

WILKES, J
JMW-MINE
TYPE: IQP
DATE: 5/01

46

01D117I

LRN: 01D117I

Project Number: JMW-MINE -46

The Potential in Data Mining:

The WPI Class of 2001 as a Case Study

An Interactive Qualifying Project Report

Submitted to the Faculty


of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the degree requirements for the

Degree of Bachelor of Science

By



Jonathan Oexner

Date: 2 May, 2001

Approved:



Professor John M Wilkes, Advisor

Abstract:

The class of 2001 “case study” is a preliminary step toward an Integrated Student Information System (ISIS). The goal is to create a unified database that will allow analysis of the team dynamics and community structure implicit in data on this portion of the WPI student body. Currently the database is not complete, but the structure created will conveniently accept the database additions currently underway and anticipated in the future. As a start toward learning where different types of personality types cluster in the WPI community, a preliminary analysis of the membership of a well-known service fraternity (Alpha Phi Omega) as compared to the student body as a whole revealed a cluster quite different from the student body as a whole, in ways that were not anticipated. Hence, this approach promises to be very revealing.

The purpose of this project is to assemble a single large database which will aggregate much of the information available on WPI's class of 2001 in order to allow for an integrated statistical analysis. This would be a first step toward creating a "prototype" Integrated Student Information System, or ISIS for WPI. A secondary goal is to establish a format for further data capture in such a way as to facilitate further analysis of the larger class of 2001 – 2004 database under construction.

The first step in this process was to collect data from as many sources as possible. Since several previous projects had gathered such information for relatively small-scale analysis, the data available was in several disparate formats. It often consisted of solitary Microsoft Excel spreadsheets, which are notorious for allowing the insertion of arbitrary data into arbitrary cells, which is a tremendous problem when importing the data into a database, which imposes much stronger restrictions on the data. Consequently, the task of merging these disparate sources into one conjoint database was not an easy one. Extensive database programming has been necessary to reformat this data, through conventional means such as SQL programming, more sophisticated, custom programs written in Java using the Java Database Connectivity API (JDBC), and when all else failed, editing many of the entries by hand. However, at the conclusion of this project, it is apparent that the database contains a wealth of information on many subjects, such as cognitive style and personality type, aptitude test scores, high school background, freshman housing arrangements, social groupings, academic performance, team memberships, activity participation, survey data on self-image and expectations at the start of the freshman year and the complex interactions of all of these traits in the formation of the WPI student community and social organization.

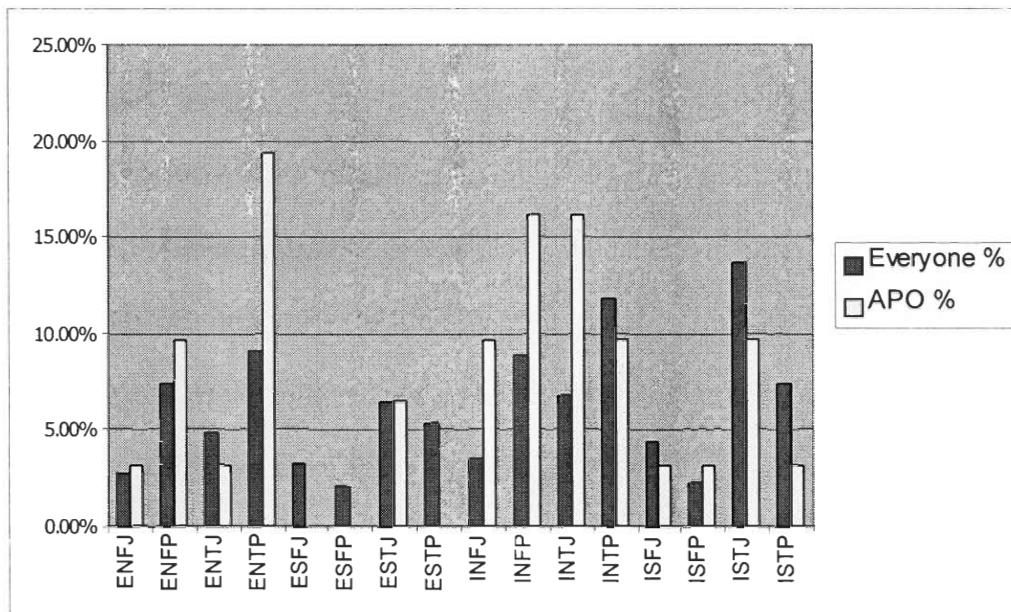
The difficulty of doing this was somewhat exacerbated by cases that didn't quite fit into the assumptions that were made originally. Transfer students, early and late graduations that put surprising students onto MQP teams and interdisciplinary project teams all added to the potential for confusion. In the end however, I had a database that could support the study of group formation, team dynamics and the social networks that make up the class as a whole.

This wave of the project will press on. My contribution to the team effort with my partners Phil, Jerome, Bryan and Eric is over. They were (and are) responsible for data capture and analysis. My responsibility is middle step of database configuration and organization. Many possible sources of data were accessed, formatted, and inserted into the database, to be used as an illustration now and the basis for full analysis by subsequent teams. Many judgment calls on cases such as this were made in the interest of producing useful, meaningful results.

This brings us to the first illustration of the kind of question that an ISIS can address. Since the goal of this project was data integration, this is intended solely to demonstrate the utility of a single, unified database, as opposed to storing many separate pieces of data. One of the possible interesting linkages between data sets is that between MBTI types and group membership. The question is, "Do some self-selected groups on the WPI campus tend to attract and serve as a meeting ground for certain personality types (in disproportionate ratio), as compared to the general WPI population?" We call it the "birds of a feather" hypothesis. With the database assembled here, it's a simple task to answer such a question. Simply formulate a query, enter it into the database and analyze the results. Here, I exported the results to an Excel spreadsheet from the original

Access database, did a bit of statistical analysis and graphed the results for seniors in an organization dear to my heart – the service club Alpha Phi Omega.

	Everyone	Everyone %	APO	APO %
ENFJ	15	2.77%	1	3.23%
ENFP	40	7.39%	3	9.68%
ENTJ	26	4.81%	1	3.23%
ENTP	49	9.06%	6	19.35%
ESFJ	18	3.33%	0	0.00%
ESFP	11	2.03%	0	0.00%
ESTJ	35	6.47%	2	6.45%
ESTP	29	5.36%	0	0.00%
INFJ	19	3.51%	3	9.68%
INFP	48	8.87%	5	16.13%
INTJ	37	6.84%	5	16.13%
INTP	64	11.83%	3	9.68%
ISFJ	24	4.44%	1	3.23%
ISFP	12	2.22%	1	3.23%
ISTJ	74	13.68%	3	9.68%
ISTP	40	7.39%	1	3.23%



Here we see a representation of the relative percentages of the different MBTI personality types within a specific club, in this case the members of Alpha Phi Omega (APO) service fraternity in the class of 2001, and the WPI class of 2001 as a whole. As you can see, certain personality types, such as ENTP, INFJ, INFP and INTJ, are much more strongly

represented in APO than in the general population – and some are notably underrepresented. As more complete information on group, team and club membership becomes available to the project, it will continue to be a simple matter to analyze subgroups of the WPI population in comparison to the population as a whole. My teammates are adding in all of the varsity team rosters and 2 fraternity membership lists to compare to APO.

Already, extensive data on the class of 2001 is contained within the database. Out of the 678 cases which comprise the base of the class of 2001, we have acquired: 594 cases' Cognitive Style (GCSI) data, 623 cases' Freshman housing data (indexed by student, term, dormitory and room), high school grade data from the 265 seniors who released their transcripts for analysis, MBTI (personality type) data from the 542 seniors who released that information (with an effort underway to retrieve as many as 100 additional cases), SAT data from 636 seniors who released that information, and the freshman year grades of 680 seniors who released those. This is in addition to a cross-year database that tracks participation in 19 different activities on campus. This database is one of the most interesting to me as it allows us to test the cognitive cluster hypothesis, and will grow even more interesting with the addition of more complete data on more and more campus clubs and groups.

The potential for analysis is just overwhelming, but in order to live up to its full potential, the database on active students must be diligently updated and maintained. The class of 2001 database is not 1 year behind on grades at present, but 3. Major changes can not be tracked by its likely causes in performance shifts. The cost of getting behind is clear. The IQP experience of the class is also missing, but can be added later. This

means that all data, as soon as possible after initial capture, must be entered directly into the database. The most efficient method of data entry would be to create an Access “form” for each new piece of data that would ensure that all data entered is in fact valid. While this would take some substantial expertise on the part of at least one member of the data entry team, it would virtually eliminate the unnecessary data loss and corruption that have plagued researchers attempting to maintain such large databases in the past. As an example, data entry team members might be presented with a drop down list for the selection of a predefined user rather than manual, error-prone entry of the name directly. Membership in a new group would be a simple task of filling in check boxes next to the names of group members. Essentially, the use of forms would make it virtually impossible for the data entry team members to mistakenly corrupt the database.

A somewhat less complicated approach might be to simply have the data entry team enter the information into the database directly via a row-edit method. However, this would also necessitate the use of careful checks on the newly-entered data immediately after entry, via the use of custom queries that would highlight suspected data entry errors and allow their repair while the original data source is still available for verification.

However, the ISIS team sincerely believes that this effort will pay off in time. CIRP data, which measures student expectations upon arrival at WPI, high school grade data, standardized test scores, activity participation/group membership, personality type, cognitive style and freshman housing data can all be linked to and combined with freshman academic performance in ways that have never been possible before. This will enable the administration of WPI to better ascertain which types of students it best serves,

and which groups need more attention, and do its best to better serve all students. This is, after all, the ultimate goal of the university, and is well worth the investment. ISIS shows promise as a versatile tool of use to campus organizations, administrators and educators as well as advisors and the admission and career development office. Any group interested in knowing what types of learner WPI most and least effectively will benefit from their new capability. I have created to rapidly produce single !!!!!!! analysis that can be loaded directly into a statistical software package, with just the variable you need.