Open Access Library Journal
2022, Volume 9, e8324
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Construction of Classification Model of Academic Library Websites in Jiangsu Based on Decision Tree Algorithm and Link Analysis Method

Qingxu Liu, Xiaoqing Xia

Pujiang Institute, Nanjing Tech University, Nanjing, China
Email: 1501342473@qq.com

How to cite this paper: Liu, Q.X. and Xia, X.Q. (2022) Construction of Classification Model of Academic Library Websites in Jiangsu Based on Decision Tree Algorithm and Link Analysis Method. *Open Access Library Journal*, **9**: e8324.
https://doi.org/10.4236/oalib.1108324

Received: December 24, 2021
Accepted: January 11, 2022
Published: January 14, 2022

Abstract

Purpose/Significance: This article constructs a website classification model for academic library websites in Jiangsu. From a quantitative point of view, this paper uses a combination of decision tree algorithm and link analysis method in order to analyze how many categories academic library websites can be divided into? Which indicators can have a greater impact on academic library websites classification? And how the indicators are ranked according to importance. **Method/Process:** The link analysis method is used to collect the index data of academic library websites in Jiangsu. After the data is cleaned, the decision tree algorithm is used to analyze the data, and the classification decision tree model is constructed. **Result/Conclusion:** The result shows that academic library websites can be divided into four categories. PC word count, indexed monthly, and Baidu weight have an important influence on the construction of academic library websites' decision tree. Among the various indicators, the impact of PC word count, mobile word count, Baidu weight, indexed monthly and the rest of others decrease in order.

Subject Areas

Library and Information Science

Keywords

Academic Library Website, Decision Tree Algorithm, Link Analysis Method, Classification

1. Introduction

In March 2012, the Ministry of Education issued the "Ten-Year Development

Plan for Education Informatization (2011-2020)", which focused on the construction of educational informatization resources such as "promoting the construction and sharing of high-quality educational resources" [1]. In 2018, the Ministry of Education issued the "Education Informatization 2.0 Action Plan", emphasizing the construction of a new learner-centered education ecology led by informatization, realizing fair and quality education, and proposing to basically achieve the "three alls and twos" by 2022. The development goal of "Higher One", among which "Two High" refers to the general improvement of the level of information application and the information literacy of teachers and students [2]. The university library website is an important part of the implementation of the Ministry of Education's plan and the active promotion of education informatization. The construction level of the university library website is directly related to the degree of advancement of university informatization.

At present, the research of university library websites is a key field of academic research, and scholars at home and abroad have given special attention and research. White E. *et al.* [3] studied the educational effects of portal content on academic librarians and other stakeholders. At the same time, using Nkrumah University of Science and Technology as a case study, library portals were used to enhance user academic exchange training. The role played by aspects. Desmarais B *et al.* [4] described a study conducted at Northeastern State University in the spring semester of 2020. The study detailed the initial stages of the website redesign process, which was collected from online surveys of website design and functionality Feedback. The results obtained after the user experience survey are used to measure respondents' satisfaction with the library website and provide information for future redesign strategies. The data will be used to plan and evaluate the next phase of the university website design project. Brunskill A [5] interviewed 12 college students with disabilities to find out their views on the navigation, search terms, and web interface of academic library portals. These interviews revealed many important considerations surrounding the accessibility and inclusiveness of websites. This compiled a list of recommendations. Domestic scholar Zhang Chao [6] has carried out a multi-dimensional classification of university library websites from the perspective of users. Liu Huiling [7] selected 30 agricultural university library websites to conduct an investigation, analyzed the columns of traditional services, personalized information services, reference consulting services and user classification services, and put forward some suggestions based on the current situation of website services. Song Ailin [8] designed a composite classification navigation system containing TAG tags based on the results of previous investigations, and elaborated the implementation plan, striving to create a three-dimensional digital resource navigation mode that allows readers to participate.

Judging from the existing research results, domestic and foreign scholars have maintained a high degree of attention and research enthusiasm for library websites, but there are still few research results using data mining methods to apply decision tree algorithms to the classification of university library websites. In this paper, the decision tree algorithm and link analysis method are combined and applied to the research of university library website classification, in order to solve the following three problems: 1) Which indicators can have a greater impact on university library website classification? 2) How are the indicators ranked according to their importance? 3) How to classify university library websites based on the decision tree algorithm and link analysis method? The research in this paper can promote the theoretical analysis of the classification research of university library websites in our country from the theoretical level, and can play a reference and reference role for the construction and classification of university library websites at the practical level.

2. Theoretical Basis

2.1. Decision Tree Algorithm

The decision tree algorithm is a classification algorithm based on the data structure of the decision tree [9]. Its basic idea is to gradually dichotomize and refine the original data set through some judgment conditions. Among them, each bifurcation point represents a decision-making judgment condition, and there are two leaf nodes under each bifurcation point, which respectively represent satisfying conditions and dissatisfying conditions [10]. The goal of the decision tree algorithm is to discover the potential classification rules in the data, so its core content is to construct a high-precision, small-scale decision tree. By automatically constructing the decision tree from the data set, it can be used for any instance based on this decision tree to make judgments [9].

2.2. Link Analysis Method

The so-called link analysis method refers to a method developed based on citation analysis, using network links as the research object, using search engines, network databases, and mathematical statistical analysis methods to analyze the distribution law of network links and the links between network information units A quantitative analysis method for regular analysis and research [11]. Link analysis is one of the more important methods in objective quantitative evaluation. It mainly uses the positive affirmation of links between network sites to indirectly evaluate the scientificity and rationality of the website's own information organization and disclosure, as well as the website's influence [12].

3. Data Collection

3.1. Data Source

The list of colleges and universities this time comes from the colleges and universities published in the Jiangsu University Ranking List in 2021. A total of 137 colleges and universities in Jiangsu Province have been collected. As some websites cannot be accessed or 19 data cannot be collected, a total of 118 colleges and universities have been collected. This research will be based on the collected data from 118 university library websites.

3.2. Statistical Indicators

The data collection indicators include 12 items, which are total pages, total links, network impact factor, PC word count, mobile word count, the number of anti-links, index volume, indexed monthly, Baidu weight, mobile weight, and 360 weight, Sogou weight. Total pages, total links, network impact factor index collection methods and their meanings come from existing literature. The total page index data collection method is "site + domain name", and total links number index data collection method is "http:// + domain name", the network impact factor index data is the ratio of the total links to the total pages. PC word count, mobile word count, the number of anti-links, index volume, indexed monthly, Baidu weight, mobile weight, 360 weight, and Sogou weight are all obtained by querying third-party webmaster tools. Baidu weight, mobile weight, 360 weight, and Sogou weight have similar meanings, but considering that Baidu PC, Baidu mobile, 360 search, and Sogou search all have a large number of user groups in country's search engines, the author has not done any indicators, so all the data are collected. The specific meaning of each indicator is as follows.

ZB1 total pages [12]: Refers to the total number of web pages in the website, which reflects the scale of the website and the richness of content to a certain extent.

ZB2 total links [12]: That is, the number of links to the website, which is commonly used to measure the influence and network radiation of the website.

ZB3 network impact factor = total number of links/total number of web pages, reflecting the ability of web pages to be linked.

ZB4 PC word count: Refers to the number of keywords ranked on the computer side.

ZB5 mobile word count: Refers to the number of keywords that have rankings on the mobile terminal.

ZB6 the number of anti-links [13]: Refers to the number of links imported from other websites to a website. Imported links are a very important process for website optimization. The quality of imported links directly determines the number of links that a website has in search engines, weights.

ZB7 index volume: Refers to the number of pages selected by search engines after crawling webpages and filtering through layers, that is, when the website is submitted to the search engine, it will send spiders or robots to the target website to crawl the pages, and the obtained URL will be Sorting out and assigning them to the index library according to a certain level is of great help to website SEO optimization.

ZB8 indexed monthly: Refers to the number of times a certain page of the website was indexed by Baidu in a month.

ZB9 Baidu weight: Refers to third-party platforms such as Webmaster Tools to classify websites based on the estimated traffic brought by Baidu keyword rankings.

ZB10 mobile weight: Refers to third-party platforms such as Webmaster Tools to classify websites based on the estimated traffic brought by Baidu's mobile keyword rankings.

ZB11 360 weight: Refers to third-party platforms such as Webmaster Tools to classify websites based on the estimated traffic brought by 360 keyword rankings.

ZB12 Sogou weight: Refers to third-party platforms such as Webmaster Tools to classify websites based on the estimated traffic brought by the ranking of Sogou keywords.

4. Empirical Analysis

4.1. Data Description

This time, a total of 118 Jiangsu university library websites have been collected to participate in the research. Statistics on the regions and levels of universities have found that Nanjing accounts for the largest proportion, with a total of 39, followed by cities such as Suzhou and Changzhou, which have the largest number. For higher vocational colleges, the number of general undergraduate colleges, private colleges, and key colleges and universities is decreasing in order. The specific statistical results are shown in Figure 1. Use R language software to count and summarize the data of various statistical indicators. The statistical results include the minimum, first quantile, median, average, third quantile, and maximum of each indicator, and various statistics. The specific description results of the indicators are shown in Table 1.

4.2. Build a Regression Tree

Decision tree is one of the most classic data mining methods. It displays the classification process in a tree structure, which is simple, direct and highly interpretable. The decision tree presents an inverted tree shape, that is, the top end is the root of the tree, the bottom end is the leaf of the tree, and a leaf node represents a decision.

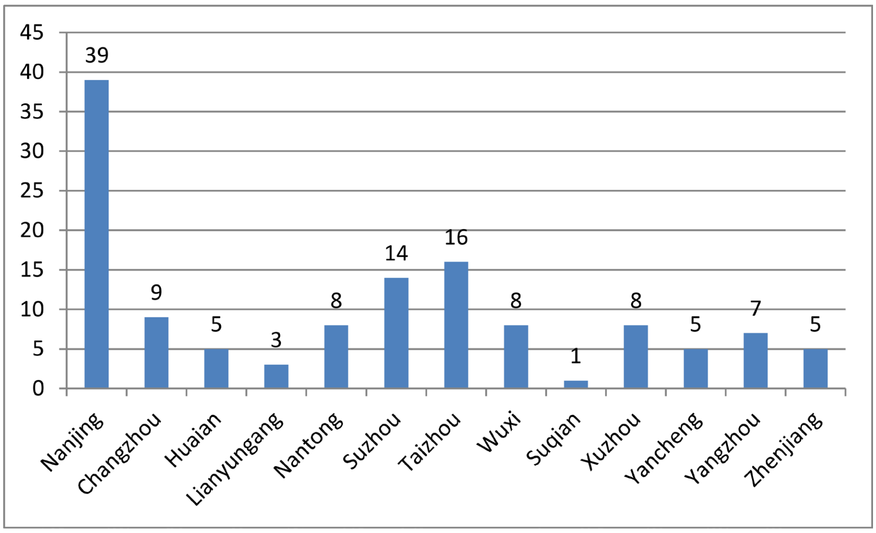

Figure 1. Statistical map of territories to which colleges and universities belong.

Table 1. Raw data statistical index data table.

total pages	total links	network impact factor	PC word count
Min.: 0	Min.: 0	Min.: 0.000	Min.: 0.00
1st Qu.: 0	1st Qu.: 6	1st Qu.: 0.000	1st Qu.: 2.00
Median: 87	Median: 12	Median: 0.000	Median: 11.00
Mean: 13,661	Mean: 174,787	Mean: 323.294	Mean: 58.53
3rd Qu.: 13,675	3rd Qu.: 37	3rd Qu.: 0.435	3rd Qu.: 80.50
Max.: 173,000	Max.: 8,060,000	Max.: 22,555.560	Max.: 587.00
mobile word count	**the number of anti-links**	**index volume**	**indexed monthly**
Min.: 0.00	Min.: 0.00	Min.: 0.0	Min.: 0.0
1st Qu.: 1.00	1st Qu.: 1.00	1st Qu.: 0.0	1st Qu.: 0.0
Median: 8.50	Median: 7.00	Median: 55.5	Median: 0.0
Mean: 29.07	Mean: 21.75	Mean: 17,979.6	Mean: 551.3
3rd Qu.: 26.00	3rd Qu.: 20.75	3rd Qu.: 2196.8	3rd Qu.: 7.5
Max.: 334.00	Max.: 203.00	Max.: 645,000.0	Max.: 35,300.0
Baidu weight	**mobile weight**	**360 weight**	**Sogou weight**
Min.: 0.0000	Min.: 0.000	Min.: 0.0000	Min.: 0.0000
1st Qu.: 0.0000	1st Qu.: 1.000	1st Qu.: 0.0000	1st Qu.: 0.0000
Median: 1.0000	Median: 1.000	Median: 1.0000	Median: 1.0000
Mean: 0.9492	Mean: 1.127	Mean: 0.5678	Mean: 0.8559
3rd Qu.: 1.0000	3rd Qu.: 2.000	3rd Qu.: 1.0000	3rd Qu.: 1.0000
Max.: 4.0000	Max.: 4.000	Max.: 2.0000	Max.: 4.0000

Step 1: Import and read data. First, create a folder, put the collected data, and rename the collected data to gaoxiaoshuju. In order to facilitate data reading, the data file type is converted to CSV format. In order to quickly view and process the data, set the header to a non-reading format, and the read data range is all rows and the seventh to eighteenth columns.

Step 2: Select the core function. In the decision tree algorithm, there are four software packages that are often used, namely rpart, rpart.plot, maptree, and RWeka. The rpart software package is mainly used for the construction of decision trees and the realization of related recursive partitioning algorithms. The basic format of the function rpart is: rpart (formula, data, weights, subset, na.action = na.rpart, method, model = FALSB, x = FALSE, y = TRUE, parms, control, cost, ...), where, Place the formula you want to build the model in formula, the format is $y \sim x_1 + x_2 + x_3$. When the output variable is all variables except y, it can also be represented by y~.; data is the data set to be trained.

Step 3: Generate a decision tree

This decision tree construction uses the rpart function. Compared with other indicators, the "total pages"can better reflect the scale and quality of the website construction, so select “total links”, “network impact factor”, “PC word count”, “mobile word count”, “the number of anti-links”, “index volume”, “indexed monthly”, “Baidu weight”, “mobile weight”, “360 weight” and “Sogou weight”. The variables establish a decision tree, and the type of the selection tree is a regression tree. In the model construction, the total number of web pages is used as the target variable, the data set “dat” is imported into the program, and the classification method is anova.

4.3. Analysis of Decision Tree Results

The decision tree constructed this time is summarized through R language software, and the following research results are obtained. The Variable importance is as follows: It can be seen from the results generated by the software that in order of variable importance (Table 2). In the process of building a decision tree, the variables are ranked in order of importance: PC word count, mobile word count, Baidu weight, indexed monthly, mobile weight, total links, index volume, and number of anti-chains.

It can be seen from Figure 2 that in the process of constructing the classification model, the three indicators of PC word count, indexed monthly, and Baidu

Table 2. Variable importance.

PC word count	mobile word count	Baidu weight	indexed monthly	mobile weight	total links	index volume	the number of anti-links
20	17	15	15	12	7	7	7

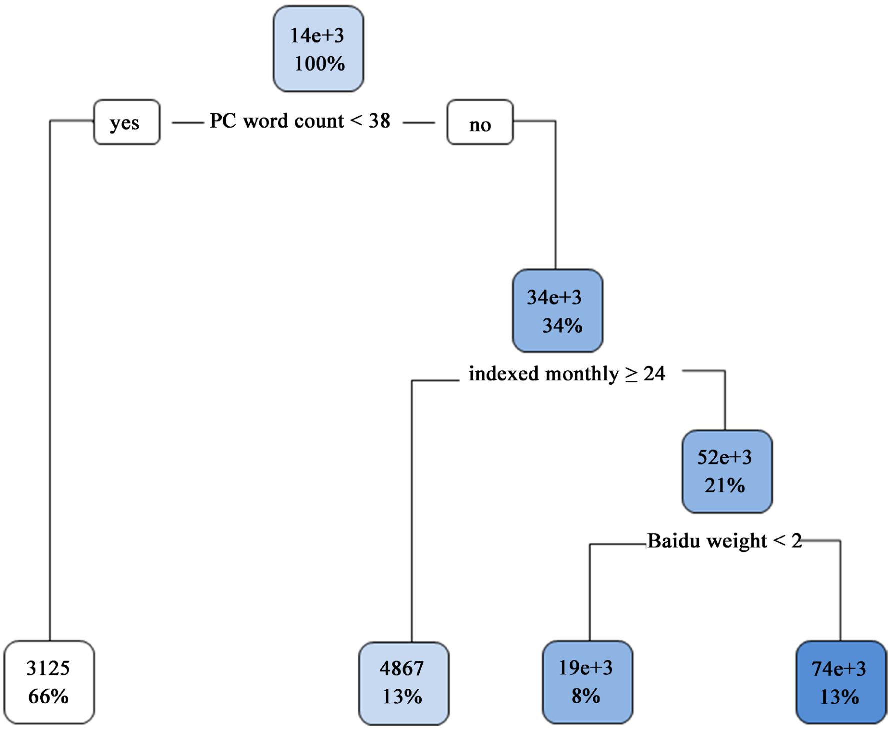

Figure 2. Classification model diagram of Jiangsu university library website.

weight play a vital role in the construction of the entire decision tree. Through the use of decision tree algorithm and link analysis method to analyze the data collected by the Jiangsu University Library website, it is found that the decision tree construction of the university library website produced a total of four leaf nodes. The weight is used as the main classification criterion, and the final classification results are divided into four categories. The first category is that PC word count is less than 38, accounting for 66% of the sample size; the second category is that PC word count is greater than 38 and indexed monthly is greater than or equal to 24, accounting for 13% of the sample size; the third category is that the number of PC word count is greater than 38 And the indexed monthly is less than 24 and the Baidu weight is less than 2, accounting for 8% of the sample size; the fourth category is PC word count is greater than 38 and the indexed monthly is less than 24 and Baidu weight is greater than or equal to 2, accounting for 13% of the sample size. Among the four classifications, the best result is the second category, PC word count and indexed monthly are both high, the poor classification result is the first category, PC word count is less, the first category accounted for a relatively large number, reaching 66%, indicating that the overall quality of the Jiangsu university library website is not high so far, and there is still a lot of room for improvement.

5. Conclusion

In this paper, the decision tree algorithm in data mining is introduced into the study of college library website classification, and the data of relevant indexes of college library websites are collected and analyzed by using the combination of decision tree algorithm and link analysis method. The research results show that this college library website can be divided into four categories, and among all the indicators involved, three indicators, PC word count, indexed monthly and Baidu weight, occupy an important position in the construction of decision tree classification model. Besides, the importance of PC word count, mobile word count, Baidu weight, indexed monthly, mobile weight, total links, index volume and the number of anti-links indicators decreases in order to the construction of the whole decision tree classification model. Admittedly, there are still many shortcomings in this study, such as the small amount of data and the small number of index dimensions, etc. In future research, we will further collect website information, increase the data sample size, and broaden the range of index dimensions to improve the scientificity of the research results.

Fund

This paper is one of the research results of “Construction of Classification Model of Academic Library Websites in Jiangsu Based on Decision Tree Algorithm and Link Analysis Method” (Project No. 2021JSETKT026) of Jiangsu Higher Education Technology Research Association for University Education Informatization.

This paper is one of the research results of “Multi-dimensional Analysis of Information Literacy Research and Perspectives on Disciplinary Growth" (Project No. njpj2021-2-28) of Nanjing Tech University Pujiang Institute 2021 school-level project.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Wang, Y., Zheng, Y.M., Jia, Y.M., *et al.* (2014) Research on the Development Strategy of ICT Resources for Education. *Distance Education Magazine*, 3-14.

[2] Zheng, X.J. (2018) Analysis and Literature Review: The Interpretation of Educational Informatization 2.0 Action Plan (Part 1)—From the Perspective of Teachers in Vocational Colleges. *Journal of Guangxi Vocational and Technical College*, 51-61+2.

[3] White, E. and King, L. (2021) Investigating the Development of a Research Portal as Part of an Academic Library Website for Scholarly Communication Guidance in a Public University in Ghana. *International Information & Library Review*, **53**, 157-169. https://doi.org/10.1080/10572317.2020.1805577

[4] Desmarais, B. and Louderback, P. (2020) Planning and Assessing Patron Experience and Needs for an Academic Library Website. *Journal of Library Administration*, **60**, 966-977. https://doi.org/10.1080/01930826.2020.1820283

[5] Brunskill, A. (2020) "Without That Detail, I'm Not Coming": The Perspectives of Students with Disabilities on Accessibility Information Provided on Academic Library Websites. *College & Research Libraries*, **81**, 768-788. https://doi.org/10.5860/crl.81.5.768

[6] Zhang, C. (2011) Multi-dimensional Classification and Application of User-Based University Library Website FAQ. *Library Journal*, **33**, 107-109.

[7] Liu, H.L. (2018) Investigation and Analysis of the Service Columns of Agricultural University Libraries—Taking 30 CALIS Agronomy Center Member Library Websites as Survey Objects. *Henan Library Journal*, **38**, 105-108.

[8] Song, A.L. (2014) Design of Library Digital Resource Compound Category Navigation System. *Journal of Xichang College* (*Natural Science Edition*), **28**, 70-72.

[9] Wang, Z.J., Zhou, P. and Han, Z.B. (2013) Research on the Model of Competitor Recognition Based on Decision Tree Algorithm. *Information Theory and Practice*, 1-5+24.

[10] Su, W., Jiang, F.F., Zhu, D.H., *et al.* (2015) Extraction of Maize Planting Area Based on Decision Tree and Mixed-Pixel Unmixing Methods. *Journal of Agricultural Machinery*, 289-295+301.

[11] Qiu, J.P. and Li, J. (2007) The Shortcomings of Current Link Analysis Tools and Their Solutions. *Information Science*, 641-647.

[12] Huang, K.M., Fan, Z.J., Lu, S.J., *et al.* (2014) A Comparative Study of Chinese and American Think Tank Websites Based on Link Analysis. *Information Theory and Practice*, 129-133.

[13] Cheng, H.Y. (2016) Comparative Research on My Country's Science and Technology Media Websites—Taking Tiger Sniffing Network, Titanium Media and 36Kr as Examples. *Journalism and Communication*, 7-8.