Knowledge Mining in Higher Education

Nittaya Kerdprasop, Ekkachai Naenudorn, Jatsada Singthongchai, Wilairat Yathongchai, Chusak Yathongchai, and Kittisak Kerdprasop

***Abstract*—**The term *higher education* refers to learning stages conducting at colleges, universities, and institutes of technology. Learning at this stage includes undergraduate, postgraduate, vocational education, and training. In this paper we discuss the roles of a novel computational technology known as *knowledge mining*, or data mining, that can facilitate the improve in learning competency and the support for administration in many aspects. Automatic knowledge acquisition and discovery from learners' profiles and institutional repositories are expected to be major sources of intelligence not only to support a better learning object design, but also to help administrative decision making. We review the knowledge mining technology, its potential applications to higher education, and then present case studies of knowledge mining to support student recruiting and student retention planning.

***Keywords*—** Knowledge mining, Knowledge discovery, Data mining, Higher education, Intelligent learning system.

I. Introduction

During the last few years, we have witnessed an important shift in educational systems. The adoption of automatic learning ability from the field of knowledge mining and the availability of the World Wide Web have resulted in state of the art Web-based intelligent learning environments. The term environment in this context refers to a suite of computer programs that facilitates electronic learning (e-learning) and distance education. These facilities include course creation and delivery, learning tracking and assessment, enrollment, and administration.

This kind of system has many other names such as Learning Management System (LMS), Course Management System (CMS), Learning Content Management System (LCMS), and Virtual Learning Environment (VLE). The well-known systems such as Moodle (http://moodle.org), ATutor (http://www.atutor.ca), and WebCT and Blackboard (http://www.blackboard.com) have been adopted by many institutions and organizations worldwide to deliver online courses.

Despite the success and popularity of Web-based online courses, people soon realized that these online materials are nothing more than a group of static hypertext pages [26], [42] designed once and used by any learner regardless of their diversities in capabilities, needs, and perceptions. Since then, the issues of content adaptability and intelligent curriculum sequencing have become the major research goals for the development of advanced Web-based educational systems [21], [34], [45]. The intelligent learning environments have also been introduced [4], [18], [22], [27], [29], [35] to be the integrative approach of intelligent technology and Web-based educational systems.

Automatic knowledge acquisition and discovery from learners' data repositories and knowledge bases are major sources of intelligence facilitating the creation of adaptive functionality of the most current intelligent learning environments [2], [11], [24], [38]. It is thus the main objective of this paper to discuss advances and trends in intelligent learning environments with the emphasis on the many roles of knowledge-mining technology contribute to the development of intelligent infrastructures.

The rest of this paper is organized as follows. In section 2, we briefly review essential concepts of knowledge mining. The potential applications of knowledge mining to educational data have been demonstrated trough simple example in section 3. Discussion of intelligent techniques applied to education is in section 4. The proposed framework of integrating knowledge mining as intelligent modules in educational systems is in section 5. We then present case studies, in section 6, applying knowledge mining concepts to the real-world problems in education, that is, student recruitment and retention. Section 7 concludes the paper with discussion regarding possible future research directions.

II. Preliminaries on Knowledge Mining Technology

A. Basic Concepts

Knowledge mining [19], [31], [33] is the discovery of hidden knowledge stored in various forms and placed in large data repositories. Hidden knowledge refers to models and patterns that implicitly exist in the data set and are unknown a priori. For instance, consider the set of data instances {(0, 3), (1, 6), (2, 15), (3, 30)}. The explicit knowledge is that this

Manuscript received March 10, 2012: Revised version received June 12, 2012. This work was supported by grants from the National Research Council of Thailand (NRCT) and Suranaree University of Technology through the funding of Data Engineering Research Unit.

N. Kerdprasop is an associate professor and the director of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology, 111 University Avenue, Muang District, Nakhon Ratchasima 30000, Thailand (phone: +66-44-224-432; fax: +66-44-224-602; e-mail: nittaya@sut.ac.th).

E. Naenudorn, J. Singthongchai, W. Yathongchai, and C. Yathongchai are doctoral students with the School of Information Technology, Suranaree University of Technology, Muang District, Nakhon Ratchasima 30000, Thailand (e-mail: ekkachai.n@acc.msu.ac.th, jatsada_007@hotmail.com, y_wilairat@hotmail.com, y_chusak@yahoo.com).

K. Kerdprasop is with the School of Computer Engineering and Data Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima, Thailand (e-mail: KittisakThailand@gmail.com).

data set contains four data, represented as a (x, y)-pair. The implicit knowledge that are hidden in the data set is a pattern $y = 3x^2+3$, and a model $y = ax^2+b$. A pattern is an expression describing a subset of the data, whereas a model is a representation of the source generating the data. In the knowledge-mining context, we refer to both patterns and models as new knowledge automatically discovered from data sources. This emerging technology was originally coined [14] as knowledge discovery in databases or KDD, also known as data mining in the field of statistics. In this paper, we use the term "knowledge mining" to state the fact that the process discovers hidden knowledge from not only raw data, but also from previously discovered patterns and meta-data. Knowledge mining, thus, can be characterized by the following scheme: *(stored data and meta-data) + previously discovered knowledge → new knowledge*.

The KDD technology often reveals interesting patterns and dependency relationships hidden in large data. It is thus beneficial to educators to understand how this emerging technology is applicable to the development of intelligent learning environments, as well as the insight that such knowledge-mining tasks provide. Researchers and educational software developers should also gain benefits from the awareness of the emergence of this intelligent-related technology, as it will soon be a major part of computer-assisted learning tools.

B. Process of Knowledge Mining

The whole process of knowledge mining works around data, meta-data, and previously discovered patterns. It can be conceptually shown as in Figure 1. The initial step (problem defining) of knowledge mining focuses on setting the goal or specifying the problem, which can be achieved through understanding the task objectives and organization requirements. Defining the problem is important because it will guide activities in subsequent steps to collect only relevant data, perform mining with the appropriate algorithm, and keep only pertinent and actionable knowledge.

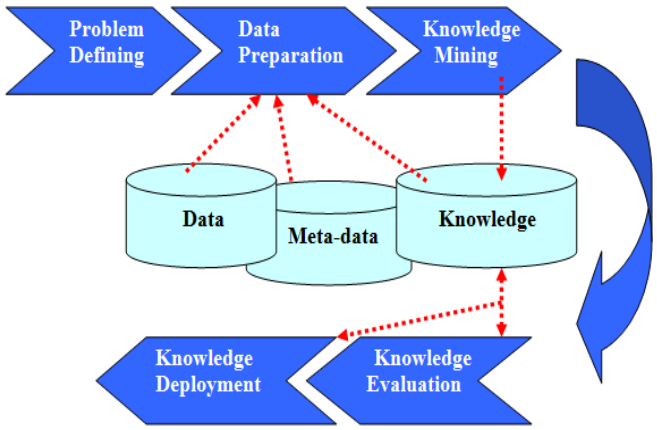


Fig. 1 The process of knowledge mining

A clear problem statement should define what is to be accomplished and what the desired outcome is likely to be. For instance, if the objective of mining is to reduce the number of dropouts at junior college level, the specific problem statements might be, "Is there an association between financial support and students' decision to leave school?" or, "Are there interdependencies among different groups of dropouts?"

The second step (data preparation) covers all activities necessary for preparing high quality data suitable for mining. Data preparation step includes collecting data from multiple sources, transforming the data format, and selecting data representatives with minimum but sufficient attributes. Data preparation is typically time consuming and likely to be performed iteratively. Meta-data and background knowledge are kinds of supportive information that can be applied in this step. Supporting tools that are useful for this step can be as simple as a notepad program, or more sophisticated tools such as those provided by SPSS or SAS data exploration packages.

The third step is knowledge mining, which is the search for and extraction of interesting patterns (local generalized structures) or models (global generalized structures) from data. Such patterns and models are called knowledge. This step is the backbone of the knowledge-mining process. Several techniques are available, but their application needs some adjustment to obtain optimal results.

The fourth step (knowledge evaluation) is to assess the accuracy and interestingness of the discovered knowledge over some threshold values. Accuracy is correctness of the induced model and it can be evaluated by using another set of data called test data. The required level of model accuracy has to be set by users or miners. Interestingness of the induced model is somehow a more subtle issue than the accuracy metric. Evaluating interestingness depends considerably on the judgment of miners or domain experts. Accurate and interesting knowledge is finally fed to the deployment step to become actionable information for an organization or to act as background knowledge for other knowledge-mining tasks.

III. Mining Educational Data: A Demonstration

Knowledge mining can be conducted with different kinds of algorithms. Despite hundreds of available algorithms, they may be grouped roughly according to the task that specific algorithm performs into three categories: *classification*, *association*, and *cluster analysis*. For the purpose of demonstration, each knowledge-mining task will be performed on the dropout data at the secondary level (7th grade through 12th grade) in the school year 2003–2004 [40].

A. Dataset Detail

The dropout data was reported annually by the state education agencies throughout the United States including the District of Columbia, Puerto Rico, American Samoa, Guam, the Commonwealth of the Northern Mariana Islands, and the U.S. Virgin Islands. The data files have been collected and

made publicly available by the National Center for Education Statistics (NCES, http://nces.ed.gov), U.S. Department of Education. Data file for school year 2003–2004 includes only dropout data for school districts enrolling 1,000 or more students. A dropout is an individual who enrolled in school at some time during the previous school year but does not enroll at the beginning of the current school year. The school year is the 12-month period starting at October 1.

NCES does provide a caution that the data file should not be used to compute state-level estimation of public school dropouts because the reported data are restricted to state education agencies with membership of 1,000 or more. However, we can investigate the dropout patterns and dependencies among different dropout groups with the knowledge-mining technology. For simplicity, we will use only dropout data reported by school districts from the California and Florida states.

Each data record (also called data instance in the data-mining community) is a report from each school district and the record contains 15 variables (or attributes). The first three attributes are Locale (i.e., location of the school divided into eight categories: 1 = large city, 2 = midsize city, 3 = urban fringe of a large city, 4 = urban fringe of a midsize city, 5 = large town, 6 = small town, 7 = rural, outside metropolitan, and 8 = rural, within a metropolitan), LO-offered (i.e., the lowest grade of the school), and HI-offered (i.e., the highest grade offered by the school). Attributes 4–9 are the total number of student enrollments in grade 7 through 12.The nonzero number was grouped into four intervals: 1–999, 1,000–4,999, 5,000–9,999, 10,000–infinity and encoded in the dataset as [1–1K), [1K–5K), [5K–10K), [10K–UP), respectively. Attributes 10–15 are the total number of dropouts at grade 7 through 12.The nonzero number was grouped into five intervals: 1–50, 51–99, 100–500, 501–999, 1,000–Infinity and accordingly encoded in the dataset as [1–50], [51–99], [100–500], [501–999], and [1,000–UP]. The number "–2" may appear in the dataset and it represents the "non-applicable" case, number "–1" means "data was not reported by the agency" and "0" encodes "no occurrence of the data element." The first three data instances of California school districts are shown as follows:

8, KG, 12, [1–1K), [1–1K), [1–1K), [1–1K), [1–1K), [1–1K), 0, 0, 0, 0, 0, 0

3, KG, 8, [1–1K), [1–1K), –2, –2, –2, –2, 0, 0, –2, –2, –2, –2

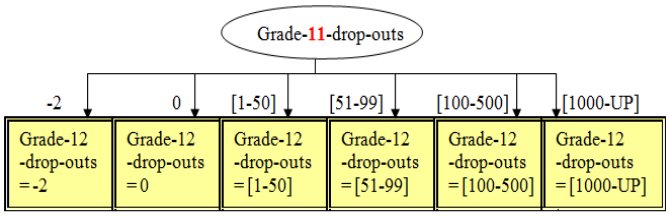
7, KG, 12, [1–1K), [1–1K), [1–1K), [1–1K), [1–1K), [1–1K), [1–50], 0, 0, 0, [1–50], [1–50]

In this knowledge-mining demonstration, we adopt the Weka (Waikato environment for knowledge analysis) system (http//www.cs.waikato.ac.nz/ml/weka/), which is open source software (i.e., free software that provides source code for anyone to modify it) used for knowledge mining and data analysis.

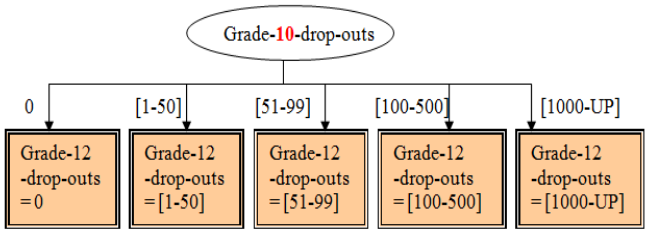
B. Mining for Classification Trees

The *classification* task is to find major characteristics that can classify correctly the specified target attribute. The algorithm normally applied for this task is J48 [44], which is known as a decision-tree induction algorithm because it can find (or induce) a pattern from the given data and display the pattern graphically as tree. The leaf nodes at the bottom of the tree are the predicted decision, whereas the nodes at the upper levels are conditions to be fulfilled prior to making the decision.

Our objective is to draw some patterns from the dropout data of 565 public school districts in the California state compared with the patterns induced from the dropout data of 70 public school districts in the Florida state. The results of mining the 12th grade dropout patterns in the California state as compare to the Florida state are shown in Figure 2. The two patterns are slightly different. In the California pattern, the total number of 12th grade dropouts can be predicted from the total number of 11th grade dropouts. This pattern is 77.2% accurate measured with the 10-fold cross-validation method. But in the Florida pattern, the total number of 12th grade dropouts can be predicted from the total number of 10th grade dropouts (accuracy = 82.8% with the same measurement method as in the California mining). These results reveal the nature of knowledge-mining technology that it is data dependent. If the data change, the mined knowledge may change.



(a) California state



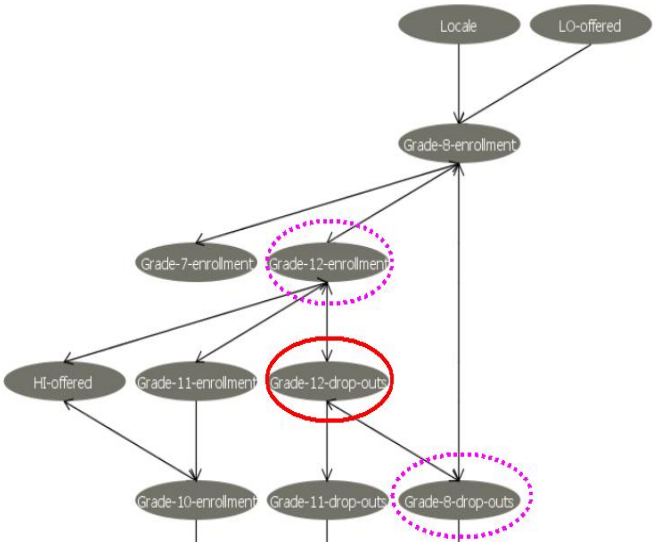
(b) Florida state

Fig. 2 Decision trees for classifying and predicting the attribute *Grade-12-drop-outs*

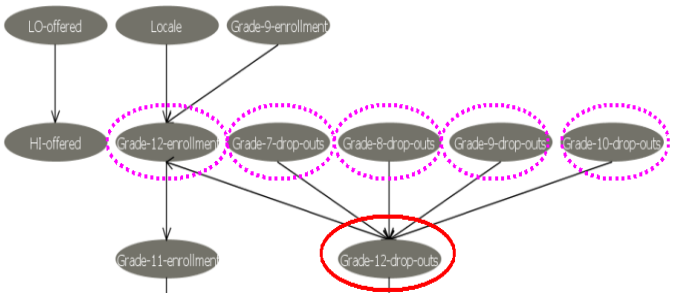
Besides decision-tree induction, the classification task can be performed using various kinds of machine-learning techniques such as artificial neural network, genetic algorithm, fuzzy rule induction, rough set theory, support vector machines, case-based reasoning, and many others. To derive cause-and-effect relationships or causal knowledge with some level of uncertainty, *Bayesian belief network* or *Bayesian*

network is utilized because it can convey both qualitative and quantitative information.

Bayesian network [15], [23] is a directed acyclic graph with nodes to represent variables or attributes, arcs to represent probabilistic correlations or dependencies between attributes, and conditional probability tables associated with each node. The table associated with the root node, which is an independent variable, contains unconditional probabilities. If there is a directed arc from node *X* to node *Y*, then *X* is a parent of *Y*, and *Y* is called a descendant of node *X*. Given parent node(s), a variable is said to be conditionally independent of its non-descendants in the network. Decision tree (as shown in Figure 2) can be used to predict the future outcome given the learned pattern from the past data. But if we are interested in finding explanation about dependency relationships among data attributes or to learn cause-and-effect relationships, Bayesian network should be employed.



(a) California state



(b) Florida state

Fig. 3 Bayesian networks showing that (a) in California state, the numbers of *Grade-12-enrollment* and *Grade-8-drop-outs* are two direct causes of the occurrence of value in the attribute *Grade-12-drop-outs*, and (b) in Florida state, numbers of *Grade-12-enrollment*, *Grade-7-drop-outs*, *Grade-8-drop-outs*, *Grade-9-drop-outs*, and *Grade-10-drop-outs* are all direct causes of the occurrence of value in the attribute *Grade-12-drop-outs*

The Bayesian network-learning result of California 12th grade dropouts versus the result obtained from the Florida state data is shown in Figure 3 (conditional probability tables are omitted due to space limitation). To learn the cause-and-effect relationships using the Weka system, we have to set the search algorithm to ICS search; otherwise, with different search algorithms the learned network simply represents conditional dependencies among data attributes..

The Bayesian network in Figure 3(a) tells us that the total number of 12th grade enrollment and the total number of 8th grade dropouts are two direct causes of the number of 12th grade dropouts. Figure 3(b) depicts the Bayesian network learned from the Florida school districts. It can be noticed that in addition to the total number of 12th grade enrollment and the total number of 8th grade dropouts, the other three attributes (i.e., the total numbers of 7th, 9th, and 10th grade dropouts) also affect the number of 12th grade dropouts.

C. Mining for Association Rules

With the same set of data, we can perform *association analysis* to discover correlations or relationships that hold among data attributes. The results of association analysis are reported as the "IF-THEN" rule. The "IF" part states the antecedent or condition of the proposition that must be true. The "THEN" part is the conclusion. In Weka, the "IF-THEN" rule is displayed as *Antecedent ==> Conclusion*.

Association rules do not convey the cause-and-effect relationships; they simply state that if the event(s) in antecedent part does occur, the event(s) in the conclusion part must also occur. Therefore, they represent the co-occurrence relationships. The following rules show the best five results (in terms of confidence value or accuracy of the rule) of association analysis conducted with the Apriori algorithm [1] on California and Florida school districts data, respectively.

California school district

1. *Grade-7-enrollment = [1–1K) ==> LO-offered = KG conf:(0.99)*
2. *Grade-7-enrollment = [1–1K) & Grade-8-enrollment = [1–1K) ==> LO-offered = KG conf:(0.99)*
3. *Grade-8-enrollment = [1–1K) ==> LO-offered = KG conf:(0.99)*
4. *Grade-7-enrollment = [1–1K) ==> Grade-8-enrollment = [1–1K) conf:(0.99)*
5. *LO-offered = KG & Grade-8-enrollment = [1–1K) ==> Grade-7-enrollment = [1–1K) conf:(0.99)*

Florida school district

1. *HI-offered = 12 ==> LO-offered = PK conf:(1)*
2. *LO-offered = PK ==> HI-offered = 12 conf:(1)*
3. *Grade-8-dropouts = [1–50] ==> LO-offered = PK conf:(1)*
4. *Grade-8-dropouts = [1–50] ==> HI-offered = 12 conf:(1)*
5. *HI-offered = 12 & Grade-8-dropouts=[1–50] ==> LO-offered = PK conf:(1)*

Each association rule is annotated with confidence value (conf) to give information regarding how accurate the rule is. Given the association rule *X* ==> *Y*, the confidence value of this rule is the proportion of number of data instances that contain both *X* and *Y* events to the number of data instances that contain *X* (the occurrence or absence of *Y* does not affect the count of event *X*).

The value 1 is the most accurate association, whereas a value less than 1 implies that the rule might contain some error. Taking the last association rule in Florida school district as an example, the rule states the relationship that "if the highest grade offered by the school is 12 and the number of 8th grade dropouts is in the range [1–50], then the lowest grade offered by this school is pre-kindergarten."

D. Data Cluster Analysis

Another commonly performed mining task is *cluster analysis*. Several algorithms exist that perform cluster analysis. However, the most fundamental and widely used algorithm is the k-means algorithm [28]. The parameter *k* is the number of clusters that users have to specify.

The result of running k-means is the *cluster centroids* (or center points) that report the characteristics of representatives in each cluster. If we set *k* to be 5 and run k-means on our data, we will obtain 5 centroids reporting characteristics of each data subgroup.

For instance, the first cluster of Florida school districts might be: a group of public schools located in rural area not within a metropolitan; the lowest grade mostly offered by these schools are pre-kindergarten; the highest grade mostly offered are grade 12; numbers of students enrolled in 7th, 8th, 9th, 10th, 11th, 12th grades are in the same range, that is 1–999; numbers of 7th through 12th grade student dropouts are also in the same range of 1–50.

Other groups of students in different school districts can be interpreted in the same manner. If we change the number of clusters (i.e., vary *k* to different values), characteristics of each data cluster should also change.

IV. Knowledge Mining Roles in Intelligent Learning Systems

Most intelligent learning systems are composed of five main components [17], [34]: student modeling, content and domain expert modeling, structure and communication module, pedagogical strategy module, and supporting module. These components cover activities and people involved in intelligent learning environment. Currently, knowledge-mining technology has been adopted to introduce intelligent behavior associated with various parts of such environments. In this section, we discuss the available technology supports and recently proposed approaches to mine knowledge; the discussion is based on these five components.

A. Student Modeling

Web-based education is a form of computer-supported learning in which course contents are delivered via the Internet. In such an educational setting, learning activities can take place pervasively through the support of Web servers. The server mediates courseware transmission and records Web access in log files. The Web logs normally contain learners' navigation on Web sites including accessing time, Web traversal paths, and responses. These Web logs are major data sources widely used to perform mining in order to gain some knowledge about learners' behavior. Zaiane [46] is among the first data-mining researchers interested in acquiring knowledge from Web logs. The early work of Zaiane focuses on the data preparation step of Web usage mining. Web mining is a powerful technology to obtain knowledge about how students learn the course materials, in which order the students study subtopics, which topics have been skipped, how much time the students spend on each topic, and how a pedagogical strategy affects students with different learning paces and styles.

Chen, Hsieh, and Hsu [9] were also interested in gaining knowledge about students' understanding of course contents. This research team adopted association analysis to mine the learners' profiles for capturing deviation from common learning patterns. The deviations that reflected misconceptions of students during the learning process could be inferred from incorrect testing items. Chen et al. also proposed a remedy approach for the students to correct their comprehension on the subject matters by delivering course materials with easier difficulty level but covey the similar learning concept. Ceddia, Sheard, and Tibbey [6] developed a tool named WAT (Web log analysis tool) to analyze sequences of Web site interactions of students to determine whether they understand the core concept of a specific topic.

Lee, Chen, Chrysostomou, and Liu [25] mined students' behavior by means of decision-tree induction with the main goal of discovering cognitive style of learners in order to manage flow of course contents suitable for each student. Decision tree is a data structure used to classify the cognitive style (field dependent/independent) of students during their learning sessions. Web navigational behavior of the students was observed and transformed into training data to induce a decision tree.

Romero, Gonzalez, Ventura, del Jesus, and Herrera [37] proposed a genetic algorithm approach to mine usage data of the Moodle course management system. Moodle is a free and open source licensed software designed to help educators create online courses and assessments. The Moodle system normally accumulates a great amount of information that is valuable for the mining and analyzing of students' behavior. The analysis results were in the form of descriptive rules such as "*IF course = CS101 & number_assignments = High & number_posts = High THEN mark = Good*." Such a rule is a student model describing relationships between student activities on modules provided by the e-course and the final mark obtained in that course.

B. Content and Domain Expert Modeling

An educational system that is meant to be intelligent has to provide at least two functionalities: content adaptiveness and personalization. AHA! [12] is an adaptive educational system working with hypermedia data in the form of XML files. This system has a log file designed to record, in addition to normal timestamps and user's identification, a flag field marking access to the concept fragment of each user. Concepts representing domain knowledge fragments are defined as knowledge attributes whose value is updated when users read related Web pages. By analyzing the knowledge attribute values, the user's level of knowledge on the studied topics may be assessed. Kristofic and Bielikova [20] proposed to improve the adaptation technique by means of knowledge discovery. The techniques proposed are association analysis, sequential patterns, and traversal pattern discovery. Association rule mining was employed to learn relations between concepts. The results were used in the process of recommending relevant concepts. The authors also extended the association algorithm to discover the sequential patterns that students navigated the concepts. Their final outcome was a recommendation system to suggest learning concepts relevant to each user.

Wang, Tseng, and Liao [43] proposed an adaptive system designed particularly for teaching English as a second language course. Given students' profiles, the system adopts a decision-tree induction algorithm to discover the most adaptive learning sequences suitable for each group of students for particular teaching content. The students' profiles have been created from the pretesting, posttesting, and student models that contain five attributes: gender, personality (e.g., introverted, neutral, extroverted), cognitive style (i.e., field dependent, field independent), learning style (i.e., sensing thinking/feeling, intuition feeling/thinking), and students' grades from the previous semester (i.e., low, medium, high). The final output of the system is a recommendation such as, "*For a female student with neutral personality, mildly introverted and has an intuition-feeling learning style, Suggestion is she should be assigned a learning sequence as <Main idea, Details, Vocabulary, Inference, Critical reading>*."

Biletskiy, Baghi, Keleberda, and Fleming [5] proposed an adjustable personalization approach to deliver learning objects to learners. The authors argued that current learning objects have been created by various suppliers targeting several groups of learners such as students, employees, and professionals. To serve the specific needs of each learner, some kind of personalization has to be adopted. The authors propose to compare the learner profile and the learning object descriptions and deliver objects most relevant to each learner. A comparison metric such as similarity measurement can be used in this approach. Feedback about content usefulness and suitability from the learner is also obtained to adjust the learner's profile on the issue of preference.

C. Structure and Communication Support

Current Web-based educational systems provide services that are transmitted synchronously (such as video conferencing) and asynchronously (such as discussion Web boards and e-mail). Online students can register in many courses and seek course materials from several lecturers that may be located at different sites. Such a physically distributed environment needs advanced technology in order to handle the dynamic aspects of sites and information resources.

Handi [16] proposed a MASACAD system designed with a multi-agent approach to customize information presented to students. An advantage to adopting the multi-agent paradigm is that the changing environment of a course, for example, may be accommodated, as in students leaving a course or registering for a new course. The students' profiles may be updated accordingly without intervention from human administrator.

Chen [7] proposed an intelligent learning system with personalized learning path guidance. The system collects the incorrect testing responses of each learner in a pretesting session, and then applies the genetic algorithm to generate appropriate learning paths. Personalized curriculum sequencing has been conducted through consideration of courseware difficulty level and the continuity of learning concept. The system is meant to replace the freely browsing learning mode in order to increase learning performance by reducing disorientation during the learning process.

D. Pedagogical Strategy Support

In a Web-based educational setting, the Web provides not only information and content accessible through browsing, but also interactive pedagogical models through Web boards, blogs, and other communication methods. Recently, Martin, Alvarez, Fernandez-Castro, and Urretavizcaya [32] presented the SIgMa (suggestions for improving educational aspects in Magadi) system to be an adaptable feedback generation tool for teachers. This tool analyzes students' performances using statistical calculations and data-mining techniques to discover anomalous behavioral patterns that reveal situations difficult for the students. The system provides feedback by means of a rule-based system to make suggestions for learning improvement. The suggestions are also adapted to teaching strategies and preferences.

Shih, Chiang, Lai, and Hu [41] also applied data-mining techniques, mainly decision-tree induction, in their Web-based self-assessment system. The system has been used to study disturbances of high-risk freshmen students, first-year students who have trouble following class contents and thus highly probable to fail in their study, providing counselors with the proper information to better bolster their counseling services.

E. Infrastructure Support

The supporting module provides technology in terms of hardware and software to be an infrastructure for the learning environment. Marshall, Chen, Shen, and Fox [30] developed the GetSmart system in integrating course management, digital library, and concept mapping components to support the information search process. The authors view concept mapping as a knowledge visualization tool that can provide both a course map view of learning topics and a methodology to support personal knowledge acquisition. The authors report that from their field study, they observed improvement in the scores of the students' online quizzes after including concept mapping in the curriculum. Chen, Kinshuk, and Chen [10] also based their study on concept maps. They proposed a construction of domain concept maps from academic articles by applying text-mining techniques. The domain concept map is a graphical representation of knowledge structure in which nodes represent concepts and links represent relationships between concepts. They concluded, in their study, that the concept maps could show the whole picture and core knowledge about a subject domain. This approach can support the domain experts upon constructing concept maps of the subjects.

Romero, Ventura, and Garcia [39] studied Moodle as an infrastructure of the intelligent educational system. They proposed a case study to apply data mining to education that would investigate the following aspects: assessing students' learning performance, providing adaptive courses and recommendations based on learning behavior, evaluating learning objects, providing feedback to both teachers and students, and detecting atypical students. Chen and Chen [8] presented a mobile formative assessment tool using hybrid data-mining techniques, that is, correlation analysis, fuzzy clustering analysis, k-means clustering, fuzzy association rule mining, and fuzzy inference. The objective of this tool was to identify key formative assessment rules according to the learners' portfolios.

V. Intelligent Learning Environment

During this decade, we have witnessed the development of tools customizing knowledge discovery techniques to support intelligent educational systems. Some tools are embedded in the course management system while some operate stand-alone as knowledge acquisition and representation applications.

In this section, the new design of the integrated Web-based intelligent learning environment has been proposed to achieve a full-scale integration of knowledge intensive tasks. The core of this environment is the data and knowledge repository in which the knowledge objects are moved around the four main stages: knowledge object generation, knowledge acquisition and extraction, knowledge object indexing and mapping, and knowledge application. The process of knowledge mining has been applied to acquire knowledge objects that will be subsequently processed in the indexing and mapping stage. This stage supports the search for suitable contents to present to learners. Performance and learners' preference are then captured and stored to be used later in the knowledge-mining stage.

In a framework of the Web-based intelligent learning environment (Figure 4), a repository is defined as a collection of three distinct levels of resources: data, information, and knowledge. Data is the most primitive resource storing raw representations of facts, concepts, learning objects, and other instructional materials. These basic resources are stored in the formats suitable for communication and processing by related modules in the environment. Information is a supplement to raw data, such as meta-data, to describe the meaning, relevance, and purpose of stored data. Information is also intended to be used in knowledge generation, sharing, and discovery guidance. In other words, information refers to any heuristics applied to the process of knowledge mining and management. Knowledge is the most sophisticated entity stored in a repository as knowledge objects, which represent the relationships among data, correlation, and high level of data abstraction. Relationships can take many forms such as rules, vectors, or even mathematical formulas. Knowledge is thus data with semantics.

Data, information, and knowledge stored in a repository are key components of the designed framework. The three major modules of the proposed system communicate through the repository. Some components such as the knowledge-mining engine (or a mining software) are even data-driven; the knowledge-mining engine is data dependent and the mining results may be different if the data contents have been changed. The three main modules in the proposed framework are learning management, content management, and knowledge management. This framework is proposed to support Web-based learning with several learning schemes including adaptive, autonomous, and collaborative learning.

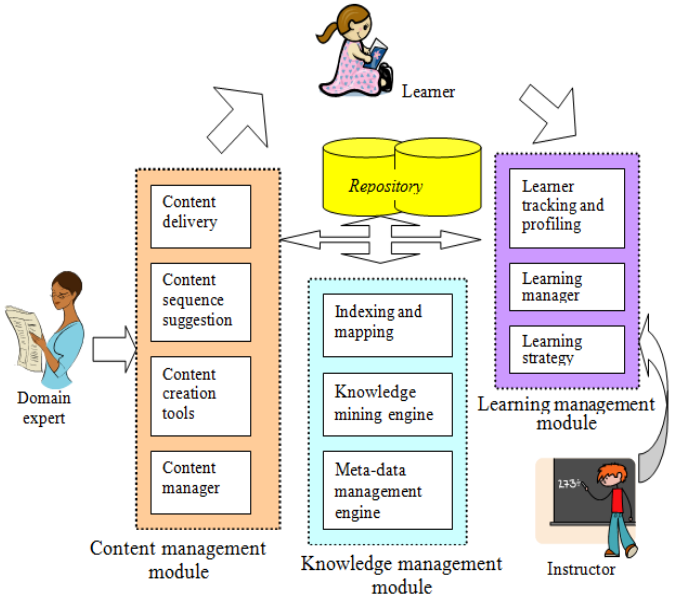


Fig. 4 A framework of infrastructures in an intelligent learning environment

VI. Case Studies on Student Recruitment and Retention

A. Knowledge Mining for Student Recruitment

Currently, various educational institutions, specially small-medium educational institutions, are facing competition with a market share of higher education to meet the needs of increasing number of student enrollment as much as possible. Factors that students take into account before deciding to enroll can be popularity of institution, quality of education, safety factors, and so on. In this section, we present the steps (shown in Figure 5) we had taken for analyzing factors to accurately predicting the enrollment decision of the high school students.

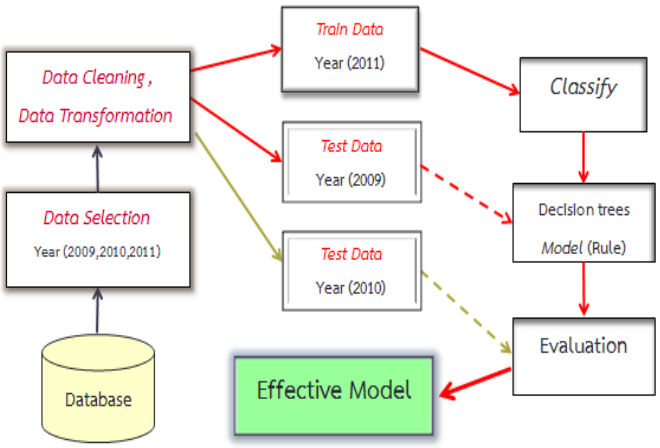


Fig. 5 Study framework of student recruitment

The original collected dataset contains 3,756 records with 27 attributes. After the data cleaning step, we obtain 2,148 records with 6 attributes, that is, province of graduation, department of graduation, field of study, occupation and income of their parents, and student decision on admission to the university. We then apply this data to the four data mining algorithms: J48, ID3, Naïve Bayes, and OneR. The accuracy performances of the four models are comparatively shown in Figure 6. The tree model is also shown in Figure 7.

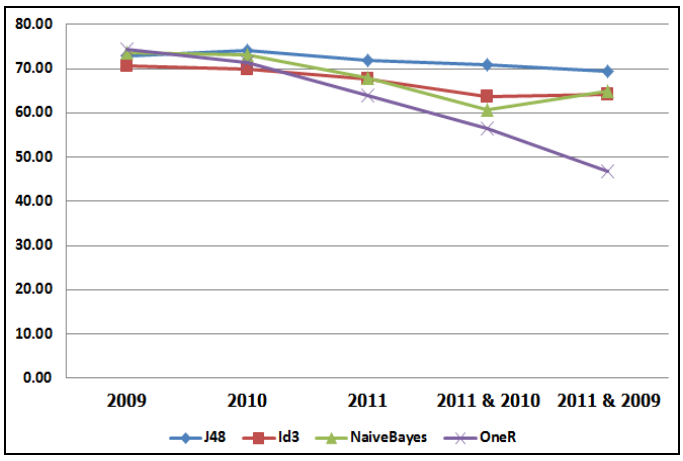


Fig. 6 Accuracy performances of mining algorithms

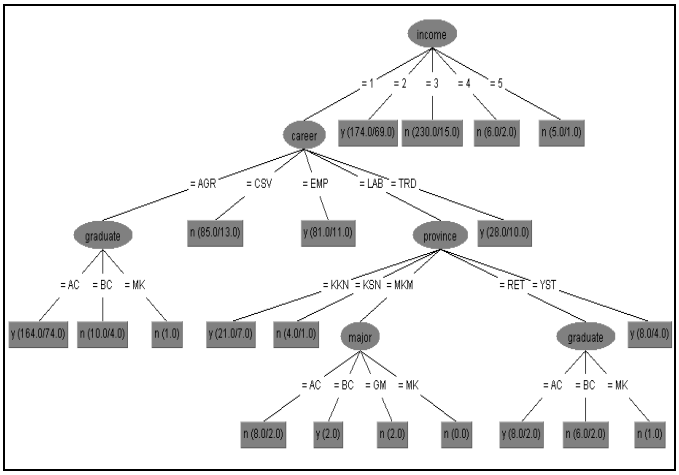


Fig. 7 A tree model summarizing student recruitment characteristics

The best classification model (J48) can be transformed into the decision rule format as follows:

(1) IF income_parent='less than 200,000 baht'

THEN admission=Yes

(2) IF income_parent='less than 100,000 baht' and career_parent='Agriculturist' and graduate='Accounting'

THEN admission=Yes

(3) IF income_parent='less than 100,000 baht' and career_parent ='Company Employee'

THEN admission=Yes

(4) IF income_parent='less than 100,000 baht' and career_parent ='Merchants'

THEN admission=Yes

(5) IF income_parent='less than 100,000 baht' and career_parent ='Laborer' and province='Khon Kaen'

THEN admission=Yes

(6) IF income_parent='less than 100,000 baht' and career_parent ='Laborer' and province='Roi Et' and graduate='Accounting'

THEN admission=Yes

(7) IF income_parent='less than 100,000 baht' and career_parent ='Laborer' and province='Yasothon'

THEN admission=Yes

(8) IF income_parent='less than 100,000 baht' and career_parent ='Laborer' and province='Mahasarakham' and major='Business Computer'

THEN admission=Yes

B. *Knowledge Mining for Student Retention*

The objective of our second case study is to use the educational information from knowledge base in an effective way by applying data mining techniques to analyze the major factors that affect the student drop out in the institutions of higher education.

The data set used in this study was obtained from the database of Academic MIS at Buriram Rajabhat University (BRU) in Thailand during the years 2008 and 2009. There are 731 students enrolled in bachelor degree: 481 students are continuing their study, whereas 251 students had dropped out. Sample data were from faculty of science which has the highest students drop out rate in this university. In this step, data stored in different tables was joined in a single table. After joining process, errors were removed. These steps are summarized in Figure 8.

The analysis points of our study are graphically displayed in Figure 9, and the mining results from different algorithms (compared to each analysis point) are in Figure 10. Some interesting rules obtained from the tree model are as follows:

- The student who has a student loan will not drop out while who has not will drop out if GPAX from high school less than 2.42.
- The student who has not a student loan and GPAX from high school more than 2.42 and studied program in high school was Science-Mathematics will not drop out while who studied other program will drop out.
- The student who drops out after finish first term would have first term GPA less than 1.6 and has not a student loan.
- The student who has first term GPA less than 1.6 and has a student loan will drop out after finish second term.
- The student who studied in Sports science (major ID=240) will not drop out.
- The student who studied in Community health or Computer science (major ID=265 or 230) will have high drops out rate during first term as a sequent.
- The student who studied in Information technology or Computer Technology (major ID=284 or 286) and has first term GPA more than 2.5 but second and third term GPA less than 1.6 will drop out after finish fourth term.
- Most of students who drop out during first term because they want to change major and will re-entrance in the next year while some of them have a finance problem, relocated, change university and have no reason.

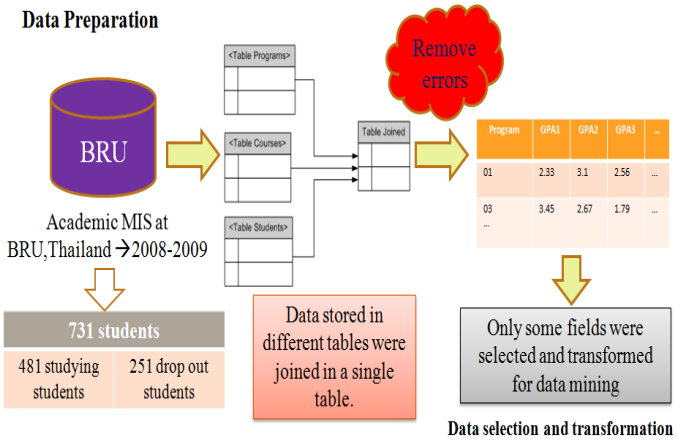


Fig. 8 Data processing steps to build model for higher educational student dropout

Factors related to the student before admission
The student background that effect to student drop out →GPAX from high school, program to study from high school and school size.

Factors related to students during the study periods in the university
The major causes of students to drop out. → study program, GPA score from the first 4 terms, and student loan.

All factors
All of above including → cause to drop out, drop term and drop out status which are the target value to be predicted for factors analysis that effect to student drop out.

Fig. 9 The analysis points related to student dropout

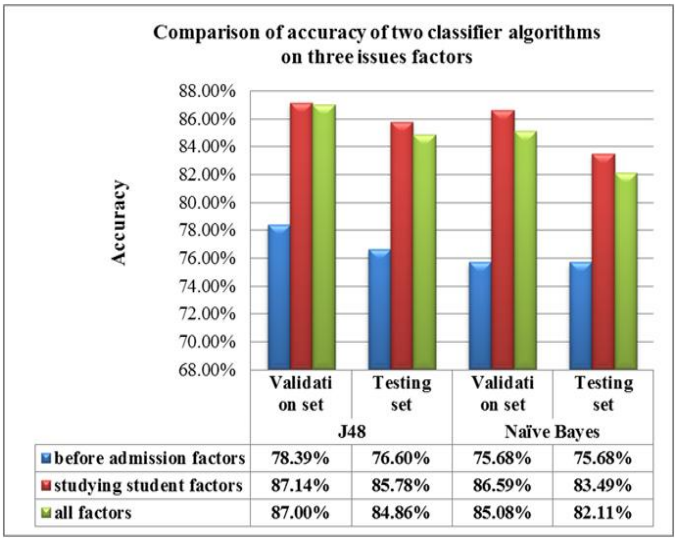


Fig. 10 A comparative result of tree model against naïve Bayes model on each analysis point

VII. Conclusions and Future Research Directions

Recent developments in information and communication technology certainly influence the design and implementation of educational systems. The emergence of the World Wide Web in the early 1990s has resulted in substantial change in both the content representation and the delivery mechanism. XML (eXtensible Markup Language) is an acclaimed technology as the main content representation format. The main advantage of XML over other hypertext languages, such as HTML, is its property of being in an interoperable data interchange format. This language technology helps improve the courseware design concept toward interoperability, which is one of the major design criteria for most software products these days.

Improvements in hardware and communication technology, such as mobile microprocessors and wireless communication, have evolved learning environments into mobile and distributed platforms. Such information infrastructures provide learners with remote access to experts and distributed resources. *Web Browser* is a software device that offers learners open and flexible accessibility to distant course contents. The organization of learning resources has also been changed from creating and delivering large inflexible course content to producing database-driven learning objects that can be reused, searched, and modified independently of the delivery media.

Interoperability, accessibility, and reusability are therefore the main design concepts of current instructional systems. These concepts have shifted the learning paradigm from static learning to collaborative, adaptive, and autonomous learning. Collaborative learning allows for direct contact between instructors and groups of learners through the communication technologies ranging from e-mail and shared workspaces to video conference systems. Adaptive and autonomous learning allow learners to take control over their own organization of learning pace and scheme.

In parallel to advancement of the Web technology, knowledge discovery in databases or data mining has also emerged as a new field. Knowledge discovery in databases is the process of searching and extracting hidden patterns from data that are too large to be efficiently analyzed by humans with simple tools such as a spreadsheet program or database software. Instead, analysis of such huge datasets has to be done automatically via a suite of complicated software. The knowledge discovery process should be done in an intelligent manner and provides useful knowledge in a reasonable time. Knowledge typically appears in a form of patterns that reflect any kind of relationships existing among data attributes. Such relationships include classification rules, association rules, and characteristics of data subgroups in summarized form. However, the discovered patterns represent local relationships, instead of the global ones, because the discovery process is performed on data samples or only some portions of the whole database. The patterns may change if the data samples are modified. Therefore, evaluation on the discovered patterns is an essential post-mining step that has to be done under the supervision of domain experts prior to the delivery of discovered patterns to the stage of knowledge deployment.

The new proposal that embeds knowledge management modules to induce knowledge and manage stored knowledge objects can efficiently supplement the content management and the learning management modules currently existing in most Web-based learning systems. In this chapter, we argue that with the matured technology of knowledge discovery in databases, the integration of knowledge-mining capability to the creation, delivery, and management of learning and knowledge objects should be the next step in e-learning. The proposed architecture of learning environments will enable the convergence of e-learning with knowledge management. A repository containing learner-related materials is a valuable source of knowledge to support personalization information for independent learners.

Despite the promising results of knowledge mining applicable to a limited domain of Web-based learning presented in the literature, the practical aspect of such applications is still in its infancy. This is due to the fact that knowledge mining is not a systematic task; it requires intuition and experiences in adjusting the techniques at every step to obtain the most relevant and actionable knowledge. Knowledge mining is still a task of experts, not at all for a novice or occasional user.

One solution that would abet the improvement of knowledge-mining techniques for typical educators that are not the experts in this field is to customize the process and make the technique more user-friendly. To achieve such a goal, constraint mining [13], in which the mining engine can be made more specific through constraint specifications, and higher-order mining [36], in which the mining results can be mined again to deliver only interesting and useful knowledge, may be used. More specifically, mining algorithms should be made more powerful to provide users with the answers they are looking for. In order to serve the specific needs of users from various and totally different fields, such as social science researchers and medical practitioners, the mining system should be made domain-specific with the intelligent user interface. Existing general mining systems such as Weka can, nevertheless, convey some useful knowledge to users at the expense of quite significant learning time.

The issue of knowledge object representation and managing for the mobile devices working on stream data is also a research challenge for the next decade. These devices are, by nature, limited in memory capacity. Knowledge caching and storing schemes need specific design for such mobile-learning environment.

Current Web-based learning environments provide tools and mechanisms to support knowledge creation and delivery, electronic document access and management, e-learning, and knowledge sharing through collaborative workspaces. Standards such as SCORM for learning objects, modeling languages, and content structures enable the rapid generation of instructional materials and other learning-related elements.

These materials in electronic form are widely dispersed and stored in various servers globally. The trends for the next generation of learning environments are the ability to efficiently exchange the available instructional materials, the functionality to create on-the-fly courseware contents that serve the specific needs of the learner, and the support for advanced technologies that allow the learning system to search Web documents semantically and organize knowledge assets intelligently.

Learning environments in the new decade require an efficient fusion mechanism to integrate new technologies such as semantic Web, smart agents, and declarative knowledge-mining engines. Semantic Web is a concept evolved from Web technology in which the semantics (or meaning) of Web documents are defined to facilitate the search on Web content [3]. A new form of Web content with attached meaning may introduce some standards for the knowledge asset format and make the knowledge exchange and sharing more feasible.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of ACM SIGMOD International Conference on Management of Data*, 1993, pp.207-216.

[2] S. Amershi and C. Conati, "Unsupervised and supervised machine learning in user modeling for intelligent learning environments," *Proceedings of 12th International Conference on Intelligent User Interface*, 2007, pp.72-81.

[3] G. Antoniou and F. van Harmelen, *A Semantic Web Primer*, 2nd edition, Cambridge, MA: MIT Press, 2008.

[4] C.B. Baruque and R.N. Melo, "Using data mining for the refresh of learning objects digital libraries," *WSEAS Transactions on Computers*, vol. 5, no. 11, 2006, pp.2662-2667.

[5] Y. Biletskiy, H. Baghi, I. Keleberda, and M. Fleming, "An adjustable personalization of search and delivery of learning objects to learners," *Expert Systems with Applications*, 2009. doi:10.1016/j.eswa.2008.12.038.

[6] J. Ceddia, J. Sheard, and G. Tibbey, "WAT-A tool for classifying learning activities from a log file," *Proceedings of 9th Australasian Computing Education Conference*, 2007, pp.11-17.

[7] C. Chen, "Intelligent Web-based learning system with personalized learning path guidance," *Computers & Education*, vol. 51, 2008, pp.787-814.

[8] C. Chen and M. Chen, "Mobile formative assessment tool based on data mining techniques for supporting Web-based learning," *Computers & Education*, vol. 52, 2009, pp.256-273.

[9] C. Chen, Y. Hsieh, and S. Hsu, "Mining learner profile utilizing association rule for Web-based learning diagnosis," *Expert Systems with Applications*, vol. 33, 2007, pp.6-22.

[10] N. Chen, Kinshuk, C. Wei, and H. Chen, "Mining e-learning domain concept map from academic articles," *Computers & Education*, vol. 50, 2008, pp.1009-1021.

[11] S. Chen and X. Liu, "An integrated approach for modeling learning patterns of students in Web-based instruction: a cognitive style perspective," *ACM Transactions on Computer-Human Interaction*, vol. 15, no. 1, 2008, pp.1:1-1:28. doi:10.1145/1352782.1352783.

[12] P. De Bra, A. Aerts, B. Berden, B. De Lange, B. Rousseau, T. Santic, D. Smiths, and N. Stash, "AHA! the adaptive hypermedia architecture," *Proceedings of 14th ACM Conference on Hypertext and Hypermedia*, 2003, pp.81-84.

[13] L. De Raedt, T. Guns, and S. Nijssen, "Constraint programming for itemset mining," *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 2008, pp.204-212.

[14] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: AAAI/MIT Press, 1996.

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd edition, San Francisco, CA: Morgan Kaufmann, 2006.

[16] M.S. Handi, "MASACAD: A multi-agent approach to information customization for the purpose of academic advising of students," *Applied Soft Computing*, vol. 7, 2007, pp.746-771.

[17] P. Karampiperis and D. Sampson, "Automatic learning object selection and sequencing in web-based intelligent learning systems," In Z. Ma (ed.), *Web-Based Intelligent E-Learning Systems: Technologies and Applications*, Hershey, PA: Information Science Publishing, 2006.

[18] S.A. Kazi, "A conceptual framework for Web-based intelligent learning environments using SCORM-2004," *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 12-15. doi:10.1109/ICALT.2004.1357365.

[19] N. Kerdprasop and K. Kerdprasop, "Knowledge mining in Web-based learning environments," *International Journal of Social Sciences*, vol. 3, no. 2, 2008, pp.80-83.

[20] A. Kristofic and M. Bielikova, "Improving adaptation in Web-based educational hypermedia by means of knowledge discovery," *Proceedings of 16th ACM Conference on Hypertext and Hypermedia*, 2005, pp.184-192.

[21] Y. Kritikou, P. Demestichas, E. Adamopoulou, K. Demestichas, M. Theologou, and M. Paradia, "User profile modeling in the context of Web-based learning management systems," *Journal of Network and Computer Applications*, 2007. doi:10.1016/j.jnca.2007.11.006.

[22] G. Kulvietis, J. Mamcenko, and I. Sileikiene, "Data mining application for distance education information system," *WSEAS Transactions on Information Science and Applications*, vol. 3, no. 8, 2006, pp.1482-1488.

[23] D. Larose, *Data Mining: Methods and Model*, New Jersey: John Wiley & Sons, 2006.

[24] E. Lazcorreta, F. Botella, and A. Fernandez-Caballero, "Towards personalized recommendation by two-step modified Apriori data mining algorithm," *Expert System with Applications*, 2007. doi:10.1016/j.eswa.2007.08.048.

[25] M.W. Lee, S.Y. Chen, K. Chrysostomou, and X. Liu, "Mining students' behavior in Web-based learning programs," *Expert System with Applications*, vol. 36, 2009, pp.3459-3464.

[26] Q. Li, R. Lau, T. Shih, and F. Li, "Technology supports for distributed and collaborative learning over the Internet," *ACM Transactions on Internet Technology*, vol. 8, no. 2, 2008, pp.10:1-10:24. doi:10.1145/1323651.1323656.

[27] Z. Liu and Y. Liu, "An adaptive personalized e-learning model based on agent technology," *WSEAS Transactions on Systems*, vol. 7, no. 12, 2008, pp.1443-1452.

[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol.1, 1967, pp.281-297.

[29] J. Mamcenko and I. Sileikiene, "Examination of e-learning system based on intelligent data analysis," *WSEAS Transactions on Information Science and Applications*, vol. 4, no. 4, 2007, pp.859-865.

[30] B. Marshall, H. Chen, R. Shen, and E. Fox, "Moving digital libraries into the student learning space: the GetSmart

experience," *ACM Journal on Educational resources in Computing*, vol. 6, no. 1, 2006, pp.1-20.
[31] P. Markellou, M. Rigou, and S. Sirmakessis, "Knowledge mining: a quantitative synthesis of research results and findings," In S. Sirmakessis (ed.), *Knowledge Mining*, The Netherlands: Springer-Verlag, 2005.
[32] M. Martin, A. Alvarez, I. Fernandez-Castro, and M. Urretavizcaya, "Generating teacher adapted suggestions for improving distance educational systems with SIgMa," *Proceedings of 8th IEEE International Conference on Advanced Learning Technologies*, 2008, pp.449-453.
[33] R.S. Michalski, "Knowledge mining: A proposed new direction," Invited Talk at the *Sanken Symposium on Data Mining and Semantic Web*, Osaka University, Japan, 10-11 March 2003. Retrieved from http://www.mli.gmu.edu/papers/2003-2004/03-5.pdf.
[34] C. Pahl, "Managing evolution and change in Web-based teaching and learning environments," *Computers & Education*, vol. 40, 2003, pp.99-114.
[35] M. Rigou and S. Sirmakessis, "Bridging personalization to online learning communities," *WSEAS Transactions on Information Science and Applications*, vol. 2, no. 12, 2005, pp.2160-2167.
[36] J.F. Roddick, M. Spiliopoulou, D. Lister, and A. Ceglar, "Higher order mining," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 1, 2008, pp.5-17.
[37] C. Romero, P. Gonzalez, S. Ventura, M.J. del Jesus, and F. Herrera, "Evolutionary algorithms for subgroup discovery in e-learning: a practical application using Moodle data," *Expert Systems with Applications*, vol. 36, 2009, pp.1632-1644.
[38] C. Romero and S. Ventura, "Educational data mining: a survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, 2007, pp.135-146.
[39] C. Romero, S. Ventura, and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, 2008, pp.368-384.
[40] J. Sable and N. Gaviola, "NCES common core of data local education agency – level public-use data file on public school dropouts: school year 2003-04 (NCES 2007-372)," *National Center for Education Statistics*, Institute of Education Sciences, U.S. Department of Education, Washington, D.C., 2007. (http://nces.ed.gov/pubsearch/).
[41] C. Shih, D. Chiang, S. Lai, and Y. Hu, "Applying hybrid data mining techniques to Web-based self-assessment system of study and learning strategies inventory," *Expert Systems with Applications*, 2008. doi:10.1016/j.eswa.2008.06.089.
[42] P. Tzouveli, P. Mylonas, and S. Kollias, "An intelligent e-learning system based on learner profiling and learning resources adaptation," *Computers & Education*, 2007. doi:10.1016/j.compedu.2007.05.005.
[43] Y. Wang, M. Tseng, and H. Liao, "Data mining for adaptive learning sequence in English language instruction," *Expert Systems with Applications*, 2008. doi:10.1016/j.eswa.2008.09.008.
[44] I. H. Witten and E. Frank, Data *Mining: Practical Machine Learning Tools and Techniques*, 2nd edition, San Francisco: Elsevier/Morgan Kaufmann, 2005.
[45] D. Xu and H. Wang, "Intelligent agent supported personalization for virtual learning environments," *Decision Support Systems*, vol. 42, 2006, pp.825-843.
[46] O. Zaiane, "Web usage mining for a better Web-based learning environment," *Proceedings of the Advanced Technology for Educational Conference*, 2001, pp.60-64.

Nittaya Kerdprasop is an associate professor and the director of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology, Thailand. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. She is a member of IAENG, ACM, and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Data Mining, Artificial Intelligence, Logic and Constraint Programming, Deductive and Active Databases.

Ekkachai Naenudorn is currently a doctoral student with the School of Information Technology, Suranaree University of Technology, Thailand. His research topic is related to educational data mining, information technology, e-learning and internet technology.

Jatsada Singthongchai is currently a doctoral student with the School of Information Technology, Suranaree University of Technology, Thailand. His research topic is related to information technology in higher education, e-learning systems and internet technology.

Wilairat Yathongchai is currently a doctoral student with the School of Information Technology, Suranaree University of Technology, Thailand. Her research topic is related to educational data mining, information technology, e-learning system, content and learning object management.

Chusak Yathongchai is currently a doctoral student with the School of Information Technology, Suranaree University of Technology, Thailand. His research interest includes educational data mining, information technology, internet technology and intelligent systems.

Kittisak Kerdprasop is an associate professor at the School of Computer Engineering and one of the principal researchers of Data Engineering Research Unit, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, USA, in 1999. His current research includes Data mining, Machine Learning, Artificial Intelligence, Logic and Functional Programming, Probabilistic Databases and Knowledge Bases.