Abstract—In this paper we give a definition of KDD and Data Mining, describing its tasks, methods, and applications. Our motivation in this study is gaining the best technique for extracting useful information from large amounts of data in an online educational system, in general, and from the LON-CAPA system, in particular. The goals for this study are: to obtain an optimal predictive model for students within such systems, help students use the learning resources better, based on the usage of the resource by other students in their groups, help instructors design their curricula more effectively, and provide the information that can be usefully applied by instructors to increase student learning.

Index Terms—LON-CAPA system, Knowledge Discovery in Databases (KDD), Data Mining

I. INTRODUCTION

Presently, the amount of data stored in databases is increasing at a tremendous speed. This gives rise to a need for new techniques and tools to aid humans in automatically and intelligently analyzing huge data sets to gather useful information. This growing need gives birth to a new research field called Knowledge Discovery in Databases (KDD) or Data Mining, which has attracted attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization.

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial Process of identifying valid, novel, potentially useful, and ultimately understandable Patterns in data" (1).

The process of data mining uses machine learning, statistics, and visualization Techniques to discover and present knowledge in a form that is easily comprehensible. The word “Knowledge” in KDD refers to the discovery of patterns which are extracted from the processed data. A pattern is an expression describing facts in a subset of the data. Thus, the difference between KDD and data mining is that “KDD refers to the overall process of discovering knowledge from data while data mining refers to application of algorithms for extracting patterns from data without the additional steps of the KDD process.” (1)

However, since Data Mining is a crucial and important part of the KDD process, Most researchers use both terms interchangeably. Figure 1.1 presents the iterative nature of the KDD process. Here we outline some of its basic steps as mentioned in Brachman & Anad (1996):

- Providing an understanding of the application domain, the goals of the system And its users, and the relevant prior background and prior knowledge (This step In not specified in this figure.)
- Selecting a data set, or focusing on a subset of variables or data samples, on Which discovery is to be performed?
- Preprocessing and data cleansing, removing the noise, collecting the necessary Information for modeling, selecting methods for handling missing data fields, Accounting for time sequence information and changes
- Data reduction and projection, finding appropriate features to represent data, Using dimensionality reduction or transformation methods to reduce the number Of variables to find invariant representations for data
- Choosing the data mining task depending on the goal of KDD: clustering, Classification, regression, and so forth
- Selecting methods and algorithms to be used for searching for the patterns in the Data
- Mining the knowledge: searching for patterns of interest
- Evaluating or interpreting the mined patterns, with a possible return to any Previous steps

Figure 1.1 Steps of the KDD Process (Fayyad et al., 1996)

Manuscript received May, 2012.
Prof. Darshana Kishorbhai Dave, Production Engineering Dept., Government Engineering College, Bhavnagar, Gujarat, India.
II Data Mining Methods

The objective of data mining is both prediction and description. That is, to predict unknown or future values of the attributes of interest using other attributes in the Databases, while describing the data in a manner understandable and interpretable to Humans. Predicting the sale amounts of a new product based on advertising expenditure, or predicting wind velocities as a function of temperature, humidity, air pressure, etc., are examples of tasks with a predictive goal in data mining. Describing the different terrain groupings that emerge in a sampling of satellite imagery is an example of a descriptive goal for a data mining task. The relative importance of description and prediction can vary between different applications. These two goals can be fulfilled by any of a number data mining tasks including: classification, regression, clustering, summarization, dependency modeling, and deviation detection. (2)

Predictive tasks

The following are general tasks that serve predictive data mining goals:

• Classification – to segregate items into several predefined classes. Given a Collection of training samples, this type of task can be designed to find a Model for class attributes as a function of the values of other attributes (3).

• Regression – to predict a value of a given continuously valued variable based on the values of other variables, assuming either a linear or nonlinear model of Dependency. These tasks are studied in statistics and neural network fields (4).

• Deviation Detection – to discover the most significant changes in data from Previously measured or normative values (5). Explicit information outside the data, like integrity constraints or predefined patterns, is used for deviation detection. Arning et al., (1996) approached the problem from the inside of the data, using the implicit redundancy.

Descriptive tasks

• Clustering – to identify a set of categories, or clusters, that describe the data (1).

• Summarization – to find a concise description for a subset of data. Tabulating the mean and standard deviations for all fields is a simple example of summarization. There are more sophisticated techniques for summarization and they are usually applied to facilitate automated report generation and interactive data analysis (2).

• Dependency modeling – to find a model that describes significant dependencies between variables. For example, probabilistic dependency networks use conditional independence to specify the structural level of the model and probabilities or correlation to specify the strengths (quantitative level) of dependencies (3).

Mixed tasks

There are some tasks in data mining that have both descriptive and predictive aspects. Using these tasks, we can move from basic descriptive tasks toward higher-order predictive tasks. Here, we indicate two of them:

• Association Rule Discovery – Given a set of records each of which contain some number of items from a given collection, produce dependency rules which will predict the occurrence of an item based on patterns found in the data.

• Sequential Pattern Discovery – Given a set of objects, where each object is associated with its own timeline of events, find rules that predict strong sequential dependencies among different events. Rules are formed by first discovering patterns followed by event occurrences which are governed by timing constraints found within those patterns.

So far we briefly described the main concepts of data mining. Chapter two focuses on methods and algorithms of data mining in the context of descriptive and predictive tasks. The research background of both the association rule and sequential pattern mining – newer techniques in data mining, that deserve a separate discussion. Data mining does not take place in a vacuum. In other words, any application of this method of analysis is dependent upon the context in which it takes place. Therefore, it is necessary to know the environment in which we are going to use data mining methods. The next section provides a brief overview of the LON-CAPA system.

Online Education systems

Several Online Education systems such as Blackboard, WebCT, Virtual University (VU), and some other similar systems have been developed to focus on course management issues. The objectives of these systems are to present courses and instructional programs through the web and other technologically enhanced media. These new technologies make it possible to offer instruction without the limitations of time and place found in traditional university programs. However, these systems tend to use existing materials and present them as a static package via the Internet. There is another approach, pursued in LON-CAPA, to construct more-or-less new courses using newer network technology. In this model of content creation, college faculty, K-12 teachers, and students interested in collaboration can access a database of hypermedia software.

LON-CAPA, System Overview

LON-CAPA is a distributed instructional management system, which provides students with personalized problem sets, quizzes, and exams. Personalized (or individualized) homework means that each student sees a slightly different computer-generated problem. LON-CAPA provides students and instructors with immediate feedback on conceptual understanding and correctness of solutions. It also provides
faculty the ability to augment their courses with individualized, relevant exercises, and develop and share modular online resources. LON-CAPA aims to put this functionality on a homogeneously distributed platform for creating, sharing, and delivering course content with emphasis on cross-institutional collaboration and intellectual property rights management.

LON-CAPA Topology

LON-CAPA is physically built as a geographically distributed network of constantly connected servers. Figure 1.2 shows an overview of this network. All machines in the network are connected with each other through two-way persistent TCP/IP connections. The network has two classes of servers: library servers and access servers. A library server can act as a home server that stores all personal records of users, and is responsible for the initial authentication of users when a session is opened on any server in the network. For authors, it also hosts their construction area and the authoritative copy of every resource that has been published by that author. An Access Server is a machine that hosts student sessions. Library servers can be used as backups to host sessions when all access servers in the network are overloaded. Every user in LON-CAPA is a member of one domain. Domains could be defined by departmental or institutional boundaries like MSU, FSU, OHIU, or the name of a publishing company. These domains can be used to limit the flow of personal user information across the network, set access privileges, and enforce royalty schemes. Thus, the student and course data are distributed amongst several repositories. Each user in the system has one library server, which is his/her home server. It stores the authoritative copy of all of their records.

LON-CAPA currently runs on Redhat-Linux Intel-compatible hardware. The LON-CAPA currently runs on Redhat-Linux Intel-compatible hardware. The current MSU production setup consists of several access servers and some library servers. All access servers are set up on a round-robin IP scheme as frontline machines, and are accessed by the students for “user session.” The current implementation of LON-CAPA uses mod_perl inside of the Apache web server software.

Data Distribution in LON-CAPA

Educational objects in LON-CAPA range from simple paragraphs of text, movies, and applets, to individualized homework problems. Online educational projects at MSU have produced extensive libraries of resources across disciplines. By combining these resources, LON-CAPA produces a national distributed digital library with mechanisms to store and retrieve these objects. Participants in LON-CAPA can publish their own objects in the common pool. LON-CAPA will allow groups of organizations (departments, universities, schools, commercial businesses) to link their online instructional resources in a common marketplace, thus creating an online economy for instructional resources (lon-cap.org). Internally, all resources are identified primarily by their URL. LON-CAPA does enable faculty to combine and sequence these learning objects at several levels. For example, an instructor from Community College A in Texas can compose a page by combining a text paragraph from University B in Detroit with a movie from College C in California and an online homework problem from Publisher D in New York. Another instructor from High School E in Canada might take that page from Community College A and combine it with other pages into a module, unit or section. Those in turn can be combined into whole course packs.

Resource Variation in LON-CAPA

LON-CAPA provides three types of resources for organizing a course. LON-CAPA refers to these resources as Content Pages, Problems, and Maps. Maps may be either of two types: Sequences or Pages. LON-CAPA resources may be used to build the outline, or structure, for the presentation of the course to the students.
A Content Page displays course content. It is essentially a conventional HTML page. These resources use the extension “.html”.

- A Problem resource represents problems for the students to solve, with answers stored in the system. These resources are stored in files that must use the extension “.problem”.

- A Page is a type of Map which is used to join other resources together into one HTML page. For example, a page of problems will appear as a problem set. These resources are stored in files that must use the extension “.page”.

- A Sequence is a type of Map, which is used to link other resources together. Sequences are stored in files that must use the extension “.sequence”. Sequences can contain other sequences and pages.

Authors create these resources and publish them in library servers. Then, instructors use these resources in online courses. The LON-CAPA system logs any access to these resources as well as the sequence and frequency of access in relation to the successful completion of any assignment. All these accesses are logged.

LON-CAPA Strategy for Data Storage

Internally, the student data is stored in a directory:/home/httpd/IonUsers/domain/1st.char/2nd.char/3rd.char/username/  
For example /home/httpd/IonUsers/gec/m/i/n/minaeibi/  

Figure 1.3 shows a list of a student’s data. Files ending with .db are GDBM files (Berkeley database), while those with a course-ID as name, for example gec_12679c3ed543a25msul1.db, store performance data for that student in the course.

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>gec_12679c3ed543a25msul1.db</td>
<td>Store performance data for student in the course.</td>
</tr>
</tbody>
</table>

Figure 1.4 Directory listing of course’s home directory

An example of course data is shown in Figure 1.4. classlist is the list of students in the course, environment includes the course’s full name, etc, and resourcedata are deadlines, etc. The parameters for homework problems are stored in these files. To identify a specific instance of a resource, LON-CAPA uses symbols or “symbs.” These identifiers are built from the URL of the map, the resource number of the resource in the map, and the URL of the resource itself. The latter is somewhat redundant, but might help if maps change.

An example of a problem is gec/dkd/parts/part1.sequence__19__gec/dkd/tests/part12.problem The respective map entry is<resource id="19" src="/res/gec/dkd/tests/part12.problem" title="Problem 2" />

Symbs are used by the random number generator, as well as to store and restore data specific to a certain instance of a problem. More details of the stored data and their exact structures will be explained in chapter three, when we will describe the data acquisition of the system.

Resource Evaluation in LON-CAPA

One of the most challenging aspects of the system is to provide instructors with information concerning the quality and effectiveness of the various materials in the resource pool on student understanding of concepts. These materials can include web pages, demonstrations, simulations, and individualized problems designed for use on homework assignments, quizzes, and examinations. The system generates a range of statistics that can be useful in evaluating the degree to which individual problems are effective in promoting formative learning for students. For example, each exam problem contains attached metadata that catalog its degree of difficulty and discrimination for students at different phases in their education (i.e., introductory college courses, advanced college courses, and so on). To evaluate resource pool materials, a standardized format is required so that materials from different sources can be compared. This helps resource users to select the most effective materials available.

LON-CAPA has also provided questionnaires which are completed by faculty and students who use the educational materials to assess the quality and efficacy of resources. In addition to providing the questionnaires and using the statistical reports, we investigate here methods to find criteria for classifying students and grouping problems by examining logged data such as: time spent on a particular resource, resources visited (other web pages), due date for each homework, the difficulty of problems (observed...
statistically) and others. Thus, floods of data on individual usage patterns need to be gathered and sorted—especially as students go through multiple steps to solve problems, and choose between multiple representations of the same educational objects like video lecture demonstrations, a derivation, a worked example, case-studies, and etc. As the resource pool grows, multiple content representations will be available for use by students. There has been an increasing demand for automated methods of resource evaluation.

One such method is data mining, which is the focus of this research. Since the LON-CAPA data analyses are specific to the field of education, it is important to recognize the general context of using artificial intelligence in education. The following section presents a brief review of intelligent tutoring systems—one typical application of artificial intelligence in the field of education. Note that herein the purpose is not to develop an intelligent tutoring system; instead we apply the main ideas of intelligent tutoring systems in an online environment, and implement data mining methods to improve the performance of the educational web-based system, LON-CAPA.

V CONCLUSION

This paper addresses data mining methods for extracting useful and interesting knowledge from the large data sets of students using LON-CAPA educational resources. The purpose is to develop techniques that will provide information that can be usefully applied by instructors to increase student learning, detect anomalies in homework problems, design the curricula more effectively, predict the approaches that students will take for some types of problems, and provide appropriate advising for students in a timely manner, etc. This introductory chapter provided an overview of the LON-CAPA system, the context in which we are going to use data mining methods.

References:


