Research and Implementation of Algorithm Based on Data Fusion Technology

Yu-zi DOU, Xi-wei FENG*, Rui ZHU, Tian-zhu GAO, Yan-bing WU and Lei MA

School of Computer and Communication Engineering, Liaoning Shi Hua University, 113001 Fushun Liaoning, China

*Corresponding author

Keywords: Data fusion, Python, web crawler, PageRank algorithm, NDC algorithm, Student evaluation of teaching.

Abstract. As the growing amount of data stored on the Internet, the work of searching for information becomes complicated. The traditional collection method cannot achieve a certain effect, it is cumbersome and time-consuming. Using natural language processing technology and web crawler technology to collect and analyze data about student evaluation, the purpose is to obtain the key factors affecting teachers' comprehensive evaluation results and propose the methods to solve the problems. For the traditional Web crawler technology, there is a lack of certain intelligence, initiative, etc. the design of the best priority crawler framework has improved and optimized its structure. And the improved PageRank value, user demand correlation degree, and NDC algorithm denoising are added, it can effectively solve a series of problems such as long retrieval time, overlapping information, incomplete information, and improve the accuracy of information collection.

Introduction

The proposal of student evaluation system is to find a solution to the current situation according to the specific needs of students and teachers' teaching requirements.

As an integrated data processing technology, data fusion technology is applied to many traditional disciplines and emerging fields, which can improve the accuracy and reliability of target rule mining and prediction. In [1] combined with the crawler technology, the acquisition and analysis of multi-source spatial data is demonstrated, which is beneficial to better assist the urban planning work; In [2], the design of the Internet public opinion analysis system based on the principle of data fusion, and the data fusion analysis processing is realized by combining the crawler technology with the natural language processing technology; In [3], it is proposed a personal credit scoring system based on multi-source data fusion, which combines the logistic regression model to improve the model estimation accuracy; In [4], the data is collected by adaptive weighted fusion method based on data fusion principle, and the Grubbs criterion is used to eliminate invalid data, so as to comprehensively deal with the problem of measuring the parameters of the inlet section of the test piece in the afterburner of an aero-engine.

Overall System Structure Design

This paper uses multi-source data fusion technology to search for keyword group information about student evaluation in the webpages. The first chapter introduces the overall chapter arrangement of this article. The second chapter introduces the proposed system architecture and optimization scheme. In the third chapter, Applied to the comprehensive analysis of students' evaluation. Chapter four gives a summary and suggestions. As shown in Figure 1.

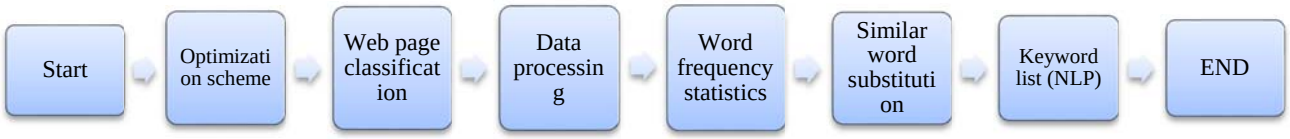

Figure 1. Overall architecture flow chart

Key Technologies and Core Processing Algorithms

Multi-source data fusion technology was first used in the military field. The technology belongs to a kind of attribute fusion. The basic goal is to mine more effective information through data optimization combinations. The framework of multi-source data fusion is mainly divided into three categories, Data layer fusion, feature layer fusion and decision layer fusion[5]. With the help of the Internet, it is one of the popular applications to use the web crawler[1] to automatically grab the programs and scripts of the World Wide Web information according to certain rules and collect the content of the target page.

Optimization Scheme of Web Crawler Based on PageRank Value

Web Crawler. Web crawlers are known as web spiders and web robots. It is a program or script that browses the World Wide Web in a systematic and automated manner. Using this technology to crawl relevant links and content from web pages according to user requirements, and continue to crawl along the link, it is a powerful information gathering program.

The Core Algorithm. The PageRank algorithm is often used to evaluate the importance of web pages, and as one of the important basis for ranking search results. The NDC algorithm is often used to remove irrelevant information in the extracted pages, and it is convenient to remove various noises present in the retrieved information.

If there is a link to page a in page *A*, it means that the owner of *A* thinks that *a* is more important[7]. Therefore, a part of the importance score of *A* is assigned to *a*, and the importance score is PR(A)/ C(A). The PR value (PageRank value) of the traditional PageRank algorithm is calculated as shown in equation (1).

$$PR(A) = (1-\mathrm{d}) + d\sum_{i=1}^{n} PR(t_i)/C(t_i) \tag{1}$$

PR(A) represents the PR score of the webpage; *d* is the damping coefficient, usually set to 0.85. PR(t_i) represents the PR score of an externally linked website t_i itself. C(t_i)represents the number of external links owned by the external link site. Extend equation (1) to get equation (2):

$$PR(A) = (1-d) + d(PR(t_1)/C(t_1) + ...PR(t_n)/C(t_n)) \tag{2}$$

At present, Google's PR value is used as a websit or page level and importance identifier. The main factors affecting PR value are: 1. Linking with PR high website; 2. Joining search engine catalogue; 3. Containing high content quality Links to websites, etc. Suppose the number of externally linked websites is 10. As the initial value of PR increases, the website level analysis value is shown in Figure 2.

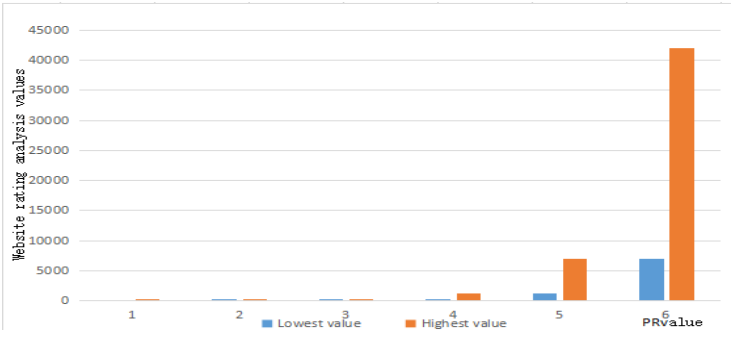

Figure 2. Comparison of PR value and website information analysis value

The higher the relevance of the web page to the user's needs, the higher the similarity between the content of the web page and the required data, the higher the order of the web pages in the search system, and the higher the quality of the collected information. At present, the proportion of popular websites with various attributes in the Internet is shown in Figure 3.

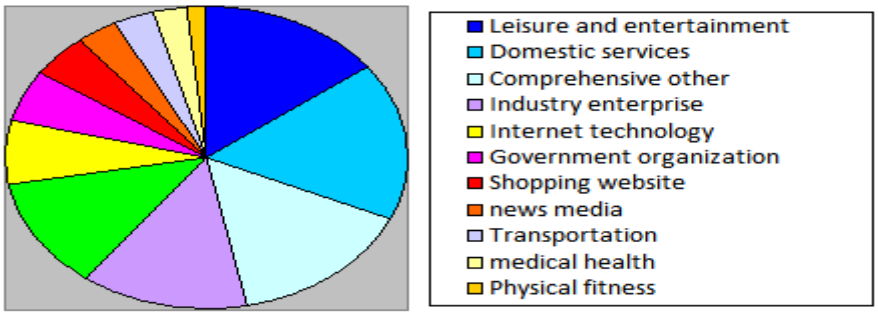

Figure 3. The proportion of various websites

Add the user demand association factor and set the U-PageRank algorithm, as shown in (3).

$$U-PR(A)=\alpha(1-d)+d(U-PR(T_1)/C(T_1)+...+U-PR(T_n)/C(T_n)) \quad (3)$$

α is the attribute of the query website in all websites, the size range is 0-1.

Assume that the initial PR of page *A*, *B*, *C* and *D* is set as 3. in terms of calculating the sorting result, the difference between the input and output of web page *C* is the largest, and the PR value is also the highest. Secondly, the PR value of webpage *A* is higher, and finally the webpages *B* and *D*. Therefore, the results presented to the user are sorted as *C*, *A*, *B*, and *D*. After calculation, the correlation factors of the website data of the webpages *A*, *B*, *C* and *D* are 0.6, 0.1, -0.5, and 0. After adding them into the correction, the PR value of webpage is ranked as *B*, *A*, *C* and *D*. It can be seen that the correlation factor between the website data and the user's demand will affect the ranking of the crawler results.

Compare the U-PageRank algorithm with the original PageRank algorithm. To find music data, Use Python programming language to build experimental environment, conduct comparative experiments on the following five types of websites to verify the effectiveness of the algorithm, as shown in Figure 4.

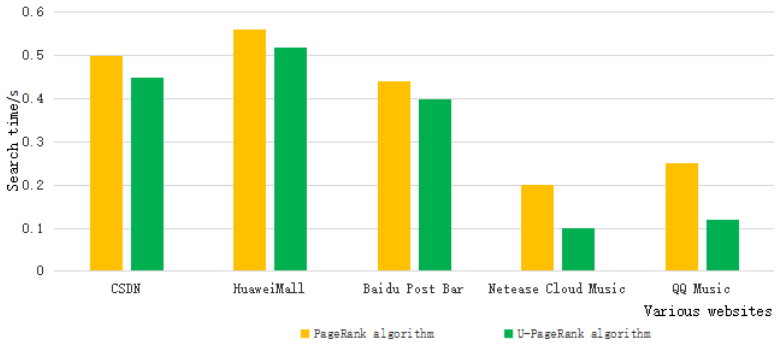

Figure 4. Comparison of U-PageRank algorithm and PageRank algorithm

As can be seen from the above figure, in the retrieval of music information, the retrieval time of CSDN and Huawei Mall Baidu Post Bar is relatively long, and the improved algorithm and the original algorithm have little similar query efficiency on such websites, while the retrieval time on the other two websites is shorter. That is to say, when the information is retrieved from the website with the same data attribute, the improved algorithm is about half of the retrieval time of the original algorithm, which improves the retrieval strength and accuracy.

Experimental Comparison

The U-PageRank algorithm and NDC performance indicators proposed in this paper have accuracy, recall and F1 indicators. For each web page in the evaluation dataset, the extracted content is compared to the gold standard[6]. The calculation formula of the recall rate (r), the precision (p), and the F1 measure, as shown in equation (4),

$$p=|c|/|g|, r=|c|/|e|, f=2\times p\times r/p+r \quad (4)$$

Where |c| is the string length of the URL pattern extractor of the most recent index page, |g| is the length of the gold standard string, and |e| is the length of the extracted content. The proposed improved algorithm is compared with the existing algorithms in the literature, as shown in Figure 5

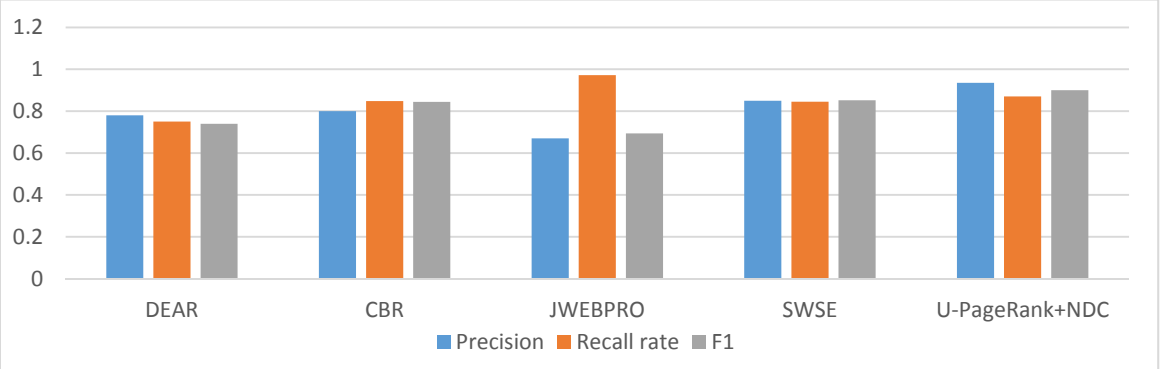

Figure 5. Web dataset customized search engine performance

For the Web dataset, the accuracy of the system for the web data set is 90%. In the literature, SWSE can only achieve the highest accuracy of 84.9%, which means that the accuracy of the U-PageRank+NDC algorithm is improved by at least 8.6%. In addition, the recall rate of the improved algorithm is only 87%, while JWEBPRO has the highest recall rate, but its precision is very low, which means it returns more irrelevant links. The F1 score of the optimization algorithm is high, and the value is 0.9. As shown in the figure above, U-PageRank+NDC algorithm is a better choice.

Results and Discussion

The U-PageRank algorithm and the NDC algorithm are combined with the optimized optimal crawling system, and the PageRank value is set to 3. Through the data retrieval of xuexin.com and so on, and then using the NLP technology analysis method based on machine learning, the key words table of influencing factors of students' evaluation of teaching was obtained, as shown in table 1.

Table 1. Keyword list

Keyword List						
Party member	Age	Education background	Professional title	Teaching age	Research project	Teacher's moral

Then use the data vectorization method to vectorize the keyword group, and get the table 2, the data in Table 2 is drawn vertically using Excel software. to get figure 6.

Table 2. Proportion of vectorized keywords

	Age	Education background	Professional title	Party member	Teaching age	Research project	Teacher's moral
1	0.12612	0.15669	0.36890	0	0.37888	0	0.44332
2	0.00364	0.46786	0.57866	0	0.56565	0	0.43234
3	0.37851	0.58945	0	0	0.66786	0.67554	0.65467
4	0	0.56668	0	0	0.35764	0.34221	0
5	0	0	0.13467	0.13432	0.44443	0	0.76577
6	0	0.63467	0.37890	0.23111	0.32278	0	0.88877
7	0.12224	0.57798	0.54322	0	0.64366	0.35222	0.15655
8	0.01199	0.76666	0.34210	0	0.14533	0.16777	0.24444
9	0	0.89000	0.12411	0.11344	0.56666	0	0
10	0	0.76555	0.67999	0.23411	0.48897	0.67443	0.66555
.....							

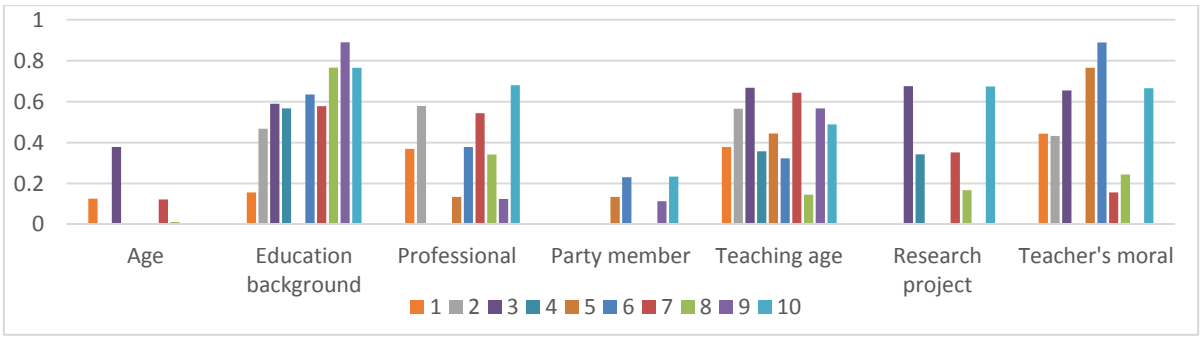

Figure 6. The proportion of students' evaluation of teaching keywords on each website

Through the analysis of Figure 6 we find that the keywords are evenly distributed in each website, and the proportion is relatively dense, such as education, teaching age, etc. On the contrary, in the data, the key words with relatively scattered proportion distribution are age, party member, etc. Indicating that such factors account for a small proportion in influencing students' evaluation of teaching performance, and their importance is low. To sum up, based on the data analysis theory, it can be concluded that education background, professional titles, teaching age, and teacher's moral are the main factors affecting students' evaluation of teaching achievements.

Summary and Suggestions

This paper proposes a Web crawler optimization scheme based on PageRank value, which reduces the retrieval time and improves the retrieval quality. However, its application system configuration is high, and the running process takes up a large memory, which needs further optimization. Based on the results of data mining combined with expert consultation and literature analysis, the main factors affecting students' evaluation of teaching achievements are teachers' educational background, professional title, teaching age and teacher's moral. We should establish a sound education team management system and develop a reasonable job title selection process to promote the development of education.

Acknowledgements

Fund project: Project supported by Liaoning Provincial Natural Science Fundation of China (20180550130)

References

[1] L.L. Pei, J.Z. Tang, X.S. Bi, The Acquisition of Multi-source Spatial Data and Its Application to Urban Planning, Geomatics World. 26(2019)13-17.

[2] X.L. Du, Internet public opinion analysis system based on data fusion, Telecommunications Engineering Technology and Standardization. 7(2017)26-30.

[3] K.G. Fang, M.L. Zhao, A Study on Credit Scoring Based on Multi-source Data Integration, Statistical Research. 35(2018)92-101.

[4] R.X. Wang, H.Y. Wang, Y.F. Sun, Research on Multi-Sensor Data Fusion Method for Afterburner Test, Aeroengine, 43(2017)85-89.

[5] Y.J. Qi, Q. Wang, Survey of Multi-source Data Fusion Algorithms, Aerospace electronic confrontation. 33(2017)37-41.

[6] P. Sahoo1, R. Parthasarthy, An Efficient Web Search Engine for Noisy Free Information Retrieval. 15(2018) 412-418.

[7] R.X. Chen, Y. Chen, C.Y. Deng, Design of Data Capture Program Based on Web Crawler Technology ,2018 International Conference on Information, Electronic and Communication Engineering (IECE 2018).2018, pp:218-221.

[8] Z.Y. Li, Discussion on the Optimization Strategy of Web Crawlers, Journal of Modern Information, 31(2017)31-35.

[9] D. Wang, J. Xu, C.Y. Du, Improved user influence evaluation algorithm based on Page Rank, Jou R Nal of Ha R Bin Institute of Technolgy. 50(2018)60-67.

[10] P. Wang, Z. Wang, S.j. Li, Improved PageRank Influence Algorithm Based on User Behavior, Computer Engineering. 43(2017)155-159.