International Journal of Computer Applications in Engineering Sciences
[VOL I, ISSUE III, SEPTEMBER 2011] [ISSN: 2231-4946]

On-Line Data Analytics

Yugandhar Vemulapalli[#], Devarapalli Raghu[*], Raja Jacob[#]

Dept. of CSE , Nova College of Engineering & Technology, Jangareddy Gudam.
Affiliated to JNTU Kakinada

Yugandhar1@gmail.com, raguau@gmail.com, rchidipi@gmail.com

***Abstract*— This paper presents an approach and a system to let tutors monitor several important aspects related to online tests, such as learner behavior and test quality. The approach includes the logging of important data related to learner interaction with the system during the execution of online tests and exploits data visualization to highlight information useful to let tutors review and improve the whole assessment process. This paper has focused on the discovery of behavioral patterns of learners and conceptual relationships among test items. For this Characterization and summarization has been used. The Characterization and summarization is implemented efficiently using Attribute Oriented Induction algorithm which discovers patterns for accessing learners behavior. By analyzing the data visualization charts, we have detected several previously unknown test strategies used by the learners. Last, we have detected several correlations among questions, which gave us useful feedbacks on the test quality.**

***Keywords*— Distance learning, interactive data exploration and knowledge discovery, Data Mining, Online tests, Data Collection, Data Visualization, Characterization, Summarization and Attribute Oriented Induction.**

I. Introduction:

In today's academic environments the tutors are playing vital role that they not only plays the role of teacher but also should play the role of guide and mentor. The tutors in corporate training and academic environments are assessing learner's ability and skills and accordingly he provides grading for their skills. Now based on these grading he suggested improvements to increase the learners learning capability, thinking ability and knowledge base. For this the tutor continuously conducts various types of tests. E-TESTING systems are being widely adopted in academic environments, as well as in combination with other assessment means, providing tutors with powerful tools to submit different types of tests in order to assess learners 'knowledge. Among these, multiple-choice tests are extremely popular, since they can be automatically corrected. However, many learners do not welcome this type of test, because often, it does not let them properly express their capacity, due to the characteristics of multiple-choice questions of being "closed-ended." Even many examiners doubt about the real effectiveness of structured tests in assessing learners' knowledge, and they wonder whether learners are more conditioned by the question type than by its actual difficulty. In order to teach learners how to improve their performances on structured tests, in the past, several experiments have been carried out to track learners' behavior during tests by using the think-out-loud method: learners were informed of the experiment and had to speak during the test to explain what they were thinking, while an operator was storing their words using a tape recorder. This technique might be quite invasive, since it requires learners to modify their behavior in order to record the information to analyze [1], [2], [3], [4], [5], which might vanish the experiment goals, since it adds considerable noise in the tracked data. Thus knowledge discovery (KDD) process is the main theme of the Thesis. The knowledge discovery strategies are used to extract knowledge from raw data [1][2]. Here raw data is nothing but large collection of data and knowledge is nothing but required small amount of data to be used in analysis. The following are various steps in KDD process.

In our Thesisthe KDD process is implemented using two phases

a. Data Collection
b. Data Visualization

A. Data Collection

It is a process of collecting or gathering learner's activities and data during online tests. For this we use think-out-loud method. The following kind of information is collected.

- Duration of the visit
- Presence and duration of inactivity time intervals during the visit
- Sequence of responses given by the learner during the visit
- Estimation of the time spent by the learner in evaluating the stem (question) and each of the options for the question.

B. Data Visualization

This is used to present the analysis data in different forms such as curves, charts, lines , pie charts ,bar charts , circle , comparison lines so on. The above may be 2D or 3D. In above presentations learner's data

may be skills, abilities and behavior. Data visualization [7] provides a graphical representation of data, documents, and structures, which turns out to be useful for various purposes. Data visualization provides an overview of complex and large data sets, shows a summary of the data, and helps human in the identification of possible patterns and structures in the data[2][3][4][7]. Thus the goal of data visualization is to simplify the representation of a given data set, minimizing the loss of information.

Visualization methods [7] can be either geometric or symbolic. In a geometric visualization, data are represented by using lines, surfaces, or volumes and are usually obtained from physical model or as a result of a simulation or a generic computation. Symbolic visualization represents non-numeric data using pixels, icons, arrays, or graphs.

The following diagram shows different visualization techniques.

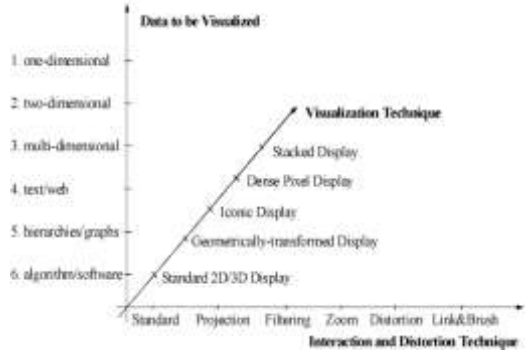

Figure 1

B.1 Problem Statement

B.2Problem Definition

To provide Index a solution that enables 'recording of learner' habits during online tests without informing them of the underlying experiment and without asking them to modify their behavior.

I.2.3 Detailed Problem Description.

This Thesis aims to present a solution enabling the recording of learner's habits during online tests without informing them of the underlying experiment without asking them to modify their behavior which potentially yields more realistic results. The Thesis deals monitoring several important aspects related online tests, such as learner behavior and test quality. The approach includes the logging of important data related to learner interaction with the system during the execution of online tests and exploits data visualization to highlight information useful to let tutors review and improve the whole assessment process. This focused on the discovery of behavioral patterns of learners of learners and conceptual relationships among test items. In Particular, by analyzing the Data visualization charts, detecting several previously unknown test strategies used by the learners. Last, this detects several correlations among Questions, which gave us useful feedbacks on the test quality. .

C. Thesis Outline:

The major contribution of this work is as follows:

- It gives an insight into the problem of identifying user behavior during the online test through visualization of user interactions with test patterns.
- It focuses on visualization where the user is directly involved in the data mining process.

.

This paper is organized as follows: Section 2 completely describes System Design and Implementation. Section 3 describes Proposed Approach. Section 4 describes Experimental Results. Section 5 presents conclusion and future scope.

II. System Design and Implementation:

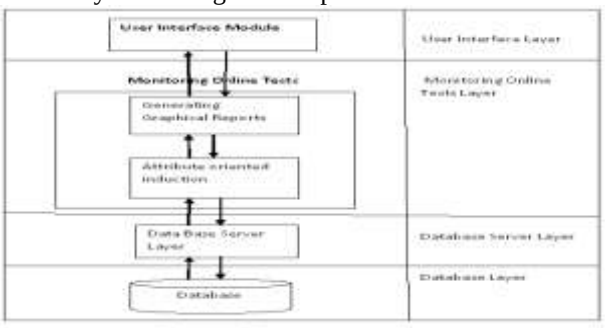

Figure 2 System Design

The System is composed of 4 layers.

A) The User Interface Layer:

It is responsible for interaction with the user and various calls to various graphical and visualization utilities. This Module provides an Interface for each user to invoke with the system and to execute Attribute Oriented Induction Algorithm for performing Data Characterization and to generate graphical reports for accessing learner's behavior.

B) Monitoring Online Tests

In this layer the Data Characterization and Summarization is performed using Attribute Oriented Index algorithm (AOI) by reading Learner's exam related data from Database.

C) Data Base Server Layer: SQL Server relational data base.Contains data base of transactional data items. Receives Queries from the user interface design and transfers results from transactional data base.

D) Data Base Layer: Currently available in the proposed system are Learner's Exam related data.

III. Proposed Approach

In this section, we describe the approach to discover knowledge related to learner activities during online tests, which can be used by tutors to produce new test strategies. In particular, we have devised a new symbolic data visualization strategy with help of Attribute Oriented Induction algorithm, which is used within a KDD process to graphically highlight behavioral patterns and other previously unknown aspects related to the learners' activity in online tests.

In this section, we describe the system implementing the proposed approach. The system is composed of a Logging Framework and a Log Analyzer application. The former, based on the AJAX technology, captures and logs all of the learners' interactions with the e-testing system interface (running in the Web browser). It can be instantiated in any-testing system and is further composed of a client-side and server-side module. The latter is a stand-alone application that analyzes the logs in order to extract information from them and to graphically represent it.

A. The Logging Framework:

The purpose of the Logging Framework is to gather all of the learner actions while browsing web pages of the test and to store raw information in a set of log files in XML format.

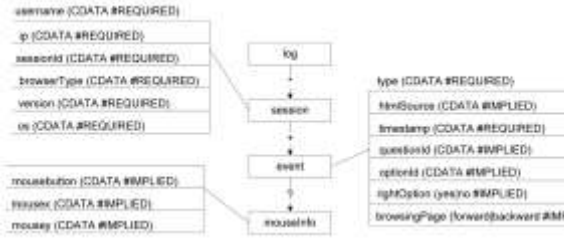

Fig. 3. The information Model for log data.

The framework is composed of a client-side and a server side module. The former module is responsible for "being aware" of the behavior of the learner while s/he is browsing the test's pages and for sending information related to the captured events to the server-side module. The latter receives the data from the client and creates and stores log files to the disk. Despite the required interactivity level, due to the availability of AJAX, it has been possible to implement the client-side module of our framework without developing plug-ins or external modules for Web browsers. JavaScript

has been used on the client to capture learner interactions, and the communication between the client and the server has been implemented through AJAX method calls. The client-side scripts are added to the e-testing system pages with little effort by the programmer. The event data is gathered on the Web browser and sent to the server at regular intervals. It is worth noting that the presence of the JavaScript modules for capturing events does not prevent other scripts loaded in the page to run properly. The server-side module has been implemented as a JavaServlet, which receives the data from the client and keeps them in an XML document that is written to the disk when the learner submits the test. The Logging Framework can be instantiated in the e-testing system and then enabled through the configuration.

B. Attribute Oriented Induction Algorithm:

In [Cai, Cercone 8z Han 1991, Han, Cai & Cercone 1992,Han, Cai & Cercone 19931, an attribute-oriented induction method for data-driven discovery of quantitative rules in relational databases is presented. It uses domain

Knowledge to generate descriptions for predefined subsets of a relational database. This attribute-oriented approach uses the concept hierarchy to direct the learning process.

From: KDD-95 Proceedings. A concept hierarchy is related to a specific attribute and is

Partially ordered according to a general-to-specific ordering. The most general point in the hierarchy is the null description (ANY), while the most specific points correspond to the specific values of an attribute in the database.

For example, assume that a university student database has the following schema. Student (Name, Status, Sex, Age, GPA)

1, the concept tree of the attribute Status will be the one

(Freshman) + Undergraduate (sophomore) + undergraduate
(Junior) + Undergraduate (senior) + undergraduate
(M.A.) + graduate (M.S.) + Graduate
(Ph.D.) + graduate
(Undergraduate, graduate) + ANY (Status)
(0.0-l .99) + Poor
(3.0-3.49) + good
(Poor, average) + Weak
(Weak, strong) + ANY (GPA)
(M, F)--+ ANY (Sex)
(16-25) + 16-25
(16-25, 26-30) + ANY(Age)
(2.0-2.99) + average
(3.5-4.0) + excellent
(Good, excellent) + Strong
(26-30) + 26-30

Learning results is called an intermediate (final)

Generalized relation. In a generalized relation, some or all of its attribute values are generalized data, that is, non-leaf nodes in the concept hierarchies. An attribute in a(generalized) relation is at a desirable level if it contains at most a small number of distinct values in the relation. This small number is specified by the user

as a desirable attribute threshold. A set of basic principles for the attribute-oriented induction in relational databases is summarized as follows.

1. Generalization should be performed only on the set of

data which is relevant to the learning task.

2. Generalization should be performed on the smallest

Attribute removal: If an attribute has too many

distinctive values and there is no higher level concept provided for further generalization, it should be removed from the relation. Concept tree ascension: For an attribute in an intermediate relation, if its values can be generalized to higher level concepts in the concept tree of the attribute, all values of the attributes are replaced by the higher level concepts. Outcome of the ascension is a generalized relation. Vote propagation: Vote of a generalized tuple indicates the number of tuples in the initial relation that are generalized to this tuple. The value of the vote of a tuple is carried to its generalized tuple and the votes should be accumulated when merging tuples.

Attribute threshold control: For an attribute, if the number of its distinct values in an intermediate relation is still larger than its desirable attribute threshold, further generalization on this attribute should be performed. By applying the above principles, an initial relation would be reduced to a generalized relation call prime relation. This prime relation has a small number of distinct values (less than or equal to the attribute threshold). This prime relation may need to be generalized further to produce the final relation. Two additional principles are used to complete the Attribute-Oriented induction process.

1. Generalization threshold control: If the number of tuples in a generalized relation is larger than the generalization relation threshold, further generalization should be performed.

2. Rule formation: A tuple in the final relation is

Transformed to conjunctive normal form, and multiple tuples are transformed to disjunctive normal form.

IV. Experimental Results

In order to demonstrate the effectiveness of the approach and of the proposed system, we have used them in the context of the Web Development Technologies course in our faculty: the eWorkbook system, equipped with the new module for tracking the learners' interactions, has been used to administer an online test to learners. They have not been informed of the experiment; they just knew that the grade obtained on the tests concurred to determine the final grade of the course exam. The test, containing a set of25 items to be completed in a maximum of 20 minutes, was administered to 71 learners, who took it concurrently in the same laboratory. The assessment strategy did not prescribe penalties for incorrect responses, and the learners were aware of that. The logger was enabled, and an approximately 4-Mbyte-sized XML log file was produced. The logging activity produced no visible system performance degradation. Then, the Log Analyzer has been used for analyzing the logs in order to extract information from them and to graphically represent it in order to trigger a visual data mining process where the tutor plays a central role. In the case of the mentioned experiments, the visual analysis of charts enabled a tutor to infer interesting conclusions about both the strategies the learners used to complete tests and the correlation between questions. In the former case, the objective was not only to understand the learners' strategies but also to detect the most successful of them under the circumstances explained above and, possibly, to provide learners with advice on how to perform better next time. On the other hand, question correlation analysis aims to improve the final quality of the test by avoiding the composition of tests with related questions.

The following shows different experimental results.

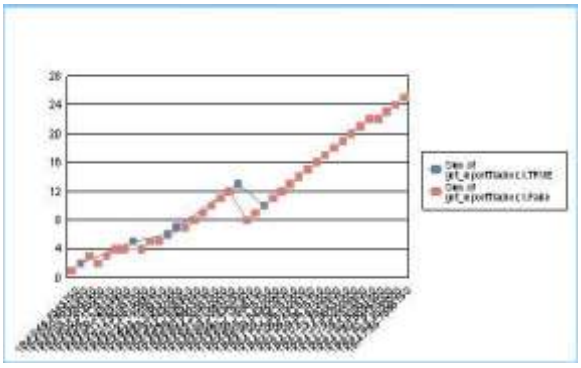

Figure4 Learner Behavior in Active State

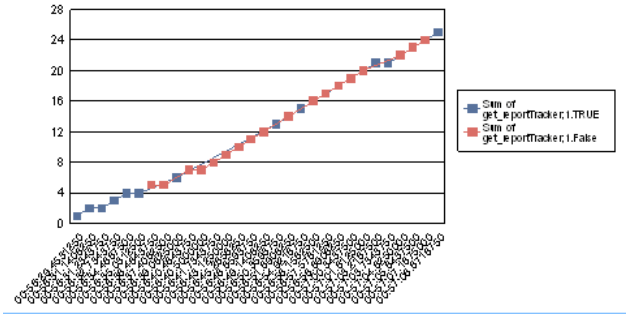

Figure5 Learner Behaviors in Passive State

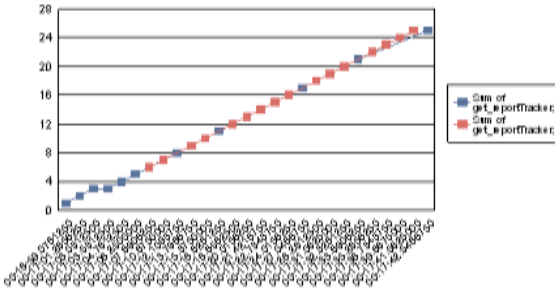

Figure6Learner Behavior in Single State

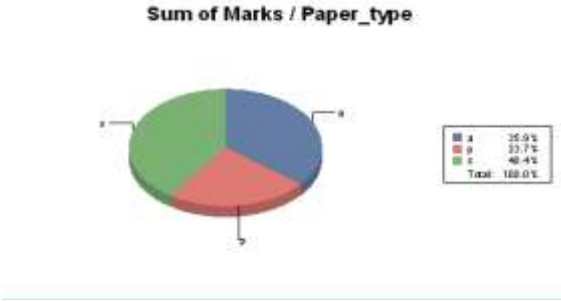

Figure7 phase average pie graph

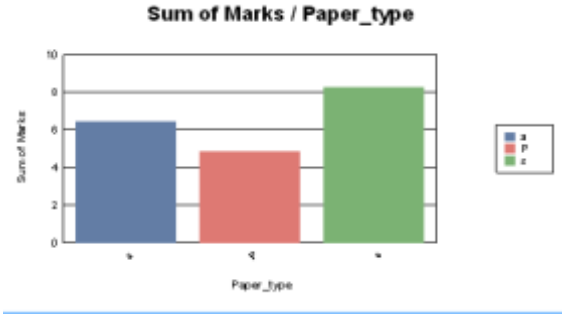

Figure8Phase Average Graphs

UserName	R2W	W2R	NOC	Score
08a81d5812	0	0	0	8
08a81d5813	2	0	2	8
08a81d5814	1	0	1	11
08a81d5815	1	0	1	0
08a81d5816	0	1	1	4
08a81d5817	0	0	0	12

Figure 9: Answer changing behavior of learners.

V. Conclusions and Future Work

It has been presented an approach and a system to let tutors monitor learners' strategies during online tests.

The approach exploits data visualization to draw the data characterizing the learner's test strategy, in order to trigger the tutor's attention and to let him/her discover previously unknown behavioral patterns of the learners and conceptual relationships among test items.

The tutor is provided with a powerful tool that lets him/her review the whole assessment process and evaluate possible improvements.

We have extensively used the implemented system experimentally to evaluate online test strategies in the courses of our faculty, in order to assess the whole approach.

This lets us discover several relevant patterns regarding the test quality, the characteristics of used strategies, and the impact on the final score.

The cheating behavior of the learner can also be visualized by tracking the mouse movements of the learner.

It is presently implemented online exams in universities but they are not represented the users behavior visually. But it should be implemented in future.

REFERENCES

[1] J. Bath, "Answer-Changing Behavior on Objective Examinations,"J. Educational Research, no. 61, pp. 105-107, 1967.

[2] J.B. Best, "Item Difficulty and Answer Changing," Teaching of Psychology, vol. 6, no. 4, pp. 228-240, 1979.

[3] Johnston, "Exam Taking Speed and Grades," Teaching of Psychology, no. 4, pp. 148-149, 1977.

[4] McClain, "Behaviour during Examinations: A Comparison of 'A,' 'C,' and 'F' Students," Teaching of Psychology, vol. 10, no. 2, pp. 69-71, 1983.

[5] C. Plaisant and B. Shneiderman, "Show Me! Guidelines for Producing Recorded Demonstrations,"pp. 171-178, 2005.

[6] D.A. Keim, "Information Visualization and Visual Data Mining,"vol. 8, no. 1, pp. 1-8,Jan.-Mar. 2002.

[7] G. Costagliola, F. Ferrucci, V. Fuccella, and F.Gioviale, "A Web Based Tool for Assessment and Self-Assessment,"pp. 131-135, 2004.