Information Collection and Data Mining Technology of Open University Distance Education Website

Junrong Guo(✉)

Hebei Open University, Shijiazhuang 050051, China

Abstract. With the development of information technology, how to reasonably develop the information collection and data mining of distance education courses in Open University to make network courses play its great role and potential is a subject worthy of study. At the same time, it also provides more reference for the design of education website. It is of great significance to use data mining technology to process the data information and access information of distance education website.

Keywords: Open University · Distance education · Information collection · Data mining

1 Introduction

It is one of the important goals of popularizing basic education and constructing lifelong education and learning socialization after the rapid development of information technology and the arrival of information society to establish modern distance education website to transfer knowledge and information to learners and provide relevant services [1].

The so-called distance education is not just a new thing. Taylor, a western scholar, thinks that distance education has gone through five stages and points out the characteristics of its technological application: the first stage is the correspondence teaching mode, which is mainly based on printed teaching materials; the second stage is the multimedia mode, which uses printed materials, sound recording, video recording, computer and interactive video recording technology for learning; The third stage is the electronic remote mode, using audio conference, video conference, acoustic communication, radio/TV/radio and teleconference technology for learning; the fourth stage is a more flexible learning mode, using interactive multimedia, Internet based access to world wide web materials, computer transmission communication technology as the media of education. In the fifth stage, the key point is to establish an online automatic response system. At this time, distance educators focus on cost-effectiveness and teaching quality. In a word, distance education in the 21st century aims at the development of open, flexible and lifelong education [2, 3]. It is not only the continuous development of traditional

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 391, pp. 158–167, 2021.
https://doi.org/10.1007/978-3-030-87900-6_19

education, but also a great change to traditional education. It breaks the limitations of traditional education and has incomparable advantages over traditional education, And take it as an important form to realize the popularization of higher education, continuing education and lifelong learning. The organizational form of Open University is shown in Fig. 1.

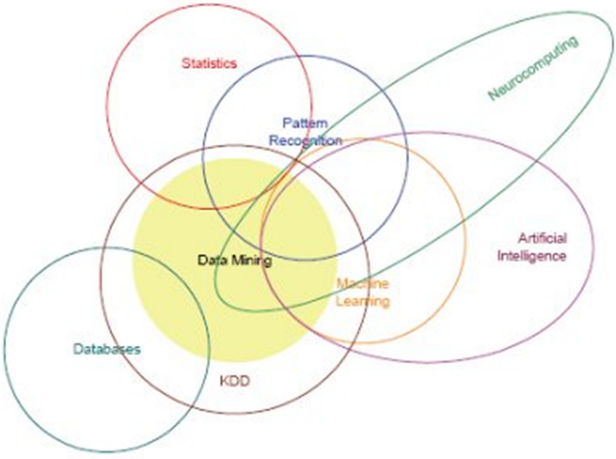

Fig. 1. The organizational form of Open University

2 Log Preprocessing and Algorithm Description of Distance Education Website

The information on distance education website is presented in the form of unstructured or semi-structured, and data mining needs structured data. The original log file is a simple flat text file, which contains some incomplete, redundant and wrong data. It needs to be processed, otherwise it will directly affect the effect of mining [4]. In addition, the implementation of mining algorithm also needs the support of standardized data sources, so the format of data storage collected in the first stage needs to be adjusted to suit the mining method. Therefore, the data preprocessing stage is the basis of the whole process of Web log mining and the premise of implementing effective mining algorithm, which plays a very important role in Web log mining. This chapter focuses on the preprocessing process and algorithm of distance education website log.

2.1 Definition and Properties of Association Rules

A typical example of association rule discovery is shopping basket analysis. This process finds the relationship between different products in the shopping basket and analyzes customers' buying habits. By understanding which products are frequently purchased

by customers at the same time, the discovery of this association can help retailers make marketing strategies [5, 6]. This is the original form of association rule mining.

The support degree of implication x → y refers to the ratio of the number of records and the total number of records that simultaneously support item set X and Y in the database. It describes the probability that X and Y occur at the same time and indicates the importance of the rule. The support degree of association rule x → y is defined as:

$$Support(X \rightarrow Y) = \text{supp}(\text{X} \cup \text{Y}) = |\text{T} \in \text{Dand}(\text{X} \cup \text{Y}) \subseteq \text{T}|/|D| \quad (1)$$

The confidence degree of implication formula x → y refers to the ratio of the number of records supporting both X and y to the number of records supporting X in the database. It can be understood as the probability of occurrence of y when x occurs. It indicates the correctness of the rule. The confidence degree of association rule x → y can be expressed by the following formula:

$$Confidence(X \rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) \quad (2)$$

2.2 Introduction of Apriori algorithm

Apriori algorithm graph is a famous association rule mining algorithm, which uses the downward closure of frequent itemsets, that is, any subset of a frequent itemset must be a frequent itemset, and any superset of a non frequent itemset must be a non frequent itemset, so as to achieve the purpose of pruning frequent itemset candidates.

Apriori is a width first algorithm, which can find all frequent itemsets through multiple scans of database D. in each scan, only all itemsets with the same length (that is, the number of items in the itemset) are considered. In the first scan, Apriori algorithm calculates the support of all single items in D and generates all frequent item sets with length of In each subsequent scan, firstly, new candidate itemsets are generated based on all frequent itemsets generated from k-1 scans. Then, database d is scanned to calculate the support of these candidate itemsets, and the itemsets whose support is lower than the minimum support given by users are deleted. Finally, all frequent itemsets with length of K are generated [7]. Repeat the above process until no new frequent itemsets are found. So Apriori algorithm is mainly composed of two processes, that is, out connection and pruning.

2.3 Partition Technology

Partition technology only needs two database scans to mine frequent itemsets. Savasere and others designed an algorithm based on partition. This algorithm first logically divides the database into several disjoint blocks, considers one block each time, and generates all frequency sets for it. Then the frequency sets generated are combined to form all possible frequency sets. Finally, the support of these itemsets is calculated. The size of the blocks should be selected so that each block can be put into main memory, and each stage only needs to be scanned once. The correctness of the algorithm is guaranteed by every possible frequency set at least in a block [8]. The algorithm discussed above can

be highly parallel, and each block can be assigned to a processor to generate frequency set. After each cycle of generating frequency set, processors communicate with each other to generate global candidate k-item set. Usually, the communication process here is the main bottleneck of algorithm execution time; on the other hand, the time for each independent processor to generate frequency set is also a bottleneck.

Confidence level:

$$C\% = Confidence(X \rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) * 100\% \quad (3)$$

The purpose of data mining is to find out the representative and credible rules. Support indicates the importance of the rule in all data, and credibility means the trustworthiness of the rule. If the degree of support is too low, the rule is not general, if the degree of confidence is too low, the rule's trustworthiness is poor.

3 Current Situation Analysis

3.1 Application Basis

This paper introduces the teaching reform of DACUM method based on CBE Mode to cultivate senior applied talents of geographic information data acquisition and processing specialty.

Higher engineering college is a part of higher education in China. It is the direction of higher education reform in China to cultivate cross century senior applied talents for the 21st century. In 1998, the Ministry of Education approved the first batch of experimental teaching reform of the specialty of geographic information data collection and processing. In order to meet the requirements of knowledge economy in the 21st century, we adopt the CBE (competency based education) ability target education mode to develop and practice the new education mode for the experimental specialty of geographic information data acquisition and processing [9–11]. Through the practice of teaching reform, this paper explores the mode and way of cultivating cross century high-tech post type talents, forms the school running characteristics of cultivating new technology post type talents in higher engineering colleges, and transports high-tech and high-quality applied talents for the construction of information superhighway and digital earth in China.

3.2 Development of Teaching Mode

In the practice of teaching reform, according to the geographical education theory put forward by experts outside school.According to the DACUM (developing a curriculum), this paper designs the teaching plan, arranges the teaching contents and methods, and establishes the ability based education mode for the specialty of geographic information data acquisition and processing in engineering colleges. The teaching reform of this major focuses on the basis of public courses, foreign language and computer teaching, with the knowledge and technology of Surveying and mapping, geography as the guide, and the cultivation of the three abilities of geographic data collection, data editing and geographic information management as the goal. The theory and practice of geographic information system run through the whole teaching process, Improve the structure of

students' knowledge, ability and quality of professional system development and maintenance spatial analysis [12]. According to the responsibility/task table, the decomposition and comprehensive choice of the curriculum system of various disciplines is the starting point of building a new curriculum. Arming the new curriculum with new high-tech knowledge, retaining the essence of the original subject curriculum system, appropriately adding new content; developing new teaching plans, syllabus and compiling some new teaching materials; in theory, striving to build a new teaching system, update and enrich the teaching content, innovate teaching methods and teaching means, and compile suitable teaching materials.

3.3 Responsibility Requirements

According to the requirements of duties/tasks, the geographic information data acquisition is divided into four modules: ground spatial data acquisition, global positioning, map digitization, digital photogrammetry and remote sensing. Among them, the ground spatial data acquisition module is composed of three major technologies: topographic survey technology, geodesy foundation, and geodetic instrument operation. Various technologies are based on the framework of modern new technology system, boldly giving up a lot of old and unused knowledge and adopting a brand-new framework. The instruments required for ground spatial data acquisition are organized independently according to the type, function and precision series, Systematically and comprehensively teach the basic principles, functions, usage and operation skills of various instruments. It avoids the repeated, scattered and unsystematic teaching of geodetic instruments in the traditional course of topographic survey and control survey. Make the new curriculum module structure system more compact, systematic and perfect. In this course, the teaching of Surveying and mapping instruments is based on the mastery of conventional instruments, with electronic total station, electronic level and electronic handbook as the theme framework, to meet the needs of the development of new surveying and mapping technology.

4 Preliminary Evaluation of cbe-dacum Education Model

In 1999, the major of GIS data acquisition and processing first enrolled 36 students. After two years of teaching practice, from the students' learning and knowledge, the effect is satisfactory. It basically meets the requirements of the syllabus, and has obvious teaching effect compared with the similar courses of Surveying Engineering Grade 99.

4.1 The Study Time Is Greatly Shortened and the Effect Is Good

According to the teaching plan, "Xincai" 99 pilot class will concentrate the series of professional courses of ground spatial data acquisition in one academic year, and carry out relevant practical teaching at the same time. According to the plan, three courses of topographic survey technology, geodesy foundation and surveying instrument operation technology were set up in the second semester. Digital mapping technology was set up in the second semester. The above courses were set up in the teaching plan of surveying

engineering class 99, which entered the school together with the "Xincai ν 99" pilot class, respectively in the first, second, third, fourth and fifth semester [13–15]. The course names and main contents of the two majors are shown in the table below.

The time of learning and mastering the content of "Surveying and mapping discipline" is greatly shortened, and the learning is more concentrated. The curriculum structure system has been adjusted, especially emphasizing the current and practical technology, paying attention to the combination of knowledge system and technological process, and improving the teaching efficiency and effect through the renovation, reduction and reconstruction.

4.2 The Practice Links Have Been Greatly Increased, and the Operation Ability Has Been Enhanced

In the teaching plan of Xincai 991, the practice link is increased, and the total number of practice weeks is 43 weeks (99 classes of Surveying Engineering), accounting for 39.2% of the total number of teaching weeks. The classroom teaching follows the practice teaching closely. For example, in combination with the practical courses, the experiment was taught independently, and five experimental weeks were set up for centralized training. According to the skill test, most of the students have a good command of it, which basically meets the requirements of the experimental syllabus. Taking the computer grade examination results as an example, 75% of the 99 classes of "Xincai" have passed the second and third level examinations for non computer majors, and the passing rate exceeds the average passing rate of undergraduate education in Jiangsu Province. In the map digitization and GPS survey professional skill appraisal organized by the professional skill appraisal department of the Ministry of labor and social security of the people's Republic of China, 45 people in 99 classes of Xincai obtained the intermediate skill appraisal certificate issued by the Ministry of labor and social security of the people's Republic of China with a high passing rate of 98% [16]. It can be seen from the analysis table 6 of the weekly examination results of professional measuring instrument operation experiment that the rate of reaching the standard at one time is more than 70% by adopting the same skill test standard of measuring engineering specialty, and all of them can meet the requirements of measuring engineering specialty through three skill tests.

5 Application of Information Collection Technology in Distance Education Website

With the accumulation of time, the webh log file in the web server will be larger and larger, which contains more and more customer information. The web log records the information of users visiting the site, including the number of users, the URL of the requested file, the protocol version number, the number of bytes transferred, the URL of the reference page, etc. Combined with the user database, the effective collection and analysis of Web log files can not only effectively evaluate the performance of the website, but also provide decision support for educational website service positioning and improving user relationship.

5.1 Application of Information Collection Technology in Distance Education Website

The application of information collection technology in distance education website mainly includes the following aspects:

(1) Through the data collection of the user's past visit history, we can know the user's frequent visit path, obtain the needs of the visitors, more fully understand the needs of the users, classify the users, and provide targeted services, which is conducive to improving the user's satisfaction and recognition, and truly realize the design of personalized website for users with the user's needs as the guide.
(2) Analyze the potential needs of users and optimize the service mode of distance education website. According to the historical data of users, we can not only predict the demand trend, but also evaluate the change of demand trend, which is helpful to improve the utilization rate of distance education website.
(3) Optimize the web site, improve the organizational structure of the web space. Website designers can no longer completely rely on the qualitative guidance of experts to design the website, but modify and design the structure and appearance of the website according to the information of visitors, find out how to optimize the organizational structure of a website, and determine which pages to pre transmit to the user, so as to improve the efficiency of the website [17].

A visual data acquisition and analysis system based on web is designed. According to the previous discussion, we use the object-oriented design method to design a visual acquisition and analysis experimental system for text content. The system can flexibly process all kinds of information, such as filtering useless information according to the needs of users, or cataloging, sorting and saving the collected information according to the needs.

5.2 Simulation Analysis

In this section, we first give the basic structure of data mining, as shown in Fig. 2, and then use this basic structure to do simulation. In this basic structure, we can see that a basic data structure contains two parts: database and mining. So next, we will simulate these data, as shown in Fig. 3. The simulation data can be obtained from Sect. 2.

6 Main Application Effects

6.1 Most of the Requirements Are Solved

Cbe-dacum teaching development mode can effectively solve the problem that higher engineering education does not meet the needs of employers for high-tech post talents under the socialist market economy environment. Through the market-oriented reform and exploration of the professional teaching and training program, the most important thing is to shift the focus of the training objectives to how to adapt to the employing units and help students to obtain employment [18–20]. The primary key of cbe-dacum mode reform is how to solve the demand of talent market and the problems of students.The interface between employment intention and school education plan.

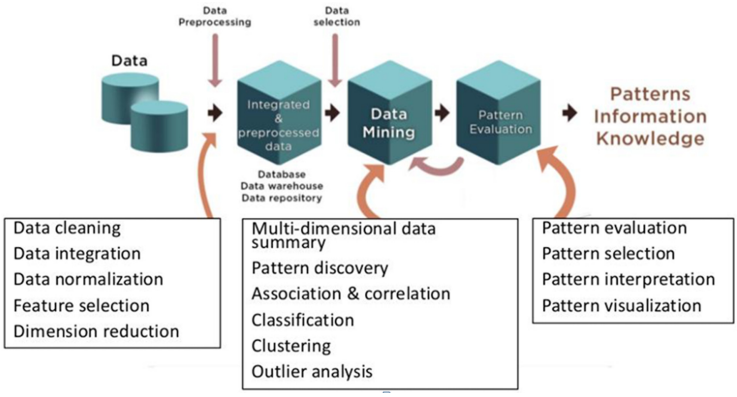

Fig. 2. Structure of data mining

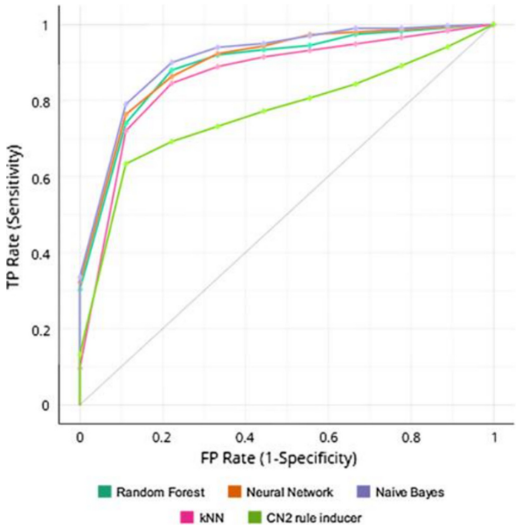

Fig. 3. The simulation with data mining

6.2 Skill Development

CBE → DACUM teaching mode has a set of standardized and standardized ability training methods, which is helpful to the cultivation of students' skills. Through the combination of production and learning, the skills training will be socialized, so that the students' skills will eventually get the identification of the application position and the market certification. It shortens the distance between employers and schools, enhances the transparency of teaching, and is conducive to the two-way choice between employers and students.

6.3 Domestic and International Integration

Combined with the national conditions and learning from foreign cbe-dacum teaching mode, the teaching goal of foreign cbe-dacum teaching mode is relatively single, which is due to the developed education and less job types. In China, geographic information data acquisition and processing is the foundation of Surveying and mapping, computer, geography, management and environment, and a platform for interdisciplinary infiltration. Students should adapt to more jobs. Therefore, we have formulated 20 responsibilities and 203 tasks, which are much broader than the training objectives of foreign majors [20–23]. Cbe-dacum mode has no clear provisions on theory teaching, and it is essential to ensure sufficient theoretical basis. How to deal with the relationship among the theoretical basis, professional knowledge and practical skills in the teaching plan according to the national conditions is a problem that needs to be deeply discussed and solved. It is also a common situation in China that the combination of industry and education is hot. How to make the industry hot, we must explore a new way according to the characteristics of China's socialist market economy.

Specialty construction is a huge system engineering, which needs to constantly update the concept and make unremitting efforts to achieve the expected goal.

7 Conclusion

Through the mining of distance education website, we can extract the useful knowledge we need from a large number of log information. Through the analysis of the total user access behavior, frequency, content and so on, we can get the general knowledge about the group user access behavior and way, provide personalized services for users, improve the service efficiency of the system, improve the website design structure, determine the user needs of the distance education website, and effectively evaluate the website, Through the understanding and analysis of these user characteristics can help to carry out personalized education service activities.

References

1. Deng, P., Xiaoshe, D., Maishun, Y.: Mining frequent access patterns from web data. J. Xi'an Jiaotong Univ. **36n0.6**, 631–644 (2002)
2. Shi Jianchen, W., Lina, W.L., Yiling, Y.: Research on mining user browsing patterns from web logs. J. Xi'an Jiaotong Univ. **35**(6), p621-624 (2001)
3. Baoshu, C., Qimin, J.: Data preprocessing in web data mining. Comput. Eng. **28**(7), 14–19 (2002)
4. Guixia, J.: Research on association rules and application in data mining, Master's Thesis. Lanzhou University of Technology, Lanzhou, Two Thousand and Six Point Four (2012)
5. Dynasty, S., Xiuying, S.: Asp.net. Comput. Knowl. Technol. **14**, 325–326 + 344 (2011)
6. Party Construction Online (within the province). Jianghuai **16**(1), 36–37 (2011)
7. Xiaohong, Y., Fangyu, L.: Research on digital resources integration of modern distance education for rural party members and cadres. China Dist. Educ. **21**(1), 88–91 (2011)
8. Qijie, G.: The enlightenment of Japanese and Korean models on online Chinese distance teaching. Capital Foreign Lang. Forum **5**, 658–665 (2001)
9. Songhe, Y.: Research on the content framework of modern distance education network course construction. J. Guangdong Radio Telev. Univ. **19**(6), 13–18 (2010)

10. Special training for application administrators of "Qilu pioneer" distance education website in Dongping County. Guide Getting Rich Sci. Technol. (34), 10 (2010)
11. Jie, C.: Comparison of modern distance education websites. Lib. Sci. Res. **4**(22), 46–48 + 80 (2010)
12. Modern educational technology. **20**(10), 153–158 (2010)
13. Aiyun, J., Baofeng, Z., Pingzhu, W., Jingchao, Z.: Design and implementation of computer aided design distance education website. China Sci. Technol. Inf. **32**(19), 232–233 (2010)
14. Yajun, L.: Analysis on the current situation of network question answering system. J. Xichang Univ. (Natural Science Edition) **24**(3), 68–70 (2010)
15. Minggang, Y.: Analysis and countermeasures of campus culture construction in modern distance education pilot colleges — taking the website construction of four pilot colleges in Shanghai as an example. China Dist. Educ. **8**(9), 53–57 (2010)
16. The characteristics of teaching Chinese as a foreign language in Gansu Province. **3**, 26–35 (2010)
17. Hong, S.: Research on the current situation and strategies of online education resources construction in primary and secondary schools. Lib. Work Res. **21**(7), 107–109 (2010)
18. Juncai, G.: Solutions for browsing distance education resources in LAN. China Educ. Technol. Equip. **22**(20), 41–42 (2010)
19. Guoqing, L.: Analysis of the characteristics of American primary and secondary education websites and its enlightenment to China. Audio Vis. Educ. Prim. Sec. Schools **15**(z2), 43–45 (2010)
20. Hafezi, S., Mehri, S.N., Mahmoodi, H.: Developing and validation a usability evaluation tools for distance education websites: Persian version. Turk. Online J. Dist, Educ. **11**(3), 22–29 (2010)
21. Moore, M., Dongjie, X.: New power of network: teaching method and organization. Open Educ. Res. **16**(3), 100–109 (2010)
22. Yan'er, T.: An analysis of the development of Chinese international education based on 3G. Res. Mod. Dist. Educ. (3), 26–31 (2010)
23. Lanlan, J.: Research on new type of farmers' Entrepreneurship Education Based on Network. Zhejiang Normal University (2010)