Optimization Study of Multimedia Education in the New Era Based on Computer Platform

Xiaodong Shu(✉)

School of Music and Performing Arts, Mianyang Teachers' College, Mianyang, Sichuan, China
silent157768526@163.com

Abstract. Web multimedia educational resources refer to multimedia teaching resources that exist in the Internet. With the development of network and multimedia technology, multimedia teaching resources in the Web, especially audio, video and animation, have become increasingly abundant and become an important part of the education field. How to quickly and accurately find multimedia teaching resources on specific topics, so that they can play a full role in information-based education, is an urgent problem for educational technology workers, but also makes the traditional search engine faces a huge challenge, all kinds of multimedia search engine also came into being and received widespread attention. The search algorithm of the topic searcher, which is the core of the search engine, determines the search efficiency and quality of the search engine. This paper proposes a multimedia topic search algorithm based on URL link rules from the characteristics of topic web pages and web pages containing multimedia distributed in the Web and focuses on the problem of improving the efficiency of multimedia topic search.

Keywords: Web Multimedia · Topic Finder · Topic Search Algorithm · Multimedia Education

1 Introduction

Multimedia is a comprehensive information resource, which is a collective name for media elements such as Text, Graphic, Sound, Animation, and Video [1]. With the development of computer technology and the Internet applications have become popular, the information construction of basic education teaching resources has also been flourishing. With the emergence of various types of specialized websites for teaching resources, the Web has become the main way for people to obtain teaching resources. According to the 19th Statistical Report on the Development of China's Internet released by CNNIC, by the end of 2006, the number of web pages on Chinese websites was 4.47 billion, among which text and images were still the most important content forms of web pages, occupying 70.2% and 29.5% respectively; video web pages accounted for 0.3% of the total number of web pages. And according to the classification of multimedia formats: the pages in swf format account for 1% of the total number of web pages, and the pages in mp3 format account for 0.1% of the total number of web pages. Meanwhile, according to the report released by CNNIC in August 2008, the number of websites in China has reached 1.919 million, with an annual growth rate of 46.3%, and continues to maintain the momentum of rapid growth [2].

C. F. Peng et al. (Eds.): EIMT 2023, AHSSEH 8, pp. 491–499, 2023.
https://doi.org/10.2991/978-94-6463-192-0_65

2 Processing of Text Information

In the process of theme search, the text information extracted by the HTML parser has to be processed before it can be added to the calculation process of the algorithm, and the information such as anchor text and web page title needs to be split into Chinese, and the web page link URL characters need to be translated into English and Hanyu Pinyin. The following is a detailed description of the related technologies.

2.1 Chinese Word Separation Algorithm

At present, Chinese word separation algorithms can be broadly classified into three categories: dictionary matching-based word separation algorithms, word frequency statistics-based word separation algorithms, and knowledge understanding-based word separation algorithms. This system uses a forward maximum matching algorithm (MM algorithm) based on a dictionary of educational topics. The idea of the algorithm is that a correct splitting result should consist of legitimate words that are in the current sentence to be cut and that belong to a split in the lexicon. In the word separation process, the maximum length first matching is used according to the positive scanning direction. The specific description is: assuming that the number of Chinese characters contained in the longest word in the automatic word separation dictionary is MaxLen, the MaxLen words in the current string of the processed material are taken as the matching fields to find the word separation dictionary.

2.2 English and Pinyin Translation Technology

The URL of a web page is usually composed of strings, either English words, Chinese pinyin, or abbreviations of Chinese pinyin, etc. For example, http://www.chuzhong2wuli/kejian/index.axp can be represented as a website for middle school, physics, and coursework. Web page URLs are also an important component in characterizing the content of a web page when determining the similarity between the content and the topic. Therefore, when analyzing this part, we need to first convert the URL of the web page into Chinese words, and then match it with the educational subject dictionary to determine the content similarity of the web page. The web URL string is divided into individual words, and then each word is translated into Chinese words by Chinese and English dictionaries. The specific process of the technology is: Chinese and English translation dictionary loading, word separation based on defined separators, word separation based on capitalized initials, Chinese and English translation and irrelevant word filtering. Due to the uncertainty of web page URL strings, we try to encompass as many strings as we can gather while the system is running. It also reduces the weight of this part in the overall similarity when calculating the similarity of web content.

3 Analysis and Processing of Hyperlinks

In the process of subject search, the subject searcher first starts from an initial set of hyperlinks, puts all these hyperlinks into an ordered queue of hyperlinks to be extracted, and then takes them out from this queue in order to obtain the pages pointed to by the

hyperlink URLs through the protocols on the Web, and then analyzes and extracts new URLs from these obtained pages, and continues to put them into the queue of URLs to be extracted, and then repeats the above process until the Web information extractor stops collecting them according to its own search strategy, which shows that hyperlinks are the most critical information in the subject searcher [3].

3.1 Implementation of the Page Parser

In this paper, we introduce the HTML parser of the reference. The HTML parser of the system design is roughly divided into five parts: protocol converter, web page reader, information extractor, URL path transformer, and information memory. The implementation process is as follows:

(1) set a time variable to save the time of access to the page and store this time variable value in the database; set a Boolean variable is HTTPS to determine whether the site uses HTTPS (Hypertext Transfer Security Protocol) to exchange data with the browser.
(2) If the value of is HTTPS is true, change the value of protocol in the URL from HTTPS to HTTP, set the port number to 443 and use SSL (Encrypted Socket Layer) to establish a connection to the server side, otherwise, simply create a socket connection.
(3) Sending an HTTP request to the web server at the specified URL.
(4) read in the server-side response title and determine whether the value of the title field Location is empty, if not, the value of the title field Location as the URL of the current web page to be accessed, that is, URL redirection, to (2); if the value of the title field Location is empty, then to (5).
(5) establish a dynamic link to the data source specified by the URL through URL Connection.
(6) if the connection is successful, read the HTML code of the web page into the reader Reader through the input stream to (7); otherwise, generate an error message and store this error message in the database.
(7) instantiate the parser class HTMLEditorKit. Parser object, and then call its parse method, which will read the HTML code from the Reader until the entire document is read.
(8) Instantiate the callback class HTMLEditorKit.ParserCallback object and pass it as a parameter to the parse () method of the parser HTMLEditorKit.Parser object; this class will do the actual parsing of HTML.
(9) If the tag < Meta > is encountered, the callback function handleSimpleTag is called to extract the value of the tag attribute name, and if the value is Keywords or Description, the value of content is extracted and stored in the database; if the tag < Title > is encountered, the callback function handleText is called to extract the title of the page and store it in the database.
(10) If the tags < A >, < Area >, < Map > are encountered, the callback functions handleStartTag and handleEndTag are called to extract the value of the attribute HREF of the tag; if the tag < Iframe > is encountered, the callback functions handleStartTag and handle End Tag are called, extracting the value of the attribute Src

of the tag; if the tag < Base > is encountered, the callback function handleSimpleTag is called, extracting the value of the tag attribute Href; if the tag < Frame > is encountered, the callback function handleSimpleTag is called to extract the value of the tag attribute Src; if the tag < Img > is encountered, the callback function handleSimpleTag is called to extract the value of the tag attribute Src; If the tag < Param > is encountered, the callback function handleSimpleTag is called to extract the VALUE value of the attribute of the tag, while each < Object > tag will generally nest multiple < Param > tags, where only the one whose VALUE value is the link address is extracted.

(11) If the extension of the extracted Href value, src value or Value value is not .html or .htm or .shtml or .asp or .jsp, it will be directly discarded; Otherwise, if it is one of the above formats and is a relative URL address, the relative URL address needs to be parsed through the link address of the web page using the constructor provided by the URL class to convert it into an absolute URL address so that the URL address can still be accessed correctly when leaving the web page.

(12) In order to have a general prediction of the content of the page to which the link points, the anchor text (the anchor text is the carrier of the hyperlink, which the user clicks on to link to a new page or another location on the same page) is extracted as a description of the content of the link; Since only the anchor tag may have anchor text (because if the image is used as the carrier of the hyperlink, the anchor tag will not have anchor text, and the image is called the link image here), only the anchor text of the anchor tag is extracted, and the link text of other tags or the anchor tag with the link image as the carrier is set to null; the method to extract the anchor text is as follows: set a boolean variable When the anchor tag is encountered, the value of this variable is set to true, and then when the HTMLEditor Kit.Parser object processes the link text, the handle Text method of the callback class is called, and the text is read out only when the value of is Link is true. After processing the text, the variable is Link is automatically reset to false, thus realizing the one-to-one correspondence between the link URL value and the anchor text.

3.2 Organization of Page Data

Information about the extracted Web page is stored in the link database along with the hyperlink to the page. The extracted Web page related information includes: the words of the Web page URL translated from Chinese to English, the anchor text of the link, the title of the Web page, all the link URLs contained in the Web page, the number of physical and logical layers of the Web page, the extraction time, etc. All these information are used in the calculation of the Web page content similarity and link similarity. Each page corresponds to a record in the record from which the necessary information can be obtained during the subsequent processing.

4 Calculation of Link Similarity Based on URL Rules

4.1 Multimedia Resource Theme Adjacency

In the same website, the subject pages containing multimedia exhibit the characteristic of "resource proximity". That is, pages containing multimedia resources are often present in one or more sections of this website [4]. And the topics of multimedia resources

in the same area are also the same. Based on this feature we can make the following assumptions:

(1) If a web page is a topic-related page containing multimedia, then the sublinks of this page are likely to be topic-related pages containing multimedia [5].
(2) If a web page is a topic-related page containing multimedia, then the sibling link of this page in the parent page is likely to be a topic-related page containing multimedia [6].

In order to better verify the feasibility of the algorithm, we manually selected 10 high school physics websites as seed links and opened 10 threads to collect web pages containing multimedia Flash in a blind search, and the experimental results are shown in Table 1.

By analyzing the experimental results, we can conclude that the Flash files are mainly concentrated in layers 2–9 of the physical layer of the website and tend to be concentrated on a certain physical layer.

Table 1. Results of blind search experiments

Flash			
Number of seeds		10	
Total number of internal pages		38820	
Number of error pages		2045	
Number of active pages (pages that include Flash)		1316	
Efficiency		3.390%	
Total number of Flash files: 5089			
Physical Layer		No restrictions	
Number of physical layers	Total pages	Valid Pages	Active pages/total active pages
1	621	3	0.2280%
2	1010	72	5.471%
3	5250	178	13.53%
4	1824	103	7.827%
5	18851	775	58.89%
6	730	20	1.520%
7	2734	84	6.383%
8	2129	64	4.860%
9	3498	5	0.3800%
10	1325	8	0.6080%
11	838	0	0%

4.2 Solving the Tunnel Problem

To solve this problem, the theme searcher first crawls each internal link linked by the seed page using a width-first strategy during the search process. In the URL regular expression learning phase, the main purpose of the topic finder is to detect the physical directory and the number of physical layers where the multimedia web pages are located to form URL regular expressions.

$$NUM_{pn} = NUM_{(p2)} * (n - 1)\,(n \leq 10) \tag{1}$$

NUM_{pn} indicates the physical layer n the pages that should be crawled, $NUM_{(p2)}$ indicates the number of pages crawled in the 2nd physical layer, n indicates the number of physical layers, Stop blind search when n = 0.

4.3 Extraction of URL Regular Expressions

For URL clusters whose URL distance meets certain conditions, extraction is performed to obtain URL regular expressions. The specific process of extraction is first decompose the URL of the same site into three parts: host, path and query, and decompose path into a series of directory, and query into a series of key-value pairs. Since the host part is definitely the same, just write the host in the regular expression as it is. Align each directory in the path part and add the value of this part into the regular expression if the directories in the corresponding position are the same, otherwise add it into the regular expression with * instead. A similar approach to the path part is used for the query part [7]. Finally, we can get a regular expression. The flow chart is shown in Fig. 1.

4.4 Implementation of URL Rule-Based Topic Search

Based on the above facts, we propose a solution for URL rule-based topic search, which consists of roughly two steps:

(1) It is the experimental search phase, where for each seed site, a blind search is first performed to learn some URL regular expressions from the searched pages containing multimedia [8].
(2) is the guided search phase, which uses the URL regular expressions learned from the experimental crawler phase to guide the crawler in the actual web crawling.

The flow of the experimental search phase is shown in Fig. 2.

The seed site list is used as the starting site of the search, and the theme searcher continuously obtains web links and stores them in the InterLink database during the web crawling process, then for the links with the status of W, it first judges the relevance of the web page topic and whether the web page contains multimedia resources and stores the links that meet these two conditions in the Multi Media database. URL regular expression learning is performed when the number of web pages in the Multimedia database is $NUM_{URL} > m$ and the distance between web page URLs is $|D_{url}| < d$.

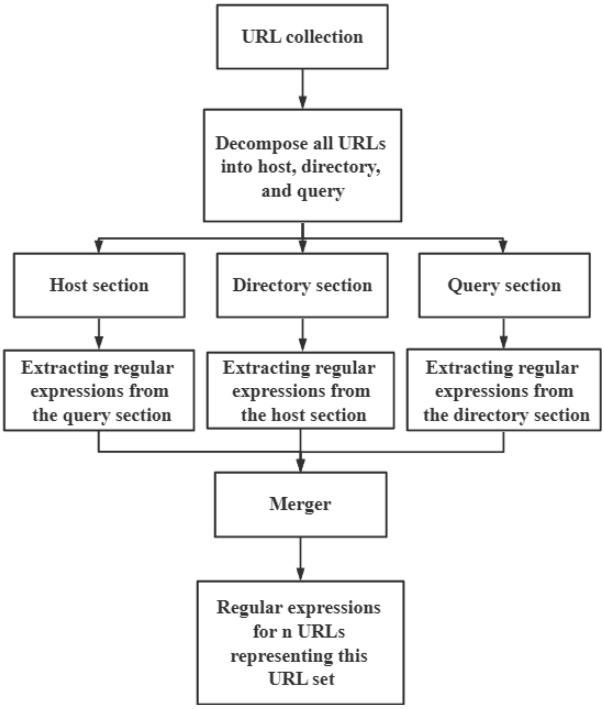

Fig. 1. Extraction process of URL regular expressions

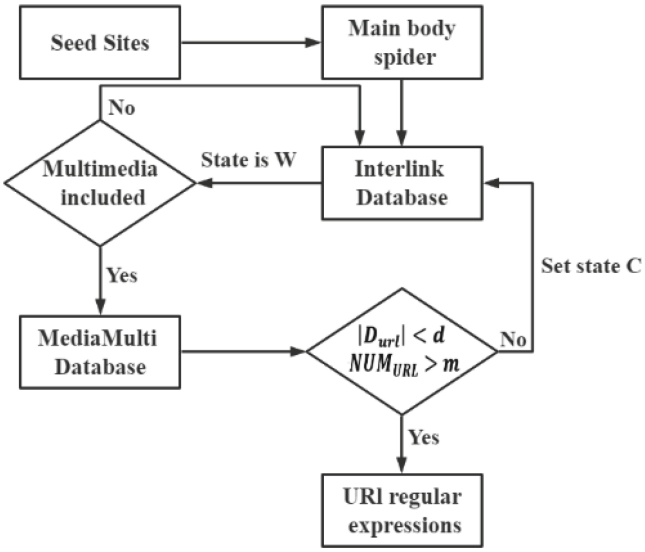

Fig. 2. Experimental search phase

5 Conclusion

In this paper, we introduce the feature of "resource adjacency" for multimedia resources in web pages, and propose a multimedia topic search algorithm based on URL linking rules, and detail the text information processing and link structure analysis and processing in the implementation of the algorithm [9]; then the process of implementing URL regular expressions in the algorithm and how to use URL regular expressions to guide the topic searcher to obtain web links are described in detail, and a specific description of the URL link rule-based multimedia topic search algorithm is given [10]. Finally, to improve the accuracy of multimedia topic search, the algorithm will perform the normalized calculation of content similarity and URL rule link similarity.

References

1. Zhang Chunmei. Exploring the innovative application strategies of multimedia technology in college music education in the new era [J]. China Journal of Multimedia and Network Teaching (Upper Journal), 2022(06):205-208
2. Qiao Ruijie. The current situation and countermeasures of China's college dance education under multimedia environment--a review of "Exploration and practice of college dance education teaching mode in the new era" [J]. Journal of Chinese Education,2021(12):127.
3. Chen Aihua. The multimedia classroom in the new era[C]//. Exploration of curriculum and teaching reform under the policy of "double reduction", 2nd series. https://doi.org/10.26914/c.cnkihy.2021.068337.
4. Li Zonglin. Strategies and implementation of multimedia-assisted high school art teaching in the context of the new era [J]. Art Education Research,2021(20):168-169.
5. Cang Wei. Exploring the innovation of multimedia teaching of ideological and political edu-cation in new era colleges and universities--Review of "Practical Teaching of Microfilm: An Exploration of the Innovation of Teaching Mode of Civics and Political Science Class in Colleges and Universities" [J]. Science and Technology Management Re-search,2021,41(17):221.
6. Lin Cunlong. Basic art education in the new era of multimedia [J]. New Course,2021(21):155.
7. Fan Lili. Discussion on Internet-based ideological and political multimedia education for col-lege students--Review of "Internet+" Research on Teaching Ideological and Political Theory Education in Colleges and Universities [J]. Science and Technology Management Re-search,2021,41(04):219.
8. Ji Yi Yun. The application of multimedia technology in teaching in the new era[C]//.Proceedings of the workshop on "Classroom teaching reform based on core litera-cy" in 2020. https://doi.org/10.26914/c.cnkihy.2020.041006.
9. Li Yang. The literacy requirements of editors and journalists for multimedia integration in the new era [J]. Satellite Television and Broadband Multimedia, 2020(13):306-307.
10. Luo Yingji. The path of multimedia network teaching of college English in the context of mutual communication in the new era [J]. Southern Agricultural Machinery,2020,51(07):210-211.