A Business Intelligence Training Document Using the Walton College Enterprise Systems Platform and Teradata University Network Tools

Jeffrey M. Stewart
College of Business
University of Cincinnati
stewajw@mail.uc.edu

Roger H.L. Chiang
College of Business
University of Cincinnati
chianghl@mail.uc.edu

Abstract

This paper discusses the process of integrating materials from Walton College Enterprise Systems (WCES) at the University of Arkansas and the Teradata University Network (TUN) in the Business Intelligence curriculum at the University of Cincinnati. The opportunities and challenges of teaching students the knowledge discovery in databases (KDD) process using a real-world data warehouse from WCES and BI software tools from TUN are discussed. The result of the project was an all-in-one training document to assist students with remote access, connecting to the datasets, and performing BI analytic techniques.

Introduction

This paper presents the opportunities and challenges associated with incorporating resources from the Walton College Enterprise Systems (WCES) at the University of Arkansas into undergraduate and graduate classes on Business Intelligence (BI) at the University of Cincinnati (UC). With the increasing demand for students who can align IT with business strategy (Winter et al. 2007), there is a clear and urgent demand to educate students who know how to implement real world BI applications..

The all-in-one training document created for our BI course is a synthesis of materials available from WCES and the Teradata University Network (TUN). It enhances student understanding of the knowledge discovery in databases (KDD) process by accessing and implementing BI tools and data warehouses provided through the partnership. Students are asked to assume the role of BI analysts for Sam's Club and use the software and data available to discover knowledge and provide business insights. This document can be used as a reference document for incorporating the resources of WCES at other university for their BI courses.

Some challenges for incorporating BI tools in the classroom are: finding a software tool that students can learn BI methods on; locating a data warehouse for students to analyze; and creating instructional materials that guide learning. Our selection criteria were low cost access and use due to limited educational funds available to purchase software. Therefore, obtaining software at zero to very little cost was important. Also, many students in both the graduate and undergraduate programs at UC have full-time employment and require off-campus access to computing resources. UC is on the quarter system, so any software used must be able to be taught to students in a 10-week period, with much of that 10 weeks being used to teach the concepts and methods that they will apply in the software. Finally, the course is taught to both graduate and undergraduate students; concepts and software applications needed to be approachable by a diverse group of students. These needs offer a challenge for teaching any BI application and are not dissimilar to any university.

Through contacts with faculty at the University of Arkansas, Walton College of Business, the instructor was made aware of an instructional site at the University designed to meet these specific needs. The WCES (http://enterprise.waltoncollege.uark.edu/) offers remote access to enterprise BI services from four vendors: IBM, Microsoft, Teradata, and SAP. In addition to those services, there are real-world data warehouses available from Dillard's, Hallux, Sam's Club, and Tyson Frozen Foods. The educational opportunity available to educators and students is to use real-world data using commercial software applications such as Teradata Warehouse Miner. WCES site offers training documents and remote access materials that can be adapted for use in the classroom. Instructors can register themselves and obtain student accounts.

Teradata Applications

Considering the time constraint of a 10-week course, the Teradata tools were chosen as a pilot case. The software tools available through WCES are offered as part of the Teradata University Network (TUN). The goal of the TUN is to serve as a central source for data mining, business intelligence, and decision support systems knowledge and to be the bridge between the academic classroom and the real world of BI (Winter et al. 2007). Through a partnership with WCES, TUN offers the use of Teradata SQL Assistant and Teradata Warehouse Miner which can access the real world datasets available at WCES. Through the WCES and TUN collaboration, students can implement BI tools that they are likely to encounter in practice, and real world datasets that accurately reflect the "messiness" of actual business data.

The Teradata SQL Assistant uses ANSI-standard SQL format, allowing students with limited SQL knowledge to successfully query a database. The tool is available for download and connects through an ODBC to the WCES datasets, or can be accessed remotely through the WCES server, where the ODBC is already set-up. Student can browse the table structure of multiple databases, including primary keys, foreign keys, and other data columns. Also, student queries are stored, allowing students to recall previously specified queries.

The Teradata Warehouse Miner is an extension of the Teradata SQL Assistant, and provides an intuitive graphical user interface for students to build the complicated SQL queries for implementing BI analytic techniques, including data-mining and statistical techniques. Students have extensive options for query refinement, can run multiple queries at once, and obtain graphical interpretations of analytic results. Students can build projects, store past queries and continuously revise results as they build an understanding of the data warehouse and analytic techniques. For our BI hand-on implementation purposes, the designed exercises focus on three primary analytic techniques provided by the Teradata Warehouse Miner. The analytic techniques are cluster analysis, association (market basket) analysis, and decision tree analysis.

Sam's Club Dataset

The instructor and teaching assistant have examined the datasets available at WCES for fit with course goals, and decided to use Sam's Club dataset. The primary benefit of working with the Sam's Club data set is that students can easily understand the customer experience and operational aspects of a Sam's Club store. This goal is that students can more rapidly translate knowledge discovered in the database into business decisions. Other benefits of using the Sam's Club dataset are: size, ease of understanding data structure, existing relational table of star schema. The Sam's Club dataset was provided to the University of Arkansas by Wal-Mart, and consists of 6 tables and over 55 million rows of data (see Figure 1). The analytic focus of the Sam's Club dataset is individual member transactions. Each register transaction captures the items a Sam Club member bought, transaction details, and the member's ID. Other tables provide information about the store, the member and the items.

The Course Project

The next challenge was creating a concise, usable all-in-one training document for students and an assignment that enhanced their understanding with the hands-on implementation of the subject matter. Registered users of the WCES have access to a wealth of training information that includes how to access and use software tools, how to perform analytic techniques, and in-depth explanations of analytic techniques. The materials were created for in-class use at the University of Arkansas. The challenge for courses outside of the University was to consolidate and integrate the information into an easy to follow, all-in-one document.

The goal in creating a usable, all-in-one guide was to create effective instructional materials that would enhance the students understanding of KDD techniques using the tools and data sets available. One criteria for the guide was that it go beyond asking students to perform analytic techniques, and encourage them to interpret the results of those techniques in terms of actual business implications. Therefore, a critical outcome of the course was that students understand the business context of the analytic techniques, and approach the interpretation of results with a creative business lens.

To help in the creation of the training document, a basic story was created. Students were asked to assume the role of a Business Intelligence Analyst for a group of Sam's Club stores in a geographic region. Students would follow the step-by-step guide in moving from million of rows of data in a data warehouse to concise business knowledge (see Figure 2) for strategic decision making. Our all-in-one training document has five sections. The first section introduces the BI concepts and KDD process.

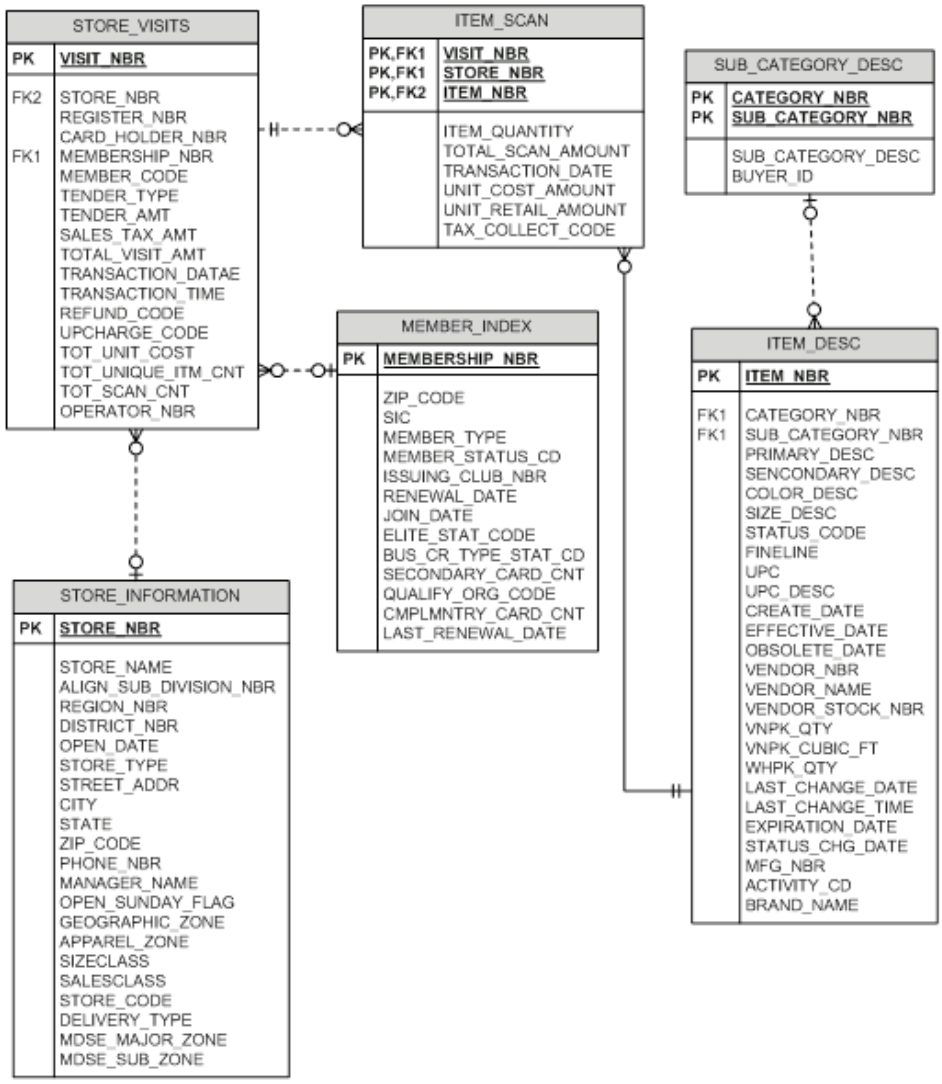

Figure 1 - Sam's Club ERD. Retrieved from: http://enterprise.waltoncollege.uark.edu/1682.asp

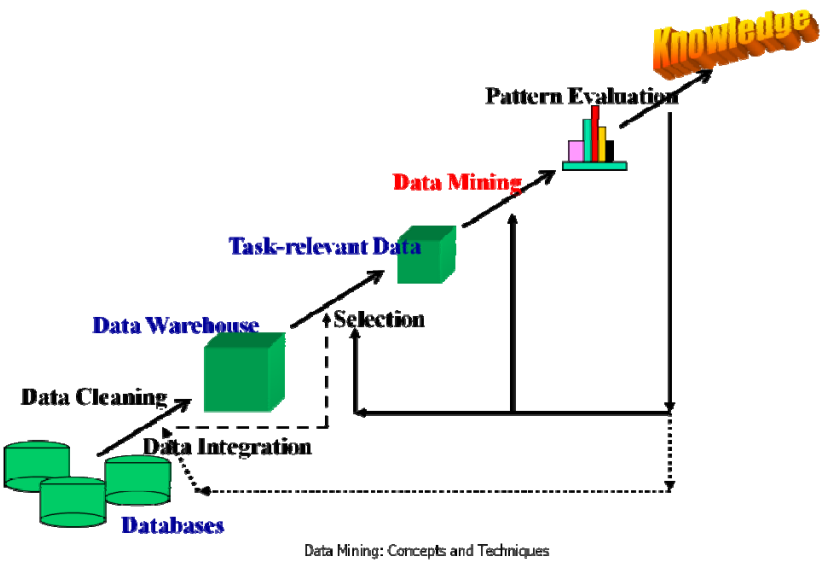

Figure 2 - KDD Process (Han et al. 2006)

The second section introduces the Sam's Club dataset. First, the dataset description available on the WCES is presented so students can see a verbal description of the dataset. When students learn dimensional modeling, assignments often ask them to present the simple data warehousing requirements as a star schema. The data warehousing requirements for the Sam's Club dataset helps students understand the checkout process of a Sam Club member, and provides enough information for students to understand the data warehouse they will be using in the BI assignment. Additionally, students are given the star schema using crow's foot notation taken from the WCES website. The first part of the hands-on implementation is to understand the data stored according to the provided star schema. The goal for this part is to reinforce concepts introduced earlier in the course. In addition to the dimensional tables of the star schema, there is a fact table in the table called "WAREHOUSE" and students are asked to use the Teradata SQL Assistant to query the fact table and describe the attributes of this fact table.

The third section, then, illustrates the process of accessing the University of Arkansas remote server, where students will access the two Teradata tools. The instructions use screenshots for every step of the process, and detail every field the students will need. Then, this section walks students through the process of opening the Teradata SQL Assistant and connecting to the Sam's Club dataset. After the completion of this section, students can begin executing queries on the dataset and gain a deeper understanding of the data warehouse in preparation for accessing the Teradata Warehouse Miner.

The three analytic techniques that students will use are discussed in the fourth section. Students are shown how to run an association (market basket) analysis, create a decision tree, and perform cluster analysis. The association analysis asks students to use the unique identifier of the dataset (visit_nbr) with the item primary description as the dependent column. The other available options are outlined and the necessary settings to get accurate, interpretable results are also described in the document. In addition, students are given the screenshot of an example output to assist them in obtaining analytic results.

The decision tree analysis asks students to create a decision tree that identifies a member's status code based on several factors, including total scan count, total unique item count, and total visit amount. Students are given six columns to enter as the dependent columns (see figure 3). This decision was based on instructor experimentation with numerous combinations of columns. Obtaining a decision tree that balanced complexity with interpretability was a difficult process, and leaving students to find the suitable columns for creating a decision tree greatly increased assignment completion time without simultaneously increasing the student's understanding of the KDD process. The settings used were adopted directly from assignments available through the WCES materials.

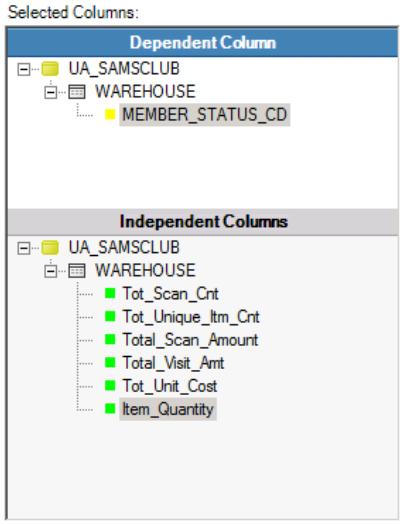

Figure 3 - Decision Tree analysis

Finally, the document demonstrates how to perform a clustering analysis. Again, in the interest of balancing complexity with interpretability, students are asked to limit their analysis to two columns, and number of clusters to two. As with the two previous analytic techniques, students are given an example input and output from the analysis so they understand the settings necessary to get a readable output, and a guide for what a readable output should resemble.

Assignment

Both graduate and undergraduate students were given the same hands-on BI implementation assignment. The assignment asked students to perform each of the analytic techniques, discuss the process of obtaining results, display and interpret the results, and offer business recommendations based on the results. The goal of the assignment was for students to understand the "messiness" of real world data, emphasize the importance of understanding the data in the data warehouse, and emphasize the linkage of results to business strategy. The last goal of the assignment is the most important; ultimately, students must understand that data mining is a collection of analytic techniques, and knowledge discovery is the work to support business intelligence (Fayyad et al. 1996).

Instructors encouraged students to think beyond even the obvious conclusions and find real business insights. As an example, after performing a market basket analysis, students often saw that milk and eggs had a high probability of appearing in the same transaction. A small percentage of students failed to even see how this could even be applied in a business context; those that did insisted that milk and eggs should be close together so customers could easily find two items most likely to be purchased together. The real learning curve for students was in creating store design, marketing, and pricing insights, such as placing a high-profit item between milk and eggs, locating the milk and eggs at the back of the store, or not placing the two on special at the same time.

Students often exhibited a pre-conditioned need to find a single "right answer." The instructors worked diligently with students in and out of class to engage them in creative business problem solving. Students frequently looked for validation that their insight from the data was the pre-defined correct answer, and were timid to embrace the notion that any answer was correct when justified with data-based insights. Instructors pushed students to ask themselves "so, what?" and "what does that mean to the business?" In terms of their role as a business analyst for Sam's Club, students were asked to provide a specific recommendation to a Sam's Club store manager and justify their recommendation with output.

The potential business insights, the knowledge discovered, are the real benefit of working with the real world dataset from WCES. Students are not asked to think of solutions to the fabricated problems of a hypothetical business from a randomly generated dataset; instead, students can use the results from the assignments to create insights for a real world company. Instructors continually reinforced to students to consider an actual Sam's Club store when making

recommendations. This real world context helped students better understand how the business intelligence and KDD process differs from data mining by bridging between IT and the business.

Lessons Learned

The integration of materials from TUN and WCES was our first attempt for including BI tools available through WCES partnerships like IBM, Microsoft and SAP in designing hand-on exercises for the BI course. Overall, we believe our experience was successful. Students expressed their appreciation for the opportunity to apply three analytic techniques they have learned to a real world dataset. Several students expressed an interest in continuing to explore the data and BI tools as their accounts remained active for the remainder of the school year.

Based on our first experience of incorporating materials from WCES into the BI curriculum, future BI classes should have extended amount of time devoted to understanding the analytic techniques and interpreting the output from decision making point of view. If time allows; a one-hour lab session is not enough to comprehend the KDD process for the hand-on BI implementation. We recommend subsequent classes or/and lab sessions devote to learning the material with small assignments between classes that help ensure students have learned the basics of BI tools.

Secondly, we recommend that graduate students and undergraduate students be given different assignments that reflect the differing levels of knowledge between these two groups of students. For the assignment, undergraduate students scored an average of 38.6 out of 50, with a standard deviation of 11.3; while the graduate students scored an average of 48.1 with a standard deviation of 1.9. This is a small sample of students and based on the first attempt at using the document and assignment, and it seems clear that undergraduate students struggled more with the assignment. A possible reason is their lack of exposure to real-world business concepts. The BI course is more beneficial for students with greater exposure to multiple business disciplines like marketing and economics that allows them to recognize business knowledge when it is discovered. Oppositely, graduate students are likely to have full-time employment, an undergraduate business education, and business experience. Graduate students are better able to understand BI results in a business context.

The materials available through TUN and WCES provide an excellent bridge for spanning the gap between BI knowledge and real world BI applications. Students can learn the KDD process using BI tools and working on data warehouses obtained from real-world businesses. This practical context for the classroom benefits students seeking careers in BI, and generally better prepares IS/IT students to understand how data and technology and play a vital role in business.

References

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. *From Data Mining to Knowledge Discovery in Databases*, 1996.

Han, J., and Kamber, M. *Data Mining: Concepts and Techniques*, (2nd ed.) Morgan Kaufmann Publishers, 2006.

Winter, R., and Gericke, A. "Teradata University Network: A Resource for Preparing and Teaching Business Intelligence and Data Warehousing Courses " 2007 Computer Science and IT Education Conference, Mauritius, 2007.

Appendix A: Assignment

University of Cincinnati
College of Business
Business Intelligence

Homework Assignment: Knowledge Discovery from Databases

For this KDD homework assignment, you will use Teradata SQL Assistant and Teradata Warehouse Miner to perform data mining tasks. You should refer to the Data Mining document prepared for this class along with what you have learned from the in-class demonstration to conduct three data mining tasks:

1. Perform a market basket (association) analysis using the UA_Samsclub data warehouse on the store that was assigned to you.
    a. Present and interpret the results.
    b. Suggest changes by referring to the result of this analysis.

2. Create a decision tree using the UA_Samsclub data warehouse.
    a. Present and interpret the results.
    b. What are the inherit flaws and limitations of the resulting decision tree?
    c. Discuss what a usable decision tree would look like.
    d. What other data from the WAREHOUSE table should be included, and how could it be transformed?
    e. With your suggested improvements, what information and insight could it provide to the organization?

3. Perform a clustering analysis using the UA_Samsclub data warehouse.
    a. Present and interpret the results.
    b. How could the tables and data be improved or transformed to create a more useful analysis?
    c. With your suggested improvements, how could this clustering analysis provide information and insight to the organization?