole webpage: A = 4× B+C.

5. Through detailed calculation, the threshold value of relevance is set to 2, and the relevance between the web page and the topic is A>2, then the web page is considered to be related to the topic.

Enter: topic set text a.txt, web page url

Output: topic relevance

(1) Get topic (String path)// get the topic text collection according to the path

(2) Compulate Topic Weight (String Topic)//Calculate the topic combination weight

(3) SortandDelRepeat (int [] count)//Delete duplicate elements and sort them

(4) delRepeat(String[] segment)// Delete repeated elements after segmentation

(5) delRepeat(Vector url)// Delete the duplicate elements in the obtained URL

(6) getParser(String url)// get Parser instance

(7) String titlestr = p.geturltitle ()//Get the title of the webpage

(8) String body str = p. getparagraphtext ()//Get the webpage text

(9) String titlestr seg = segment.segment (titlestr)//Web page title word segmentation

(10) string bodystrseg = segment.segment (bodystr)//webpage text segmentation

(11) compulate title.length, body.length//calculate the length of title vector and the length of web page text vector

(12) set topicweight1 ,titleweight1 ,bodyweight1; //set the weight

(13) Last compute relative //calculate topic correlation

(14) Return relative; //return result

According to the system settings, the first step is to download all the web pages, and then determine the topic relevance. If it is related to the topic, it will be placed in the relevant URL library, and the irrelevant web page will be discarded.

3.2.3. Save web page information

1. Establish a URL connection first.

```
URLConnection url_C = url_test.openConnection();
```

2. Create a new Pagepro class.As follows:

```
    private String
    Host;
```

private
int
Port ;
private String
ContentType ;
private
int
ContentLength ;
private String
Date;
private String
Url ;

3. Save the data into the newly created Pagepro class.
4. Save the data to the address entered in advance.

The crawler algorithm is shown in Figure 2.

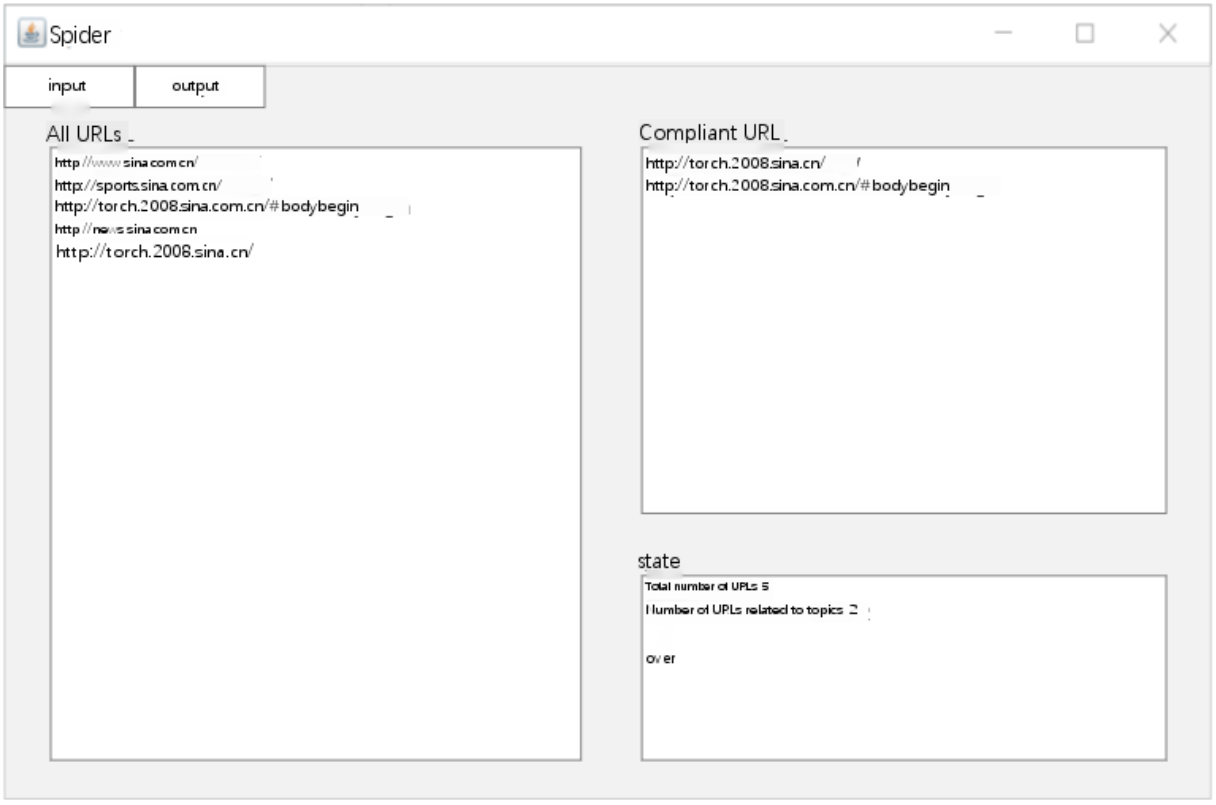

Figure 2. The effect of crawler algorithm

4. Epilogue

With the development of network technology, tourism data information has also begun to produce explosive growth, accurate crawling network information has become a huge challenge.This paper first introduces the private customization function, and then analyzes the web crawler workflow to judge the topic relevance. Finally, it concludes that when the topic relevance meets the requirements, it can improve the crawling accuracy of web pages. Although there will be some missing pages, the comprehensive effect is better. Therefore, the analysis of topic relevance is very important in crawler design.

Acknowledgments

The project was funded by Liaoning University of science and technology 2021 innovation and entrepreneurship training program, project number: x202010146487.

References

[1] Chinese search engine technology decryption: Web spider [M]. Beijing: People's Posts and Telecommunications Press. 2004

[2] Li Xiaoming, Yan Hongfei, Wang Jimin. Search engine: principle, technology and system -- the academic library of Huaxia talent foundation [M]. Beijing: Science Press, April 2005

[3] Tian Jun. discussion on Key Technologies of theme web crawler [J]. Journal of Tianjin Vocational Colleges, 2017 (3): 78-85

[4] Yu Juan, Liu Qiang. A review of topic based web crawler research [J]. Machine engineering and science, 2015 (2): 231-237

[5] Luo Gang, Wang Zhendong. Writing web crawlers by themselves [M]. Beijing: Tsinghua University Press, October 2010