Intelligent Learning Systems – An automated data mining approach

Prof. Omprakash L. Mandge
Institute of Computer science
Mumbai Educational Trust
Bandra, Mumbai
Olm.deep@gmail.com

Abstract: ***The advent of communication and technology has led to the widespread use of interactive, personalized and web enabled learning environments in different colleges, universities and institutions, to construct knowledge. The e-Learning systems serve the needs of learners to acquire knowledge and helps teachers to enhance the teaching capabilities. The Learning Management Systems (LMS) can accumulate a large amount of information which can be useful to analyze student's behavior, preferences, likings and disliking. Data mining and web mining techniques can be used to perform such analysis. We have realized that the business objectives are same across institutions and so is the information which is gathered by the learning management systems. This means that the models extracted may differ across different institutions, but the data mining processes are same. This paper studies on implementing "An Intelligent Learning System" in such a way that it provides a more generic approach towards the mining process in Learning Management Systems. The system can be ported to other learning environments to get a faster implementation experience.***

Keywords: Data Mining, Web Mining, E learning, Association rules, KDD

I-INTRODUCTION

Electronic learning ("e - Learning") is emerging as a standard to redesign the entire teaching and learning process with the use of latest technology. e – Learning allows flexible scheduling of learning /teaching process at widely affordable level. As a result many e-Learning systems have been developed and commercialized; these are based on client-server, peer to peer and web based architectures. This has led to the widespread use of Learning Management Systems (LMS) with e- learning capabilities in colleges, universities and institutions, in order to supplement/replace the traditional face-to-face courses [1].

LMS Systems accumulate a vast amount of information which can be used to analyze student's behaviors, their likings, disliking and preferences [2]. They can record all student activities such as reading, writing, taking tests, communication with the teachers, and discussions with other students. However due to the vast of amount of data these systems can generate, it becomes very difficult to analyze this data manually. Data mining techniques can play a very important role in such analysis.

Data mining is the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns for a business advantage [3]. Luan [4] has proposed that in (higher) education, data mining can have an added scientific value in fostering the creation and modification of theories of learning. The entire process of data mining is known as Knowledge Data Discovery (KDD). KDD is usually a very expensive process, especially in determining business objectives, data mining objectives and preparation. Each time data mining is applied to a LMS, many meetings have to be held with the director of the institute, administrators, faculty members etc, to establish the objectives, prepare the data, the mining views and for training the users.

We have realized that, given a business area, in our case learning systems, many data mining implementations repeat the same business objectives, data mining objectives, needs of data, feature construction etc., as compared with previous implementation. Therefore, a lot of work needs to be repeated each time we apply the data mining process. Although the time required in delivering the project is shorter than the first project, most of the work is still manual and hence most of the work involved in previous project is not reused for subsequent projects. We see that models can be very different between different institutions, but the process from data to rules is almost same for every institution.

In this paper we analyze which parts of data mining project for learning management system are equal or highly similar across different institutions. This allows us to design several data mining modules which can be portable across various institutions, thus dramatically reducing the time to implement a data mining program in a new institution. The proposed tool can be implemented on any LMS, but the models to be deployed needs to be retrained.

This paper is organized as follows: In section 2 we explain the data mining virtuous life cycle and propose the structure of automated tool. Section 3 discusses the different business objectives in LMS and conversion of those objectives into data mining objectives. In subsequent sections we concentrate on data collection, transformation and model creation. The final section emphasizes the model evaluation and selection.

II-THE VIRTUOUS LIFE CYCLE OF DATA MINING

The complexity of the data mining process is well captured by CRISP-DM methodology [5]. On the basis of this methodology we have formalized the data mining virtuous life cycle [6] as shown in Fig 1. Mainly the life cycle consists of 1) Translate business opportunity (problem) into DM opportunity (problem) 2) Select appropriate data 3) Get to know the data 4) Create a model set 5) Fix problems with the data 6) Transform data to bring information to the surface 7)Build models 8) Assess models 9) Deploy models 10) Assess results 11) Begin again

The first step in any data mining process is to set up the business objectives. These objectives include: Understanding the student preferences, identifying the most popular courses, improving the performance of the system, improving the utilization of online material, assessing student performance etc. Objectives like these are of major interest for the management of any institution. So these objectives can work as a set of generic objectives for the initial analysis.

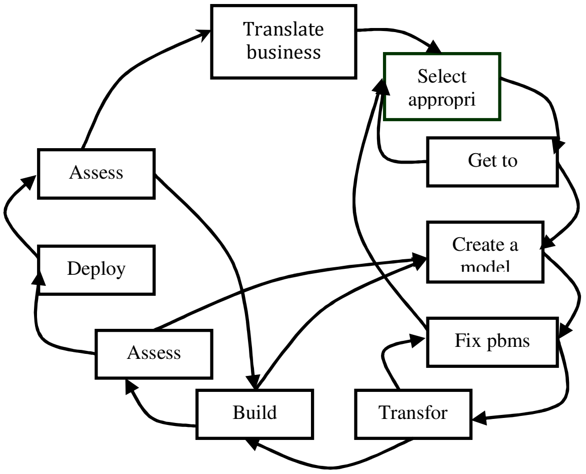

Fig 1 Data mining virtuous cycle

Similarly, the data collected by the institution can be similar such as the student personal data, courses data, teacher's data, chat discussions, test data. The only difference is the format in which the data is stored in the DBMS. This gives a hint for the automation process. We need to characterize only on the data load process in getting the data and transferring it to a data warehouse. Subsequently, the data preparation processes such as grouping, cleansing, attribute selection can be same. Even the models generated for one institute can be utilized for other institutes as well. Considering all these issues we propose the following structure for the automated tool.

The structure comprises all of the steps defined in the virtuous life cycle. The tool is structured in order to carry out the generic operations with ease. The white part shows the processes that can be fully automated and grey part shows the processes that can be semi automated

III-BUSINESS OBJECTIVES

In this section we identify the business objectives in institutes... The administrators, faculty members, users etc in a college plays a major role in deciding the objectives. We study existing learning management systems and consider the inputs from the major players in an institute's management identify the few of the objectives as follows:

- To optimize the performance of the system
- To improve the utilization of resources
- To improve the basic process of learning
- To identify student preferences, liking, disliking
- To evaluate the course material
- To assess student performance

Now these objectives need to be converted in data mining objectives as follows:

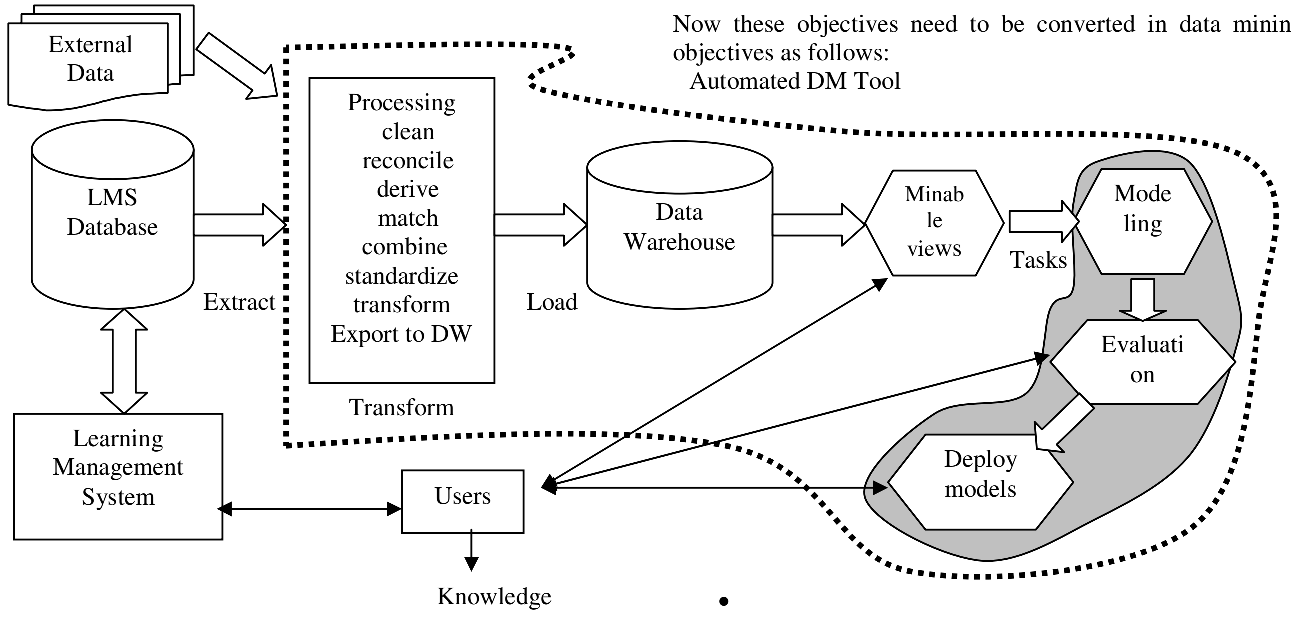

Fig 2 Structure of proposed tool

- To carry out global model to perform analysis of students access to the system at different time periods
- A model to predict the course materials required at different time periods
- A model to analyze each students transaction data to identify their preferences
- A model to perform chat analysis
- A model to analyze and predict the student performance
- A model to find students at high risk
- A model to evaluate contents using a gradation system

IV- DATA COLLECTION

Once the objectives are defined, we need to identify and collect the data to be used. In our case we need two types of data: Internal (collected by the LMS) and external data (such as demographic details of the students, area, culture, attitude of the people towards education etc). The internal data can be collected very easily, but it's very difficult to get the external data. The Generic Structure of an E- learning system is as shown in Fig 3.

Most of the current LMSs normally use a relational database that stores all the systems information: personal information of the users (profile), academic results, the user's interaction data, etc... Databases are more powerful, flexible and bug-prone than the typically textual log files for gathering detailed access and high level usage information from all the services available in the LMS. The LMSs keep detailed logs of all activities that students perform. Not only every click that students make for navigational purposes (low level information) is stored, but also test scores, elapsed time, etc. (high level information).

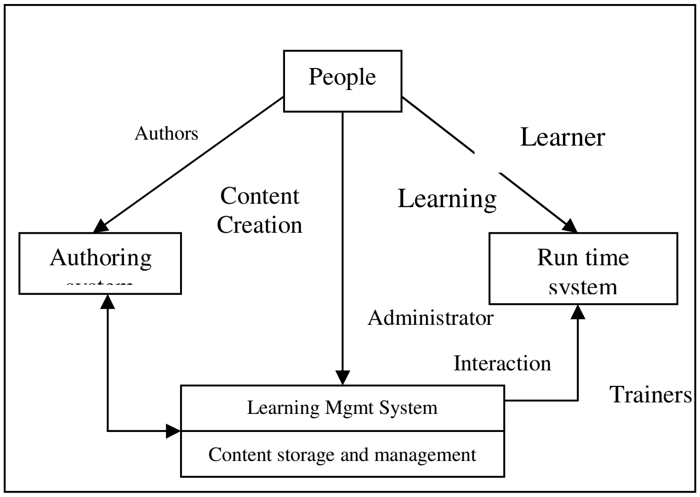

Fig 3 A generic e- Learning System

We collect the data which may contain a number of variables. We identify the variables of our interest such as student, course etc. The next step is to prepare a model set, which will be of interest for the management of any institution. From the collected data set we formalize the structure of data warehouse.

V- DATA PRE-PROCESSING

The next step in the data mining method is the preprocessing of the data collected in the model set. The preprocessing comprises the stages of identifying, collecting, filtering and aggregating data into a format required for the data warehouse. Most of these traditional data pre-processing tasks are not necessary in LMS. Data pre-processing of LMS data is simpler due to the fact that most LMS store the data for analysis purposes, in contrast to the typically observational datasets in data mining, that were generated to support the operational setting and not for analysis in the first place. LMSs also employ a database and user authentication (password protection) which allows identifying the users in the logs. Some typical tasks of the data preparation phase are: data discretization (numerical values are transformed to categorical values), derivation of new attributes and selection of attributes (new attributes are created from the existed ones and only a subset of relevant attributes are chosen), creating summarization tables (these tables integrate all the desired information to be mined at an appropriate level, e.g. student), transforming the data format (to format required by the used data mining algorithms or frameworks).

Some processes for extraction of data are adapted for a particular LMS, but the other many pre-processing operations are similar for every LMS. This part of preparation process can be reused from one LMS to another LMS, through automation of all common processes in data preparation module.

We define scripts for extracting data from different LMSs into the Data Warehouse. These scripts must be slightly different from LMS to LMS. From the DW, since the data definition (multidimensional schema) is the same for every LMS, we used SQL scripts to generate the minable views, which are exactly the same. All these complex queries are highly time-consuming. With our approach, these queries are portable from one LMS DW to another, and all this effort is reused.

For example, to identify the student preferences we collect a data set with the structure given in fig 4. Using this data model we identify the attributes of our interest to prepare a common data format for the data warehouse. In order to identify the preferences, we consider the format of material accessed by the students and compare it with the performance of the student. The SQL script for the same can be as follows.

```
Select format,count(*)
From topic_info, access_by
Where access_by.topic_id= topic_info.topic_id
And access_by.student_id  IN( select student_id
            From student_res where student_grade > 3)
Group by format;
```

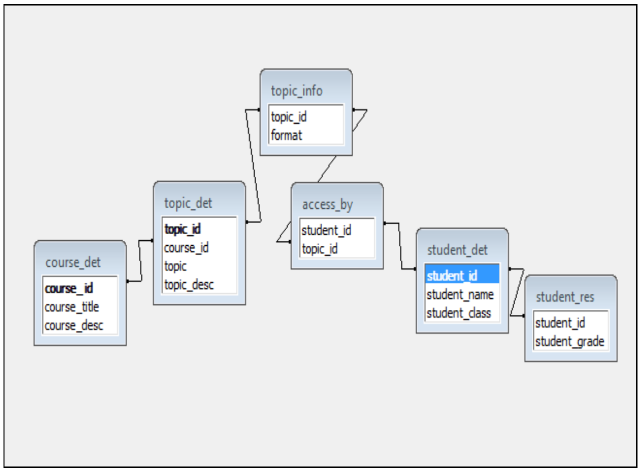

Fig 4 Sample Model Set

We need to define more such scripts to prepare the data warehouse. Once the Data Warehouse is prepared, we generate different minable views of the DW using DM models as described in the next section.

VI - LEARNING DATA MINING ALGORITHMS AND MODELS

Once the data have been properly filtered, cleaned and transformed, we can proceed with the induction of the different models for analysis and prediction in order to generate knowledge. We need to generate different minable views from the DW and retain the best of minable views to achieve the business objectives defined in section 3. As we have discussed earlier LMS datasets are comparatively small, but the number transactions can be quite large depending on how much information LMS stores about the interaction of the student with the system. In this paper we explain how the a) classification and b) association rule mining techniques can be applied to LMS.

a) We show the minable view generated to identify the student preferences. To generate the minable views we consider: the different course, topics in the course, various formats of material (audio, text, hypertext, video, diagrams, short notes) etc.

Table 1 Minable view to evaluate the course contents

Attribute	**Description**
Student ID	Students ID
Topic	Name of the topic
course	Course to which topic belongs
numText	Number of times text contents were accessed
numshort notes	Number of times short notes were accessed
numvideos	Number of times videos were accessed
numaudio	Number of times audio files were accessed
numdiagrams	Number of times diagrams were accessed
Numhypertext	Number of times hypertext material was accessed
numAccess(class)	Number of times the topic was accessed

We prepared this minable view considering 200 students; five courses each including 5 topics. So, this view contains 200*5*5=5000 rows. With this initial minable view, we use different learning methods such as LinearRegression, LeastMedSq, SMOreg, MultilayerPercepton, K- DBC, LWL, Tree DecissionStump, Tree M5P and IBK and identify the best methods in order to prepare decision model.

b) An association rule X ⇒ Y expresses that in those transactions in the database where X occurs; there is a high probability of having Y as well. X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by its support and confidence. The confidence of the rule is the percentage of transactions with X in the database that contain the consequent Y also. The support of the rule is the percentage of transactions in the database that contain both the antecedent and the consequent.

Association rule mining has been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset), including, e.g., the following tasks: building recommender agents for on-line learning activities or shortcuts , automatically guiding the learner's activities and intelligently generate and recommend learning materials , identifying attributes characterizing patterns of performance differences between various groups of students , discovering interesting relationships from student's usage information in order to provide feedback to course author, finding out the relationships between each pattern of learner's behavior , finding students' mistakes that are often occurring together , guiding the search for best fitting transfer model of student learning , optimizing the content of an e-learning portal by determining the content of most interest to the user , extracting useful patterns to help educators and web masters evaluating and interpreting on-line course activities , and personalizing e-learning based on aggregate usage profiles and a domain ontology . All of these analytics can be included in the proposed tool. Table 2 lists the attributes common to a variety of e- Learning Systems.

Table 2 Examples of attributes common to a variety of e-learning systems.

Attribute	**Description**
Visited	If the unit, document or web page has been visited
Total_time	Time taken by the student to complete

	the unit
Score	Average final score for the unit
Difficulty_ level	Number of attempts before passing the unit
Chat	Number of messages sent/read in the chat room
Forum_messages	Number of messages sent/read in the forum
Accounting	Number of courses booked
Number of visits	Number visits to the document
Ontology	Number of semantic searches performed

The obtained results or rules are interpreted, evaluated and used by the teacher for further actions. The final objective is to putting the results into use. Teachers use the discovered information (in form of if-then rules) for making decisions about the students and the LMS activities of the course in order to improve the student's learning. So, data mining algorithms have to express the output in a comprehensible format.

The final goal, however, is not to solve the management problems of a single institution, but to port these results to other institutions. According to our study, we need to extract which minable views, which attribute selection, which learning methods and association rule mining algorithms are best from these data. Consequently, we have to implement the best minable views and models in the automated system.

VII- Conclusion

It is still early days for the total automation of data mining processes in e learning systems and not many such real implementations are available. In this paper, we have analyzed the adequacy of designing specialized modules for data mining for Learning management system and we have also identified which are the stages in the KDD process which could be reused and automated across different LMSs. The design of such an automated tool could turn data mining into an available technology to many institutions at technical and economic affordable level.

Further, we propose the design of the tool in such a way that, it hides the details of the data mining processes from the normal user. The use of proposed tool could avoid repetition of efforts and could provide a great expertise, knowledge with ease.

The future work in this area could include extension of the modules to define or modify new data, add new models, add new business objectives etc.

References

1 Rice, W.H.: Moodle E-learning Course Development. A complete guide to successful learning using Moodle. Packt publishing (2006).

2 Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., Heiner, C.: An educational data mining tool to browse tutor-student interactions: Time will tell! In: Proc. of the Workshop on Educational Data Mining (2005) 15–22.

3 SAS Institute: http://www.sas.com/feature/4qdm/whatisdm.html

4 J. Luan. Data mining and knowledge management in higher education: Potential applications. In Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada, June 2002

5 CRISP-DM, www.crispdm.org

6 Mastering Data Mining, Ron Norman

7 Data Mining Techniques, Ralph Kimball

8 Agawam R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. of SIGMOD (1993) 207-216.

9 Sahami, M.: Learning Limited Dependence Bayesian Classifiers. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96 AAAI Press (1996) 335–338

10 Castillo, G., Gama, J., Breda, A.M.: An Adaptive Predictive Model for Student Modeling.Advances in Web-based Education: Personalized Learning Environments, (2005) Chapter IV

11 Sahami, M.: Learning Limited Dependence Bayesian Classifiers. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96 AAAI Press (1996) 335–338

12 Mitchell, T.: Machine Learning. McGraw Hill (1997)

13 Carver, C.A., Howard, R.A., Lane, W.D.: Enhancing Student Learning Trough

Hypermedia Courseware and Incorporation of Student Learning Styles, IEEE

Transactions on Education, v.42, nº 1 (1999) 33-38

14 Felder, R.M., Silverman, L.K.: Learning and Teaching Styles in Engineering Education, Engr. Education, 78 (7), (1988) 674-681

15 "Discovering Student Preferences in E-Learning."C. Carmona, G. Castillo, E. MillánInternational Workshop on Applying Data Mining in e-Learning (ADML'07) as part of the Seco-nd European Conference on Technology Enhanced Learning (EC-TEL07)