Intelligent System for Personalizing Students’ Academic Behaviors- A Conceptual Framework

Azwa Abdul Aziz¹, Wan Mohd Rizhan Wan Idris¹, Hasni Hassan¹, Julaily Aida Jusoh¹

¹Faculty of Informatics, Universiti Sultan Zainal Abidin (UniSZA), Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia
{azwaaziz, wmrizhan, hasni, julaily}@unisza.edu.my

Abstract. Business Intelligence (BI) refers to skills, processes, technologies, applications and practices used to support decision making. BI technologies provide historical, current, and predictive views of business operations which are normally used to analyze business data. Online Analytical Processing (OLAP) is one of the common BI approaches in answering multi-dimensional analytical queries for analytical purpose. In addition, Educational Data Mining (EDM) is an emergent discipline for exploring data, and a method to support learning and teaching processes. In this paper, we proposed Educational Intelligence (EI) Framework by combining BI technologies with various EDM algorithm techniques. Taking UniSZA as our case study, the patterns on students’ academic behaviors and performance can be analyzed. A set of data from students’ examination results in relational database is extracted into multi-dimensional model to support OLAP query processing. The results are grouped into several subject areas. Then, the analysis to recognize the patterns on students’ academic behaviors is conducted using EDM algorithms. From the analysis, the groups of students who have excellent skills or vice versa can be identified. It also optimizes the time to perform current and historical data analysis. The weaknesses and strengths of the student can also be obtained. Finally, students’ future potential areas of studies can be predicted using the framework.

Keywords: Educational Intelligence, Educational Data Mining, Educational Data Warehouse

1 Introduction

Data Warehouse (DW) is databases devoted to analytical processing. They are used to support decision-making activities in most modern business settings, when complex data sets have to be studied and analyzed [1]. The processes of DW involve taking data from the legacy system together with corresponding transactions of the system’s database and transforming the data into organized information in a user friendly format. One of the
main components in DW architecture is Business Intelligence (BI). BI refers to skills, processes, technologies, applications and practices which are used to support decision making [2]. BI software leverages on investment in data and systems to provide easy-to-use information that improves decision-making in an organization. Online Analytical Processing (OLAP) is one of the main functions offered by BI tools to support multi-dimensional model of DW. Educational Data Warehouse (EDW) refers to a DW created to support Academic Data.

Most of DW projects have been implemented in business sector whereby information is crucial in ensuring the companies’ growth and in maximizing their profits. Therefore, the term ‘business’ refers to the BI tools such as Statistical Analysis Software (SAS), Business Object and Cognos. The problem is: Can these BI tools support educational sector with massive and rapid changes of data transactions? Can the multi-dimensional model support educational data in universities that normally contains high volumes of data? How can we utilize the benefits of BI tools in education data? Indeed, these questions become the motivation to this research. In this paper, BI technologies to be utilized in education for personalizing the students’ academic data based on their results and assessments are discussed in detail. It is anticipated that using this technique, we would be able to identify students’ levels of strengths and weaknesses in specific subject areas. Besides, from the study, the factors that influence the pattern of students’ results can also be obtained.

University Sultan Zainal Abidin (UniSZA) is one of the public Institutions of Higher Learning (IHL) in Malaysia. To date, the university has nine faculties including Faculty of Informatics (FiT) which becomes the case study for this research. FiT offers two diploma programs that are Diploma in Information Technology (Multimedia) (DITM) and Diploma in Information Technology (DIT). Data of students from existing transaction systems or any external sources has been converted into the multi-dimensional model in order to be used in OLAP analysis. The subjects of each program have been grouped into five main fields of computer science such as networking, multimedia, software engineering, programming and databases. The pattern of students’ results for each semester can be analyzed using several data mining techniques. This pattern is important to determine the students’ potentials on the subject area and to predict the best programs for the next level of students’ educations. Furthermore, a list of factors that influence student academic performance can also be obtained.

This study is important to determine whether BI technology is appropriate to be used in dealing with educational data. Previous researches show that data mining has been widely used in business area and scarce in the area of education. It is because the data in educational field is very dynamic; the data can also be much different between the samples; and teachers or lecturers cannot afford the time and do not have the expertise to do the tests on each sample especially in real time [3]. Another reason is in terms of the size of the data. If tremendous amount of data are collected, the size of the data on one sample usually becomes bigger. Therefore, this study only focuses on personalizing students’ behaviors based on the results for each semester and several selected factors in the determination of students’ achievements.
2 Related Research

Prior researches have focused more on the implementation of data mining techniques and models to discover educational data. Educational Data Mining (EDM) is an emerging discipline for developing methods to explore unique types of data from educational context [4]. In fact, EDM is an application of data mining techniques implemented in the area of education for obtaining better comprehension on students’ learning processes and acknowledging the ways they participate in it, in order to improve the quality of the educational system [5].

Most of the EDM researches have utilized e-learning environments to implement the data mining algorithm [5, 6, 7, 8, 3]. It is because the collected data is more manageable and accurate since the online systems can automatically record activities performed by the students. However, there are still studies conducted to analyze the educational data in the traditional learning environment [3]. Merceron & Yasef [3] have used the association rules to analyze the learning data. They have also conducted educational data mining case study for Logic-ITA student data [6]. Zorrilla et al. [5] have proposed a decision making system to help instructors study the pattern of student data in e-learning environment. The system used Learning Content Management System (LCMS) databases to request a pattern and interpret the results. Talavera & Gaudiso [7] have used data mining model to characterize similar behavior groups on the unstructured collaborations spaces while Hong & Zhang [8] have identified the behavior patterns of 98 undergraduate students in business course in Taiwan through online learning processes.

On the other hand, the use of BI methods for educational data is still not widely implemented. For instance, Janson et. al. [10] had proposed a conceptual framework for DW approaches for Smart School System. The study used a general method to apply DW systems into educational perceptive and did not analyze the students’ performance behaviors or classifying students’ patterns but it has implemented a normal DW architecture to improve school administrative data.

In this paper, the benefit of BI in analyzing students’ achievements and performance is utilized and discussed in detailed. It involves the studies on mining students’ behaviors to obtain individual patterns of students’ academic behavior. Consequently, the results could be used by the lecturers as an aid to enhance the teaching process in order to improve students’ academic achievements.

3 Business Intelligence vs. Educational Intelligence

BI refers to skills, processes, technologies, applications and practices used to support decision making. BI software leverages on investment in data and systems to provide easy-to-use information that improves decision-making in an organization [19]. They are many BI tools that provide several applications to suit certain purposes. Some applications produce an Executive Reporting System (ERS) while others require dashboard, graphical
reports and even more powerful analytic programs that allow user to have more manipulations and flexibility over a set of data from DW. Most BI tools such as SAS, Micro strategy, Cognos and Business Object have the abilities to do the processes mentioned above and are still enhancing their technology to meet market demands. An effective architecture must provide a full range of BI capabilities to solve real business problems across the organization without creating new ones [19]. The Fortune 1000 and many organizations use BI as part of a growing move towards integrating organizational processes for higher performance [20].

Common functions of BI technologies are reporting, dashboard, OLAP, analytic, data mining, business performance management, benchmarking, text mining, and predictive analytics. BI often aims to support better business decision-making. Dashboard is multi-layered performance management systems, built on BI and data integration infrastructures, enable organizations to measure, monitor and manage business activity using both financial and non-financial measures. Dashboards tend to monitor the performance of operational processes, whereas scoreboards tend to chart the progress of tactical and strategic goals. Dashboards are often used to display and monitor Key Performances Indicator (KPI). Another main feature of BI technologies is the reporting tool. The reporting provided by BI tools is usually able to customize to the customers’ needs. Users are able to drill down the reports, chose the type of reports that need to be produced and automatically generate reports. In summary, BI application is used to analyze business data, provide descriptive and predictive analysis and perform forecasting with the aim to help the management to make decision.

In this project, we are trying to apply the same concept for educational data. Applying multidimensional model for educational data might be an answer to solve high volume and rapid changes in educational data. As a result, the query processes become faster. It is hoped that the term “Educational Intelligence (EI)” proposed here would serve as a guidelines for EI criteria. Several features that EI needs to have are:

i. The ability of various data from different sources to be integrated in one large pool of database/Educational Data Warehouse (EDW).

The implementation of EDW will improve the process of handling large volume of data and rapid changes in educational data. Designing a physical design of database in multidimensional format is improving the query taken for producing reports. Furthermore, applying cube for OLAP engine provides efficiency and flexibility in educational data reports.

ii. The ability to perform multiple analytic processes whether descriptive analysis or predictive analysis

The main importance of EDM is the ability to provide comprehensive analysis on students’ data. Lecturers or teachers may understand students’ individual behavior and could devise strategies to improve their achievements in learning
process. The analysis provided by EI whether descriptive or predictive will help improving learning and teaching process. In traditional ways, it is difficult for educators to know each of their students. It is because some of them are having more than 100 students at certain times. Therefore, the ability to produce meaningful reports is important in learning process.

iii. A user-friendly Graphical User Interface (GUI) for EI reports

Not all teachers are a computer literate but usually most of them have basic knowledge on computers. In order to ensure EI technologies will be supported by teachers, a simple and efficient GUI needs to be provided by EI reports. Teachers can customize and use the reports with minimal technical assistance.

4 Educational Intelligence Conceptual Framework

DW is a collection of data, physically separated from the operational environment that solves real business pain, and contains data of interests to multiple user groups. It comprises data stored within well-defined period of time, thus enabling a trend analysis to be carried out. Normally, DW is divided into three main layers in which data from one layer is derived from the lower layer [11]. The layers are data sources, staging area/data repository and presentation.

Data sources layer is the transaction data used for daily or basic operations to support an organization. In educational perspectives, the data sources may come from online transaction system (OLTP) such as students’ academic system, attendance system, course registration system or students’ enrollment system. It also may come from unstructured data such as excel records of students’ assignments. The data repository layer is DW repository, the core of DW system architecture. It contains valuable data used for analytical purposes and data mining. It is designed in a star schema model which has dimension and fact tables [12]. Fact tables contain numeric values that present a measurement of business rules. Dimension tables are attributes of the fact tables. Presentation layer offers various functions and technologies in BI tools to manipulate the data.

3.1 EIDesign

For EI physical architecture framework, we proposed two main layers that are slightly different with normal DW architecture. The two main layers are Front End layer and Back End layer as shown in Fig. 1.
The Front End Layer serves as a presentation layer or data access layer. It can be divided into front end for data sources and front end for analysis and report. Data sources are any systems or procedures for providing educational data. It involves systems that support educational operation such as students’ registration system or students’ assessment system. End users especially the lecturers will deal only with the analysis and reporting where BI and query tools are used to process the data. Users do not need to know the technical parts of systems. The process of preparing data will be managed by the Back End Layer.

The Back End Layer is where data are being processed and stored in the systems. It includes a data repository area, data staging area and cleansing process that are hidden from front end users. Mining algorithm to get the reports will be applied here. This layer is the backbone for every EI application to support the front end processes. Technical staffs such as system administrator, DW designer, and ETL developer will support the layer with the requirements obtained from Front End Layer users.

EDW is the main component in EI architecture. Fig. 2 shows the architecture of the proposed EDW for UniSZA case study. This architecture is designed to personalize students’ academic performance. The data for analyzing the students’ performance may
come from structured or semi-structured data. Structured data is aggregation or generalization of items described by elementary attributes defined within a domain. Relational tables and statistical data represent the most common type of structured data [13]. It is also well organized data stored in database management system (DBMS) such as students’ record system, students’ academic system or students’ enrolment system. In UniSZA, the Information Technology Centre is responsible to handle all application systems. In order to obtain the data, a set of procedures needs to be followed in order to preserve the integrity and confidentiality of the data.

**Fig. 2. EDW Architecture for UniSZA**

The semi-structured data is the type of data that has a structure with some degrees of flexibility [13]. The data may come from XML documents or text files. For example, lecturers may store the details of students’ assessment in their own files with different formats. One of the methods for collecting semi-structured data is to provide a standard format to store the data in text files.

The selected data will be stored in staging area. The data staging area is everything between the operational source systems and the physical storage of DW. Once the data is extracted to the specified staging area, numerous potential transformations can be done such as cleansing of the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, duplicating the data, and assigning warehouse keys [11]. Extract, Transform, Load (ETL) tools will be used to perform those processes/transformations. Once done, the data can be extracted from structured data (relational, statistical) to multi-dimensional model. For example, the relational tables which contain several normalized geographical information entities such as region, country, states, district, and location can be converted into geographical dimension. The primary key of the lowest group hierarchy (in this case location table) becomes the primary key for the dimension tables as shown in Fig. 3.
3.2 EDW Multidimensional Model

Operational databases offer an integrated view of the application, so that each entity of the real world is mapped into exactly a single concept of the schema [12]. The schemas are often represented at an abstract level through the entity relationship model (ERD). ERD needs to be normalized as highest as possible to avoid data redundancy or data inconsistency. The operational databases are optimized for preservation of data integrity and speed of recording business transactions through the use of database normalization.

In contrast, DW needs to become as simple as possible since it integrates a huge set of data so that the speed of data analysis can be optimized. Usually, the data in DW is denormalized via multidimensional model (star schema). It contains the facts and dimension tables. The fact tables contain numeric values that present a measurement of business rules. A row in a fact table corresponds to a measurement at the same grain such as the students’ CGPA. The dimension table is a description of the fact tables such as students’ information, time dimension and geographical dimension. EDW multidimensional Model resides in the Back End Layer.

In our study, five dimension tables and one fact tables have been initialized in the early stage to analyze the students’ performance. Geographical Dimension (D\textsubscript{GEO}), Time Dimension (D\textsubscript{TIME}), Program Dimension (D\textsubscript{PROG}), Student Dimension (D\textsubscript{STU}), Subject Dimension (D\textsubscript{SUB}), Assessment Dimension (D\textsubscript{ASM}) and Student Result Fact (F\textsubscript{RES}) have been created to perform OLAP analysis. Each dimension may contain its own group of hierarchy. Fig. 4 shows the relationship between a table and former specifications of the multidimensional model.
In order to support the OLAP analysis, a cube (C) relation can be described as follows:

\[ C = (D_{\text{GEO}}, D_{\text{TIME}}, D_{\text{PROG}}, D_{\text{STU}}, D_{\text{SUB}}, F_{\text{RES}}) \] (1)

Given a particular dimension table Di, the members of the dimension are the values in this table and they are arranged in a hierarchy denoted \(<i>\). Given an n-dimensional cube \(C = D_1, \ldots, D_n, F\), a cell reference (or reference for short) is an \(n\)-tuple \((m_1, \ldots, m_n)\) where \(m_i\) is a member of dimension \(D_i\) for all \(i \in [1, n]\). A cell is a tuple of cube \((F)\) where cube \((F)\) denotes the data cube [14] of the fact table F.

### 3.3 Descriptive and Predictive EDW Analysis

The data available in DW will be used for analytical purposes and the OLAP application has been used to provide the descriptive analysis. Through this application, details of students’ performance over a subject and students’ achievement records can be obtained. For example, a query to obtain a student’s performance on a specific subject group based on certain amount of time can be triggered by combining Subject Dimension \((D_{\text{SUB}})\), Time Dimension \((D_{\text{TIME}})\) and Student Dimension \((D_{\text{STU}})\) with Fact \((F_{\text{RES}})\).

```
SELECT * from D_{\text{SUB}}, D_{\text{TIME}}, D_{\text{STU}}, F_{\text{RES}} WHERE D_{\text{SUB}}.ID = F_{\text{RES}}.ID AND D_{\text{TIME}}.ID = F_{\text{RES}}.ID AND D_{\text{STU}}.ID = F_{\text{RES}}.ID AND F_{\text{RES}}.ID = i
```
The query can be modeled in the relational algebra as follows:

\[ \Pi_{\text{pointer}}(\sigma_{\text{stu_id} = "i"}(F_{\text{RES}})) \]  

Where \( i \) can be any relevant student ID number from databases.

To obtain the description of students’ information or time dimension, the fact table can be linked to the dimension table as follows:

\[ (\Pi_{\text{ID}}(F_{\text{RES}}) \cap \Pi_{\text{ID}}(D_{\text{SUB}})) \&\& (\Pi_{\text{ID}}(F_{\text{RES}}) \cap \Pi_{\text{ID}}(D_{\text{STU}})) \&\& (\Pi_{\text{ID}}(F_{\text{RES}}) \cap \Pi_{\text{ID}}(D_{\text{TIME}})) \]  

The Predictive Analysis is used to forecast or predict students’ achievements or to obtain the potential areas interests of the students. This analysis must be based on the data gathered in the predictive analysis. Data mining techniques will be used to perform the predictive analysis. However, model to be used in this project is not decided yet. A study will be conducted to gather the best data mining model to be implemented in the multidimensional model such as AHP, TOPSIS and Fuzzy. Previous researches have utilized various techniques in educational data mining including statistical classification, fuzzy rule induction and neural network to classify data [9], association rules [3], and clustering techniques [8].

Some major analyses that need to be conducted are:

i. Students’ Achievement Analysis based on the subject group

The objective of this analysis is to obtain the best subject group of students. Based on Information Technology (IT) subject groups (networking, multimedia, software engineering, programming and databases), the graph of the analysis result can be produced to perform the comparison and to view achievement captured over specified time frame. Acknowledging the best subject group is important in order to suggest the best computer field for the students to further their higher learning in future. In addition, performing mining techniques to find reason of each student’s improvement or vice versa can be done. The fact result has many (m-1) to one relation with dimension assessment. Each student’s assessment (number of quizzes taken, attendance, number of assignments done, and frequency of meetings with lecturers) will be recorded as method to investigate whether the assessment can help students improve their grades.

ii. Clustering Analysis

Clustering is a process of grouping physical or abstract objects into classes of similar objects. Both clustering and classification are data mining methods. Clustering is an unsupervised classification and classification is a supervised classification [15]. The goal of the clustering technique is descriptive rather than predictive [8]. It can be done by categorizing students into homogenous groups such as groups of programs (multimedia & IT), gender or pointer. Identifying students based on similar group can help researchers perform more specific analysis against the group. Thus, decision can be made for
choosing the best learning method to be applied to a particular group of students. Talavera & Gaudiso [7] have conducted the cluster profiling in mining student of E-Learning environment. Fig. 5 shows an example of clustering students based on several variables in unstructured collaboration spaces.

### iii. Predictive Analysis

Predictive analysis is a process of predicting the pattern or future results based on current data. It is important in Decision Support System (DSS) to make better decision. One of the famous methods in EDW for the predictive analysis is association rules. The association rules find relations between items [6]. Rules have the following form: \( X \rightarrow Y \), support 40%, confidence 66%, which could mean "if students meet lecturer for consultation more than 10 times, they get pointer more than 3.00 for the subject", with a support of 40% and a confidence of 66%. Support is the frequency in the population of individuals that contains both \( X \) and \( Y \). Confidence is the percentage of the instances that contains \( Y \) amongst those which contain \( X \).

Let \( I = \{I_1, I_2, ..., I_p\} \) be a set of \( p \) items and \( T = \{t_1, t_2, ..., t_n\} \) be a set of \( n \) transactions, with each \( t_i \) being a subset of \( I \). An association rule is a rule of the form \( X \rightarrow Y \), where \( X \) and \( Y \) are disjoint subsets of \( I \) having a support and a confidence above a minimum threshold [3].

Let us denote by \( |X \cup Y| \) the number of transactions that contain both \( X \) and \( Y \). The support of that rule is the proportion of transactions that contain both \( X \) and \( Y \): \( sup(X \rightarrow Y) = |X \cup Y|/n \). This is also called \( P(X, Y) \), the probability that a transaction contains both \( X \) and \( Y \). Note that the support is symmetric: \( sup(X \rightarrow Y) = sup(Y \rightarrow X) \).
Let us denote by \(|X|\) the number of transactions that contain \(X\). The confidence of a rule \(X \rightarrow Y\) is the proportion of transactions that contain \(Y\) among the transactions that contain \(X\): \(\text{conf}(X \rightarrow Y) = |X \cap Y| / |X|\).

An equivalent definition is: \(\text{conf}(X \rightarrow Y) = P(X,Y) / P(X)\), with \(P(X) = |X| / n\). Some variables that should be included in the project are shown in Table 1.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_quiz</td>
<td>Number of quizzes taken (passed / failed)</td>
</tr>
<tr>
<td>n_assignment</td>
<td>Number of Assignments done (pass / failed)</td>
</tr>
<tr>
<td>n_consultation_hour</td>
<td>The frequency of meetings/discussions with lecturers</td>
</tr>
<tr>
<td>n_attendance</td>
<td>Number of attendance (class, lab, tutorial)</td>
</tr>
<tr>
<td>n_freq_KELIP</td>
<td>The frequency of using KELIP(online notes/ exercises)</td>
</tr>
<tr>
<td>Tota_time_KELIP</td>
<td>Total times using KELIP</td>
</tr>
<tr>
<td>Total_carrymark</td>
<td>Total carry mark</td>
</tr>
</tbody>
</table>

### 4 IMPLEMENTATION TOOLS

The costs of hardware and software involved in implementing DW projects are extremely high. The commercial BI software such as Cognos, SAS, Business Object and Hyperion usually cost thousands up to hundred thousand dollars. As budget constraint is one of the major issues in DW projects, we have decided to use Open Source BI/ETL tools. There are several open source BI and ETL tools available in the market. Most existing tools have begun in 2005. Even though their capabilities are limited and having bugs compared to the commercial products, they are rapidly being enhanced and improved. The success of using open source tools in DW projects may dramatically reduces costs. As a result, small companies will gain opportunities to apply DW. Eclipse Business Intelligence and Reporting Tools (BIRT), Jesper Reports, Jfree Reports and Palo are examples of DW projects using open source BI tools. Meanwhile, ETL Talend Open Studio (TOS), Kettle ETL and Enhydra Octopus are examples of DW projects designed specifically for data integration and migration purposes.

TOS has been used to design data quality framework for DW [16]. TOS is an open source project for data integration based on Eclipse RCP [17]. TOS leverages the open source model to make data integration available to all types of organizations, regardless of their size, level of expertise or budgetary constraints. It may be used to connect all source and target systems. TOS is distributed under GPLv2 and has been launched in October 2006. In January 2008, it has been downloaded over 1 million times. In July 2009, the product was downloaded 5 million times and used by over 300,000 users [18]. TOS
operates as a code generator allowing data transformation scripts and underlying programs to be generated either in Java (since v2.0) or Perl (since v1.0). Its GUI is made of a metadata repository and a graphical designer. TOS is a metadata-driven solution. All metadata is stored and managed in the repository shared by all the modules. The jobs are designed using graphical components. Some of the functionalities offered by TOS are synchronization or replication of databases, right-time or batch exchanges of data, ETL for analytics, complex data transformation and loading and data migration. Fig. 6 shows how to integrated data from source system to DW and Fig. 7 shows how mapping is being done.

Fig. 6. Populate Data to DIM_GEO

Fig. 7. ETL Job for DIM_GEO using TOS
Some reports produced from DW are employee’s salary report and number of employees’ based on region. Those reports are gain from BIRT. Fig. 8 shows one of example report produce by BIRT.

![Fig. 8. Sample Report using BIRT](image)

### 5 Conclusion

Inspired by research challenges in developing EDM, we have proposed an EI architecture to analyze students’ performance behavior. Data from the transactions systems as well as external sources will be extracted to EDW to design the star schema model. Then, the OLAP techniques can be utilized to obtain students’ achievements and to perform descriptive analysis. Decision Support System (DSS) technique and algorithm can be applied in order to forecast potential areas of studies for the students. We are also proposing association rules to be used as part of the predictive analysis techniques with the hope to derive the best teaching mechanisms for individual student. This project requires participation from many people in different computer backgrounds such as database, data mining, decision support system, and information retrieval. Future works can be focused on implementing the EI architecture in critical areas such as determining the most suitable course for secondary students to pursue their studies in a higher learning institution, or be of assistance in the process of selecting suitable candidates in public universities.
References