Application of Data Mining in Personalized Remote Distance Education Web System

Yue Li*, Jian Sun and Wei Qiang

North China University of Science and Technology, Tangshan, China

Abstract: This paper used web log mining technology for different users to take a different service policy and provide different and individualized services . Also the paper used basic theory of application of web data mining for remote education process in the distance education. After analysing systematic framework of distance education, the paper proposed web data mining application model in remote education and described all module function. The model used log information and user information to get interesting mode, and applied this mode to remote education system in order to improve personalized service and can be more conducive to learners. It is beneficial to help improve the content and the site topology update which reflects more dynamic and personalized than common distance learning.

Keywords: Data mining, distance education, personalized, web system.

1. INTRODUCTION

With the advent of the information age, modern distance education has better developed as the main way of e-learning and become an important foundation for building a learning society. Also it plays an important role in the implementation process of higher education popularization in China. Resources in e-learning environment which are widely shared and provide effective support for collaborative learning gave rise to a gradual change in the learning mode that is the centre changed from "teacher" to "learners" and emphasized personalization and adaptation of the learning environment [1].

The World Wide Web has been widely accepted by people due to its rich hypertext information (graphics, sound, animation and video), unified platform (the browser) as well as the easy usage. Computer-aided teaching system experienced the same change as a complement to a new school of education. Distance teaching system based on World Wide Web has been developed [2, 3]. This teaching mode has changed the traditional teaching in spatial and temporal boundaries, by letting people experience to free access of knowledge of fun, has achieved teaching process of interactive sex, fast update of teaching content , and teaching media using various valuable lesson plans. The teaching service is available for more users and students have opportunities to select the best school learning program, and select the best teacher and like most courses, various excellent of teaching resources and teaching courseware can be shared [4].

Distance education involves computer technology and application of network technology. It is based on a teaching model of modern information technology platform, being a complement to traditional education. Distance education as a learning tool is used in higher education, vocational education and adult education. Iit is a good means to provide lifelong education. There is a great deal of difference, mainly reflected by different individual learning goals, learning abilities, and different cognitive styles. This necessarily determines that the distance education must be a kind of individualized education and distance learning must also be an adaptation of the individual learning and needs individual teaching [5, 6].

In distance teaching system based on WEB, the course providers to design good software are stored on the server and wait for the user to access. The user can access at any time, using any browser on a computer with access Internet/intranet servers according to their own interest to enrol themselves as students in a study course. However, the traditional system of remote education made the system as the centre and did not fully take into account the student's needs and habits and required the user to adapt to the system instead of the system adapt to the users' needs; it was not fully in accordance with the law of education, generally lack user guide which is basically the electronic copy of a book. Moreover, creating interactive teaching was not obvious, and involved many other issues. Distance education system has low human rigidity mainly due to lack of user behaviour analysis, as it ignores the particularities of different user behaviour, and does not give user feedback,thus reducing the interest of users [7, 8].

In recent years, data mining has made it possible to deal with large database found. With the development in Internet and Web technologies, application of data mining based on Web knowledge permeated many aspects of social life at an alarming rate. Combining Web and data mining technology, the application of data mining in the Web has introduced another field of data mining which is knowledge discovery through Web data mining [9, 10].

*Address correspondence to this author at the North China University of Science and Technology, Tangshan, China; Tel: 13832969848;
E-mail: 769683614@qq.com

1874-110X/15 2015 Bentham Open

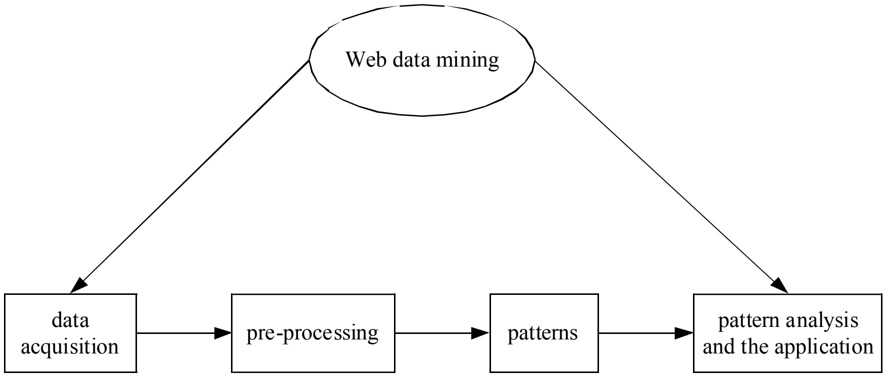

Fig. (1). Web data mining process.

Web mining is the www-related resource and extracts interesting, useful patterns and hidden information. It is the function of data mining technologies to utilize Web documents and Web services to automatically find and extract interesting information [11]. Application of web mining in distance education site, mine useful information and model for the analysis of students of distance education, their learning behaviour and learning, to improve the design of distance education sites and provide a personalized learning environment for distance education students, which is very meaningful.

Next, the data and methodology used in the study are described in more detail.

2. DATA MINING

Data mining from large-scale data has potential rules and requires skills for extracting useful knowledge. Because the database deals with knowledge, it is also known as knowledge discovery in databases (KDD). Data mining not only extracts knowledge, but is able to identify what is unknown, given that the knowledge is "explicit", and due to easy usage and application and being comprehensive, it has gained widespread attention.

Knowledge discovery is a process which requires use of integrated systems, and data mining is just one link. Data mining is just one step in the whole knowledge discovery system, but it is one of the most important and most critical steps. We can say that data mining is the core technology of knowledge discovery, data mining algorithms directly affects the quality of knowledge discovery. Good data mining algorithms can perform quick and efficient data mining because internal rules enable the whole system to extract more useful knowledge, however, the speed of data mining or knowledge discovery was observed to be inadequate for effective knowledge discovery. Knowledge discovery in databases was identified from the database effectively by using novel and potentially useful and understandable model based on advanced process.

Web data mining generally can be defined as a www-related resource which extracts interesting, useful patterns and hidden information. In general, web data mining can be divided into three categories: web content mining and web usage mining, and web structure mining.

2.1. Web Content Mining

Web content mining is the mining of web page content. It includes: (1) intelligently extracting information from the www search tools. (2) The database methods: reconstruction of semi-structured web to make it more structured, therefore standard database query mechanism can be used for the analysis and data mining methods. (3) Mining the content of HTML pages, the text in the page through text mining, multimedia information mining of multimedia information on the page including the content of the page classification, clustering and association rules.

2.2. Web Structure Mining

Web structure mining uses the link of the web document which reveals useful structural pattern contained in these documents. It is considered a web of processed data. Hyperlinks in the documents reflect a link between documents, such as its content, subordination, references, *etc*. More representative of these tools is the page rank. I Information in the links of the documents is used to find relevant web pages.

2.3. Web Usage Mining

Web usage mining is left to the user when accessing the web server access logs for mining, namely access to the user's access to the web site for mining. Mining object is included in the server, such as server log data log. Mining methods include the following: (1) path analysis; (2) discovery of association rules and sequence patterns; (3) cluster and classification. Web usage mining can automatically discover users' access to web pages from the web server mode that a group of users or an individual user finds interest in and accesses .

Data mining based on web generally includes the following processes: data acquisition, pre-processing, patterns, pattern analysis and the application.

Web mining process consistent with the preceding data mining process can be divided into following steps; data mining of business objects, data preparation, data mining and analysis and validation of results in several stages as shown in Fig. (**1**). (1) Business objects: data mining is carried out in initially before exploring issues of evaluation, prediction, goals and basic structure of data mining. (2) Data preparat-

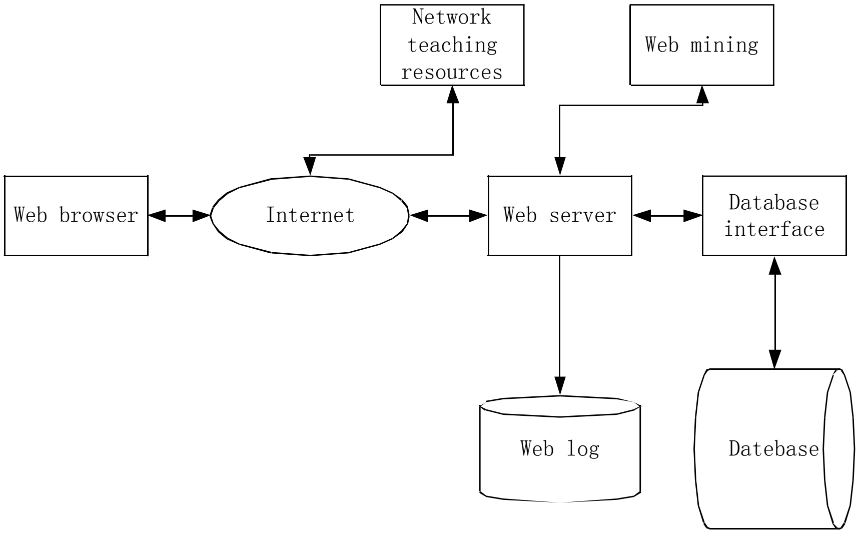

Fig. (2). Distance education model based on web mining.

ion: this involves web data mining of user's background information and Web page which consists of two parts; User's background information and u information regarding user's registration. This records user's private information, social background, information regarding profession, interests, and hobbies. As these informations are personal , many users do not register themselves on website Shang registration real content, this on to zhihou of data mining caused has obstacles;. Another section involves information about user's browsing log. This information reflects user's preferences, and can also give information about user's social background,. Web page file is a key component of the site, users through their acquisition of knowledge. Data preparation requires background information about the user and Web page files to extract information to have access to the data. (3) Data mining involves selecting and designing appropriate data mining algorithms, for data analysis and processing of data in the preparation phase to , find the expected pattern. (4) Analysis of authentication phase verifies the accuracy of the data mining phase results. In case of errors, the preceding steps are rolled back and are amended and if the results are correct, further analysis, interpretation, guide site design and renovation are followed.

Based on remote education platform combined with web data mining technology, understand and master students learning of interest, and browse mode, and learning status, and need of navigation help, get conducive to remote education of fresh mode and rules, guide teaching material of arrangements and courseware of design and improved, improve remote education of quality, building a perfect of online virtual teaching system, makes students of remote education learning mode more intelligent of, and personalized.

With the application of web mining in distance education technology, we can take full advantage of the information site , for constructing effective education system. Based on web data mining of remote education model shown in Fig. (**2**), model in traditional of based on web of remote education mode increased web mining technology through on website log and background database of integrated analysis, found user of learning law and learning preference, will these information feedback to line teachers and courses design personnel and website management personnel, these people will can proposed website of improved programme both frame site (Fig. **2**). So we can set up a personalized intelligent education platform.

In general, the application of data mining technology in the education system has made the decision makers use data mining techniques to detect shortcomings of education, to predict the future trend for educational development, and to improve the quality of education .

3. APPLICATION REAEARCH

3.1. Bayesian Classification

Bayesian classification is a very important application in personalized learning system. Classification in data mining is a very important task. The purpose of classification is to build a classification model, and this model can be mapped to data items in the database in one category . Classification is a common problem. There are many different applications for classification. For example, a classification model can be created, security classification or risk on bank lending based on e-mail headers and content check out spam; In the MRI results, to distinguish between tumours as being malignant or benign, the results are categorized based on the shape of galaxy Bayesian classification method was proposed for the classification of learner's learning style, to establish a personalized learning system to provide information for decision-making.

Briefly, classification involves learning function f, and x maps each property set to a previously defined class label y. Data classification is divided into two steps: (1) building a model, and providing description of the scheduled data collection or set of concepts. The description involves the analytical description of the attribute database tuples to construct the model. Tuples are also known as sampling, instances, or objects. Tuple form of data analysis is used for model training dataset. Each sample also has a specific corresponding label, as the provision of training samples for each class label, which is also known as supervised learning.

(2) The classification using the model, evaluates the predictive accuracy of the model followed by a, a simple method that uses class label sample test sets. These samples are randomly selected, and are independent of the training samples. The model calculates the percentage of test samples which are accurate and classifies them. For each test sample, samples of known class label and the comparison of model prediction.

Bayesian network is based on dependencies between variables, using graph theory to represent the joint probability distribution of the variables collection of graphical models. Bell leaves republican network is based on a no to no ring showing conditions of probability of distribution, allowing the variable of subset to define class conditions independence in which, each knot point is described as a random variable, and between the two knot points, there is an article arc. This two knot points phase corresponds to the random variable dependent on probability, instead of the description the two random variable is conditions independent of In the network, a node x has a corresponding conditional probability table to represent node x in its parent node each having a possible value of the conditional probability. Bayesian network has two main components: (1) directed acyclic graph showing dependency between vectors; each node represents a random variable, and each ARC represents a probability dependent. (2) Probability table with a parent node and its associates. Bayesian network modelling consists of two steps: (1) creation of a grid structure. (2) Estimation of the probability of each node in the table of probabilities. Network topology can be obtained through the expert knowledge in the field of subjective coding.

Bayesian classification is a combination of statistics and Bayesian network classification method. It is based on the following assumptions: the probability distribution of the studied variables is based on the probabilistic reasoning and observed data in order to make the best decision and the probability for a given sample belongs to a particular class. Bayesian classification method has the following main features: (1) it makes full use of knowledge in the field and other information can be explicitly calculated if probability result is the combination of domain knowledge and information data. (2) It uses digraph representations, and arc demonstrates the dependency between variables, represented by probability distributions dependent on the strength of the relationship. Representation is conducive to the understanding of the knowledge in the field. (3) Under normal circumstances, all steps are involved in the classification of the property, and potentially play a role in the process of classification. (4) It can be incremental learning.. The data can be incremented or decremented to estimated the probability of a hypothesis, and can easily deal with incomplete data. (5) Bayesian classification process is a discrete object.

Suppose that the sample space of experiment is S. A, and B represent the event. The probability of event A is $P(\mathrm{A})$, , the probability of event B is $P(B)$. The probability of event A and B is $P(\mathrm{A}\cdot\mathrm{B})$. Therefore, the following formulas are obtained:

$$P(B/A)=\frac{P(A\cdot B)}{P(A)} \tag{1}$$

The multiplication theorem calculates conditional probability :

$$P(A\cdot B)=P(B/A)P(A)=P(A/B)P(B) \tag{2}$$

The total probability formula is as follows:

$$P(A)=\sum_{i=1}^{n}P(\mathrm{A}/\mathrm{B}_i)\,\mathrm{P}(\mathrm{B}_i) \tag{3}$$

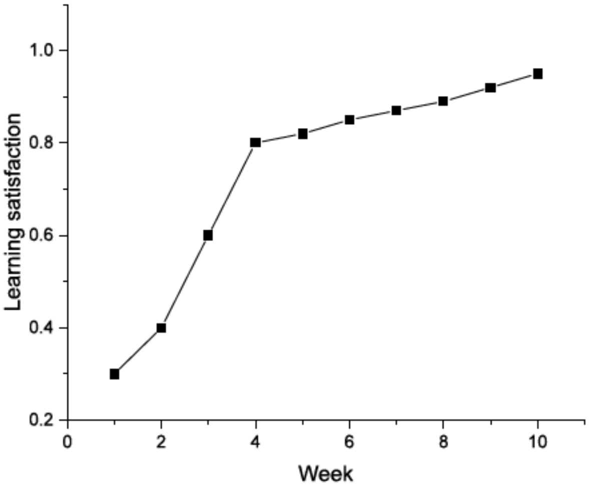

Fig. (3). Learning satisfaction based on data mining in personalized distance education.

The study used naive Bayesian classification method to categorize learning styles, and personalized learning content for personalized learning system organizations which provided information for decision-making as shown in Fig. (**3**). (1) **Identification of learners**. In learning who login to the learning system first with a user name reveals learning style database in the weather has the learning who of learning style type records, if has, system is through the learning who of learning style rendering for its learning style of learning content, if the learning style is not determined, Shi system will access the information related to learning behaviour (user name, the time of enrolment in the learning module and the time of completion, test results, selection of learning content and Learning styles, *etc*.) stored in repository. (2) Pre-processing of data. Quality of data is an important factor for successful data mining through data processing. On one hand, it guarantees that the modelling data is correct and effective, while on the other hand, by regulating the format and content of the data, it makes the model more accurate and effective. Through on source data analysis and processing, classification model of learning styles is obtained which includes training sample set of the property, that video learning times (v), and FLASH learning times (f), and text learning times (t), and video learning average results (VS), and FLASH learning average results (FS), and text learning average results (TS), and learning style type (s).

3.2. Association Rules

Associated rules mining in data mining is the most active research method , initially introduced for shopping basket analysis. Its purpose was to find trading database of the different commodities in Zhejiang o, with development in the theory of associated rules mining , Associated rules are widely used in other fields like medical, and commercial, industry, and in personalized learning systems to keep large amount of data based on learning records. This data contain personal information about learners and learning behaviour, however, its potential value has never been fully used. The association rules mining method can be applied for the analysis of learner's record, to find out learners ' preferences according to the Act of law to guide personalized learning system about personalized service, and problems that are worth studying. This study described the basic concepts and methods of association rules and made use of association rules mining to learn about courses and explore the relationship between the test scores, learning styles and learning methods of association rules method for the development of personalized learning system.

Data Association is an important aspect in the database of knowledge. Between two or more variable values there is a regularity, which is called an association. Associations can be divided into simple association associated, timing, and the causal associations. Correlational analysis of hidden Association network aims to identify the database. Sometimes, the functions associated with data in the database are not known. Despite of this uncertainty, analysis rules bring credibility through associations. . Interesting associations or related links were found between sets of items in large volumes of data through the application of mining association rules.

$$Support(\mathrm{x}) = \frac{\{\mathrm{d} \in \mathrm{D} \mid x \subset d\}}{D} \tag{4}$$

The strength of association rules can be measured in terms of its support and confidence. Support is determined by how often certain rules can be used for given data set and confidence is determined by analysing the frequency of the occurrence of y in the x. Measurement of both supports (s) and confidence (c) is defined as follows:

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N} \tag{5}$$

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{6}$$

The basic task of association rule mining involves the specification of first user-specified minimum support and minimum confidence for the mining of association rules in the database. Association rule mining process consists of two stages: at the first stage, all frequent item sets in data collection are found, while the second phase is based on mining association rules of those frequent item sets the user is interested in. (1) Through the given user minimal support to find all frequent item sets meets not less than support a subset of all the items is created. In fact, the frequent item sets may contain. In General, we are concerned only with those who did not were other frequent item set contains a collection called maximum frequent item sets . All previously observed frequent item sets form the basis of association rules. Discovering frequent item sets in the original method determines the structure of each set of options that provide support after calculation in order to complete this task. Each candidate's item sets must be compared with each transaction. If the candidate's set is included in the service, support for the candidate set a count value, in order to reduce the computational complexity of frequent item sets by using the following two methods. (2) From high frequency associated with projects, group's rules are created. Therefore, high frequency of k-groups of projects from the previous step is used to generate rules, having minimal reliability, under the threshold. If the rules of the trust meet minimum reliability, association rules are set . For example, with high-frequency, k-projects are generated by the set {a, b}with rule AB, whose reliability can be obtained by the formula. If the reliability is greater than or equal to the minimum reliability, AB is set for association rules.

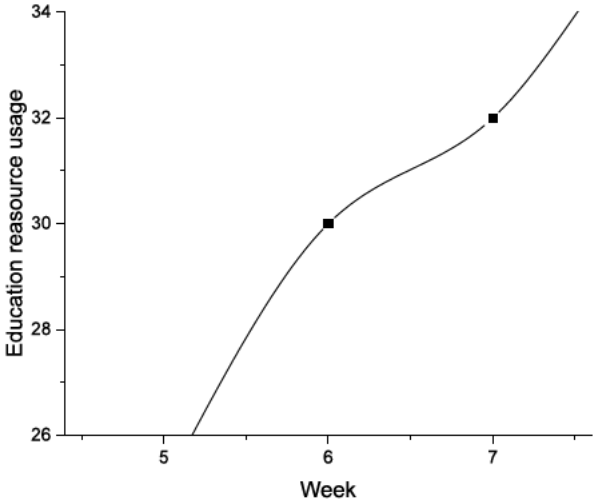

Fig. (4). Education resource usage based on data mining in personalized distance education.

At present , association rules mining technology has been widely used in the financial industry in the West, it can successfully forecast the bank's customer needs. The system uses data mining techniques by utilizing basic information regarding learning activities for analysis, sorting and mining to construct learning features. The fact that one section contains a description about the learner, while another section contains the rules for describing learner, the method of association rules is used to derive data by using the learning rules, assisted by experts in education. Moreover established operand is used to distinguish between good and bad rules to reconstruct the learners ' characteristics. In order to find the description of individual learners' behaviour rules, a variety of data mining algorithms can be used [12, 13].

For generating association rules, the process involves four steps: data processing, transaction of database, generating frequent item sets and association rules. They of function respectively for: on learning who learning behaviour data for cleaning finishing formed learning results database; from learning results database in the extraction data mining object, on data mining object for coding and will relationship table conversion for Affairs database; according to given of minimum support degrees in Affairs database of based Shang generated frequently items set; according to given of minimum confidence degrees by frequently items set generated associated rules.

CONCLUSION

The web data mining technology and distance learning are fully combined, using existing data sources and data mining algorithms to find courses, network design and other links between them. Studying favourable rules, and applying these tothe existing distance education platforms, proved best for the distance education mode , and has enabled students to change from being passive to Active students. Distance education is more personal, comprehensive and has greater advantage.

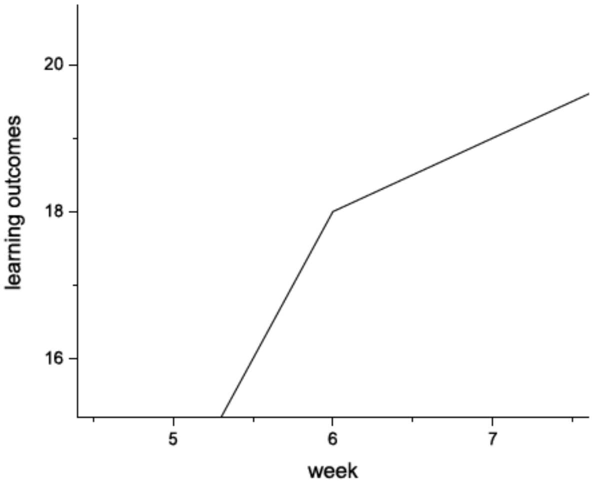

Fig. (5). Learning outcomes based on data mining in personalized distance education.

In this paper, the mining application was used in distance education. With the development in web and internet, human society has entered into the information age. Due to the arrival of this era, education also faces new opportunities and challenges. Therefore, the subject has a strong theoretical and practical application.

Modern distance education has undergone three changes; it has become more dependent on the network, but many distance education websites are not dynamic and personalized. However, the existing distance education system has accumulated a great deal of useful information on the Web, the w g b mining technology can be used from a broad array of existing aggregate data found in useful information. Moreover, the use of mining application in distance education was also studied. This log processes pre-treatment data in five stages; purification, identification of the data, session identification, path complements and recognition, introduction of filter lists, and heuristic rules, on the basis of maximum forward path, which are described one by one. In the end, basic theory and application of mining in remote education process and application of Web mining in remote education along with its all modules of function have been described. This model proved to be advantageous as it is useful for mining of object, and not only for log information. It also analysed learner's personal information, the results, content, and will mining get of results application to personalized recommended system in the, can more conducive to learning who of learning, while, also conducive to improve the structure of website topology and content of update reflects more personalized than the ordinary distance education (Fig. **4**).

This system was used to personalize learning theory for based, will education theory, and data mining technology and network technology and component technology, application to learning system in the, the system full grasp education teaching law, can according to each students of basic situation, and learning style, and learning preference, and learning requirements, itself features, to learning who provides different of learning information and the learning page, reflected has education thought and theory of guide role, also reflected has education theory and modern science and technology of mutual fusion .

The main features of this system are as follows: (1) suitable for multidisciplinary studies, and used as a common platform for individualized learning; (2) it expands the knowledge ; (3) it classifies learners ' learning styles according to the learners ' records, automatically; (4) According to the learner's basic situation, information regarding learning styles is provided to determine learners' characteristics in learning, for personalized functions; (5) It applies data mining algorithm to fully extract potential information, and provide a guarantee for personalized learning content to the learning organization.

In personalized distance education system, in addition to mining log files, users can also interact with the site database tree and get access to courseware, site, files, such as mining, students ' assignments and examination process along with the analysis and results, as well as can ask questions and receive answers , thus, carrying out a full range of personalized services. Despite initiation of developments in personalized remote education website, development also just started, It has many technological problems which need inquiry, but it is worth the people's efforts, because remote education in today's world provides lifelong education. However, this field requires in-depth research to establish perfect personalized remote education platform (Fig. **5**).

We know of no previous register-based study that has illustrated the relevance of these two crucial issues in an equally detailed manner as we have done here.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Key Project of Guangxi Social Sciences, China (No. gxsk201424), the Education Science fund of the Education Department of Guangxi, China (No. 2014JGA268), and Guangxi Office for Education Sciences Planning, China (No. 2013C108).

REFERENCES

[1] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Computers and Education,* vol. 51, no. 1, pp. 368-384, 2008.

[2] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," *Evolution of teaching and learning paradigms in intelligent environment*, Springer Berlin Heidelberg, vol. 62, pp. 183-221, 2007.

[3] S. H. Ha, M. B. Sung, and C. P. Sang, "Web mining for distance education, "Management of Innovation and Technology," In: *Pro-*

ceedings of the IEEE International Conference on, vol. 2, pp. 715-719, 2000.

[4] O. R. Zaïane, and J. Luo, "Web usage mining for a better web-based learning environment," In: *Proceedings of conference on advanced technology for education*, 2001.

[5] C. Romero, "Personalized links recommendation based on data mining in adaptive educational hypermedia systems," *Creating New Learning Experiences on a Global Scale*, Springer Berlin Heidelberg, pp. 292-306, 2007.

[6] W. Hämäläinen, "Data mining in personalizing distance education courses," In: *World conference on open learning and distance education*, 2004.

[7] M. Penelope, "Using semantic web mining technologies for personalized e-learning experiences," In:*Proceedings of the web-based education*, pp. 461-826, 2005.

[8] C. F. Lin, "Data mining for providing a personalized learning path in creativity: An application of decision trees," *Computers and Education*, vol. 68, pp. 199-210, 2013.

[9] E. Yukselturk, O. Serhat, and Y. K. Türel, "Predicting dropout student: An application of data mining methods in an online education program," *European Journal of Open, Distance and e-Learning*, vol. 17, no. 1, pp. 118-133, 2014.

[10] M. Abdous, H. Wu, and C.J. Yen, "Using data mining for predicting relationships between online question theme and final grade," *Educational Technology & Society*, vol. 15, no. 3, pp. 77-88, 2012.

[11] A. Fernández, "E-learning and educational data mining in cloud computing: an overview," *International Journal of Learning Technology*, vol. 9, no. 1, pp. 25-52, 2014.

[12] J. A. Lara, "A system for knowledge discovery in e-learning environments within the European Higher Education Area–Application to student data from Open University of Madrid, UDIMA," *Computers & Education*, vol. 72, pp. 23-36, 2014.

[13] M. Munk, and D. Martin, "Impact of Different pre-processing tasks on effective identification of users' behavioral patterns in web-based educational system," *Procedia Computer Science*, vol. 4, pp. 1640-1649, 2011.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015