A Knowledge-Oriented Framework for the New Age of Training and Learning Systems

Kittisak Kerdprasop, Nittaya Kerdprasop

Knowledge Engineering Research Unit, School of Computer Engineering,
Suranaree University of Technology, Nakhon Ratchasima, Thailand

Corresponding Author: kerdpras@sut.ac.th

Abstract

In this paper, we introduce a new concept of training and learning systems that are based on the Web and knowledge mining technologies. Knowledge mining is the process of deriving hidden and potentially useful information from the current and archival databases. Derived knowledge can possibly be in various forms of patterns, such as data summarization and relationships among data instances. We propose the design of a Web-based training and learning system that incorporates knowledge management modules to induce knowledge and manage stored knowledge objects. This new module is intended to supplement the content management and the learning management modules that exist in most Web-based learning systems. Our proposed architecture of learning environments will enable the convergence of e-learning with the knowledge management. A repository containing learner-related materials is a valuable source of knowledge to support personalization information for independent learners. The contribution of our work is the design of a knowledge-oriented system to provide an integrated, flexible, and efficient platform for intelligent learning environment suitable for both vocational training and higher educational systems.

Keywords: Web-based learning framework, knowledge mining, intelligent training system.

1. Introduction

During the last few years, we have witnessed an important shift in educational systems. 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 offering learners with 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.

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, known as e-learning, and distance education. These facilities include course creation and delivery, learning tracking and assessment, enrollment, and administration. 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 designed once and used by any learner regardless of their diversities in capabilities, needs, and perceptions^(1,2). Since then, the issues of content adaptability and intelligent curriculum sequencing have become the major research goals for the development of advanced educational systems⁽³⁻⁶⁾.

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, especially those that support higher education. In this paper, we discuss the adoption of inductive learning abilities from the knowledge mining field to create such intelligent learning environments. A brief review of

knowledge mining technology, a conceptual framework of intelligent systems, as well as running examples are also presented in the next section.

2. Knowledge Mining Tasks and Technology

2.1 Knowledge Mining Technology

Knowledge mining is the discovery of hidden knowledge stored possibly in various forms and places 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 data set contains four data; each data instance has been represented as an (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 this paper, we refer to both pattern and model as new knowledge automatically discovered from data sources. The process of knowledge mining works around data, meta-data, and previously discovered patterns (conceptually shown in Fig. 1).

The first step of knowledge mining focuses on setting the mining goal, which can be achieved through understanding the task objectives and organization requirements. A clear problem statement should define what is to be accomplished and what the desired outcome is likely to be.

The second step covers all activities necessary for preparing high quality data suitable for mining via algorithms. This step includes collecting data from multiple sources, transforming the data format and selecting data representatives with minimum but sufficient attributes. Meta-data and background knowledge are kinds of supportive information that can be applied in this step.

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

Fig. 1. The process of knowledge mining.

The fourth step is to evaluate 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. Interestingness of the induced model is somehow a more subtle issue than the accuracy metric. Evaluating interestingness depends considerably on the judgment of a knowledge miners or domain experts. Accurate and interesting knowledge is finally fed to the deployment step to be actionable information for the organization or to act as background knowledge for other knowledge mining tasks.

2.2 Example Data Set

The mining step (the third step in Fig. 1) can be conducted with different kinds of algorithms. These algorithms may be grouped roughly into three categories: classification, association, and cluster analysis. In the next sub-sections, each mining task will be demonstrated through the example data containing educational transitions of 500 Irish schoolchildren age 11 in the year 1967. This data was collected by Greaney and Kelleghan⁽⁷⁾ and stored in statlib (<http://lib.stat.cmu.edu/datasets/>). Samples of this data are shown in Table 1.

Table 1. Irish schoolchildren data set.

Sex	DVRT (Drumcondra Verbal Reasoning Test Score)	Educational level attained	Leaving certificate	Prestige score for father's occupation	Type of school*
male	113	3	not_taken	28	1
male	101	1	not_taken	28	9
male	110	9	Taken	69	1
...	...	...	...	...	...
female	109	4	not_taken	43	2

*Type of school: 1= secondary; 2= vocational; 9= primary terminal leaver

The Irish school system comprises eight years of primary education followed by six years of secondary cycle⁽⁸⁾. The primary and the first three years of secondary levels are compulsory. After the secondary school, students may take the National Leaving Certificate Examination in order to gain access to higher education, also called third-level or tertiary education. Children in Ireland attend primary school at the age ranging from four to six. Primary schools consist of two kindergarten years, junior and senior infants, followed by classes 1-6 of senior primary level.

Secondary education comprises the first three years period (called the junior cycle) in which at the end of the period, students have to take the Junior Certificate Examination⁽⁸⁾. Following the examination is a one-year program, called a transition year. During the year, students may take vocational courses such as drama, motor vehicle maintenance, or obtain work experience for a short period. The final two years in secondary school, called the senior cycle, is the preparation time for the Leaving Certificate Examination.

The data set is about educational disadvantage among Irish schoolchildren in 1967. Some students had to terminate their education at the primary level (encoded as 1 in the 'Educational level attained' column of Table 1), while others could further their study but had to leave school at different levels prior to graduation. Secondary education in this dataset includes secondary and vocational schools. The other details regarding gender, ability in verbal reasoning (displayed as DVRT scores), and father's occupation (computed as prestige score – the higher is the better) of each student are also provided in the dataset.

Our objective is to draw some patterns from the data of 500 Irish schoolchildren. We use this data set to illustrate various steps that are necessary to mine useful knowledge using the Weka (Waikato environment for knowledge analysis) system (<http://www.cs.waikato.ac.nz/ml/weka/>). The mined knowledge is a pattern, displayed in various forms such as a decision tree, a Bayesian network, association rules, or representative characteristics of each data cluster.

2.3 Knowledge Classification Task

The classification objective in our running example is to find major characteristics that can classify correctly *Educational_level_attended* of the Irish schoolchildren. The algorithm applied for this task is J48⁽⁹⁾, which is known as a decision tree induction algorithm because it can find (or induce) pattern from the given data and display the pattern

graphically as tree (as shown in Fig. 2). It may be noticed from the model that to predict the educational level, we can make predictions simply based only on the two factors: *Leaving_certificate* (taken/ not_taken) and *Type_school* (1= secondary, 2= vocational, 9= primary terminal leaver).

Consider the left branch of this tree as an example of model interpretation. This branch reveals the pattern that there are 110 students who had not taken a leaving certificate but studied in secondary school (*Type_school* = 1). According to the model, this group of students is predicted to have *Educational_level_attended* of 5, which are those who attend the secondary school but have to quit at the junior level. Out of these 110 students, there are 45 students whose actual *Educational_level_attended* value is not 5. The proportion 45/110, which is 0.409, is inherent error (also called training error) of this model for classifying and predicting *Educational_level_attended*.

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.

Fig. 2. Decision tree for classifying and predicting *Educational_level_attended*.

Fig. 3. Bayesian network representing causal relationships of *Educational_level_attended* and *Type_school* attributes.

Bayesian network⁽¹⁰⁾ 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 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 nondescendants in the network. Consider the network in Figure 3; the arc linking *Educational_level_attended* and *Type_school* implies dependency between these two attributes. The strength of this dependency is given by probability value such as:

$$P(\text{Type\_school}=9 \mid \text{Educational\_level\_attended}=1) = 0.974.$$

2.4 Association Knowledge Mining

With the same set of data, we can perform *association analysis* to discover correlations or relationships that hold among data attributes. The following rules show some results of such analysis conducted with Apriori algorithm⁽¹¹⁾.

IF *educational_level_attended* is 9,
 THEN *student has taken leaving_certificate*. (conf.=1)
 IF *educational_level_attended* is 9,
 THEN *type of school is secondary school*. (conf.=1)
 IF *sex is female and educational_level_attended* is 9,
 THEN *student has taken leaving_certificate*. (conf.=1)
 IF *student has taken leaving_certificate and score of father's occupation is 30-37.5*,
 THEN *type of school is secondary school*. (conf.=0.98)

Each association rule is annotated with confidence value to give information regarding how accurate the rule is. The value 1 is the most accurate association; whereas the value less than one implies that the rule might contain some error. Taking the last association rule as an example, there might be some case that the student has taken *leaving_certificate* and score of father's occupation is in the range 30-37.5, but type of school is not secondary school.

2.5 Cluster Knowledge Mining

Another commonly performed mining task is *cluster analysis*. This task is unsupervised learning because the clustering algorithm will not be guided what the correct cluster each data instance should be. Therefore, there is no straightforward way to grade the performance of cluster algorithm. The best we can do is to determine the cohesion

that each data cluster has formed. This means that data that are similar should be assigned into the same cluster, whereas different data should be in separate cluster. Several algorithms can be applied to perform cluster analysis task such as EM and Cobweb⁽⁹⁾, but the most widely used algorithm is the k-means algorithm⁽¹²⁾. The parameter *k* is the number of clusters that users have to specify. The result of running k-means is the cluster centroids (or mean value of a cluster) that report the characteristics of representatives (majority value) in each cluster.

In Table 2, the result of running k-means on Irish schoolchildren dataset with *k* = 2 has been summarized. It can be noticed from the clustering results that when *k* is 2, students who are in the category of primary terminal leaver (*Type_school* = 9) are absent from the characteristics of students in both groups.

We further the experiment by increasing the number of clusters to be seven (*k* = 7), this group of students has thus been captured with the value (sex = male, DVRT = 82.0244, educational level attended = 1, leaving certificate = not_taken, score father occupation = 25.0488) reported as its mean (clustering result shown in Fig.4).

These running examples emphasize the fact that knowledge mining is an iterative process in that if the mined knowledge is insufficient, irrelevant or inaccurate, we have to repeat the process with some parameter adjustment.

Table 2. Cluster analysis result with k-means algorithm (k=2).

Cluster characteristics	Cluster 0 (# students=212)	Cluster 1 (# students=288)
Sex	male	Female
DVRT	93.3208	105.1806
Educational level attended	4	9
Leaving certificate	Not_taken	Taken
Score father occupation	31.6368	40.7917
Type_school	2 (vocational)	1 (secondary)

Attribute	Full Data (500)	Cluster#						
		0 (128)	1 (88)	2 (38)	3 (62)	4 (43)	5 (100)	6 (41)
Sex	male	male	female	female	male	male	female	male
DVRT	100.152	93.5391	101.2045	114.1579	108.7903	113.5814	98.67	82.0244
Educational_level_attended	9	4	9	11	9	9	5	1
Leaving_certificate	not_taken	not_taken	taken	taken	taken	taken	not_taken	not_taken
Score_father_occupation	36.91	31.5234	39.3409	42.2632	29.9677	64.3953	36.98	25.0488
Type_school	1	2	1	1	1	1	1	9

Fig. 4. Cluster analysis result when k = 7.

3. An Intelligent Learning Environment

During the past decade, we have witnessed the development of tools customizing knowledge mining techniques to support intelligent educational systems⁽¹³⁻¹⁶⁾. Some tools are embedded in the course management system while some operate stand-alone as knowledge acquisition and representation applications. We thus attempt to achieve a full-scale integration of knowledge intensive tasks by designing the integrated environment for Web-based intelligent learning system.

In our framework of the Web-based intelligent learning environment (Fig. 5), 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. Information is a supplementary of 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 mining and management.

Knowledge is the most sophisticated entity stored in a repository as knowledge objects. Each knowledge object represents the relationship 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.

Fig. 5. A framework of Web-based intelligent learning environment.

Data, information, and knowledge that are stored in a repository are key component of the designed framework. The three major modules of the proposed system communicate through the repository. Some component such as mining engine is data-driven in which it is data dependent and the mining results may be different if the data contents have been changed. The three main modules in our proposed framework are learning management, content management, and knowledge management modules. Our framework is proposed to support Web-based learning with several learning schemes including adaptive, autonomous, and collaborative learning. Capabilities of each component are as listed as follows.

Content management module provides the following capabilities:

- Allows the content developers to import and export content through the authoring tools.
- Allows the content manager to individualize the presented content.
- Enables the content manager to archive the old but useful contents and assign number on different version of contents of the same topic.
- Creates an interface for learning management modules in getting desired form for delivered content.
- Creates an interface with a data repository that contains learners' personal information and other metadata such as knowledge assets that are created by knowledge management modules and apply this data to create a personalized sequence of content material suitable for each learner.

Knowledge management module provides the following capabilities:

- Provides tools to collect different kinds of data such as learners' personal data, tracked data of learners' performance and behavior, and data related to content sequences that were presented in the past with the evaluation results according to that content sequence. These valuable data are stored in the data repository in different files that may take different formats.
- Provides tools to discover valuable knowledge assets from the collected data.
- Supports the indexing and mapping of knowledge objects that are discovered by the knowledge mining engine.

Learning management module provides the following capabilities:

- Enables instructors to post syllabi, class schedules,

assignments, lecture notes, slides and other supplemental materials for learners to access via Web browsing tool.

- Enables instructors to conduct assessments in various forms such as online tests, surveys, quizzes using a variety of standard question formats, e.g. multiple choice, true/false, essay, short-answer, matching, etc.
- Enables learners to submit assignments remotely either as file upload or interactive through Web interface.
- Provides profiling tools to collect personal data of learner and tracking tool to observe learners' actions including like and dislike information.
- Provides tools to compare the created profile with the available contents in order to create and deliver customized learning materials suitable for learner's preference and ability.
- Facilitate instructors and learners to engage in collaborative discussion on assignments and course content.

The major actor of the learning management module is a learning manager component which acts as a conductor controlling and synchronizing every component within the module. The manager component is also responsible for interfacing with the repository. This is also the case for the content manager component in content management module. The content-creation tools in a content management module support the creation of all types of digital content materials such as word documents, spreadsheet data, video content, animation and multimedia data.

For the knowledge management module, the mining engine is responsible for the synchronizing process. Indexing and mapping is a component for storing and searching knowledge objects to be used in the learning process. Meta-data management engine is a tool to incorporate background knowledge that may be useful for the mining component. This knowledge mining and managing model can be graphically shown in Fig. 6.

A top-level of knowledge management model in web-based learning environments is the one shown in Fig. 5. From the data and knowledge repositories, 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 presented to learners. Performance and learners' preference are then captured to store as history and learnable objects to be later used in the discovery stage.

Fig. 6. A knowledge mining and managing model to handle knowledge objects in an intelligent learning system

4. Conclusions and Discussion

Recent developments in information and the internet technologies have certainly influenced the design and implementation of educational systems. The emergence of the World Wide Web has resulted in substantial change in both the content representation and the delivery mechanism. In parallel to advancement of the Web technology, knowledge mining has also emerged as a new field of intelligent techniques. Knowledge mining is the process of searching and extracting hidden patterns from data that are too large to be efficiently analyzed by human with simple tools such as a spreadsheet program and database software. Instead, analysis such huge data has to be done automatically via a suite of complicate software. The knowledge search 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 may be any kind of relationships existing among data attributes. Such relationships include classification rules, association rules, and data subgroup characteristics in summarized form.

However, the discovered patterns represent local relationships, instead of the global ones, because the search 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.

We propose the addition of knowledge management modules to induce knowledge and managing stored knowledge objects to supplement the content management

and the learning management modules currently exist in most Web-based learning systems. In this paper, we argue that with the matured technology of knowledge mining, the integration of mining capability to the creation, delivery and management of learning and knowledge objects should be the next step in e-learning. Our 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.

The contribution of our work is the design of a knowledge-oriented system to provide an integrated, flexible, and efficient platform for intelligent learning environment. Despite the promising results of intelligent Web-based learning presented in our work and others, 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 users. One solution that would abet the improvement of knowledge mining techniques for typical educators or trainers who 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 in which the mining engine can be made more specific through constraint specifications and higher-order mining may be used. Most specifically, mining algorithms are made more powerful by mining from learned patterns, providing us with the answers we are looking for.

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. Learning environments in the new decade also require an efficient fusion mechanism to integrate new technologies such as semantic Web, smart agents, and declarative mining engines.

Semantic Web is a concept evolved from the Web technology in which the semantics of Web documents are defined to facilitate the search on Web content. A formal specification method is used for the description of concepts, terms and relationships within a Web document. 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.

Acknowledgment

This research has been conducted at the knowledge engineering research unit, school of computer engineering, Suranaree University of Technology (SUT). This research unit is financially supported by research grants from the SUT and National Research Council of Thailand (NRCT).

References

- (1) P. Tzouveli, P. Mylonas, and S. Kollias : "An intelligent e-learning system based on learner profiling and learning resources adaptation", *Computers & Education*, Vol. 51, No. 1, pp. 224-238, 2008.
- (2) 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, pp. 10:1-10:24, 2008.
- (3) C. Pahl : "Managing evolution and change in Web-based teaching and learning environments", *Computers & Education*, Vol. 40, pp. 99-114, 2003.
- (4) S. Kazi : "A conceptual framework for Web-based intelligent learning environments using SCORM-2004", *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04)*, pp. 12-15, 2004.
- (5) D. Xu, and H. Wang : "Intelligent agent supported personalization for virtual learning environments", *Decision Support Systems*, Vol. 42, pp. 825-843, 2006.
- (6) 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*, Vol. 1, No. 4, pp. 603-627, 2008.
- (7) V. Greaney, and T. Kelleghan : *Equality of Opportunity in Irish Schools*. Dublin: Educational Company, 1984.
- (8) J. Coolahan : *Irish Education : Its History and Structure*. Dublin: Institute of Public Administration, 1981.
- (9) I. Witten, and E. Frank : *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. San Francisco: Elsevier/Morgan Kaufmann, 2005.
- (10) J. Han, and M. Kamber : *Data Mining: Concepts and Techniques*, 2nd edition. San Francisco, CA: Morgan Kaufmann, 2006.
- (11) 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*, pp.

207-216, 1993.

- (12) J. MacQueen : “Some methods for classification and analysis of multivariate observations”, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297, 1967.
- (13) M. Handi : “MASACAD: A multi-agent approach to information customization for the purpose of academic advising of students”, *Applied Soft Computing*, Vol. 7, pp. 746-771, 2007.
- (14) C. Chen : “Intelligent Web-based learning system with personalized learning path guidance”, *Computers & Education*, Vol. 51, pp. 787-814, 2008.
- (15) 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*, Vol. 36, No. 5, pp. 9113-9120, 2009.
- (16) C. Chen, and M. Chen : “Mobile formative assessment tool based on data mining techniques for supporting Web-based learning”, *Computers & Education*, Vol. 52, pp. 256-273, 2009.