

Data Mining And Knowledge Management

A System Analysis for Establishing a Tiered Knowledge Management Model (TKMM)

Jing Luan, Ph.D.

Terrence Willett, M.S.

Researchers are living in an era of quiet yet pervasive advancement in technology that has instigated an unforgiving expansion of data analysis techniques. This brings significant pressure to update skills and analysis tools. Researchers in higher education in particular have gradually begun to realize that yesterday's modus operandi is dated, as technology is knocking on their office doors with promises of efficiency, speed, and convenience. In response, many researchers have cautiously opened their doors to new software, new technology, and new ideas. While there is romance in being on the cutting edge, those who have been among the first to install a new program only to find it is full of bugs or not actually useful have seen the value of letting others be the first to tread into uncharted terrain. However, we are also constantly reminded that we do not want to be "left behind" and miss out on truly valuable new techniques and concepts. This paper will focus on current and nascent methods in data storage, analysis, and dissemination in the forms of Datawarehousing, Data Mining (not data "fishing"), and Knowledge Management. The paper culminates in the Tier Knowledge Management Model (TKMM) that seeks to provide a stable structure with which to organize the plethora of established and nascent technologies.

Datawarehousing

Data mining cannot be discussed without first relating it to datawarehousing. A datawarehouse is a "subject oriented, integrated, nonvolatile, time variant collection of data in support of management decisions" (Corey & Abbey, 1999). Many researchers have established datawarehouses on their campuses. If they haven't yet, they might have datamarts which is a specially constructed database for a limited research study (not to be confused with Fact Tables, which are associated with the Star Schema). The difference between a datawarehouse and a datamart lies in the size and purpose. A datawarehouse, typically derived from an OLTP system such as Peoplesoft or Datatel, is comprehensive in nature and may become too large for a study that only requires a few fields of data. This requires the introduction of another layer of data management (described later as the Middle Tier) to extract specific fields based on a query.

Datawarehousing technology has drastically expanded the amount and type of data available for analysis, therefore making data mining a possibility. It's worth noting that datawarehouses ought to be built with the research needs in mind. The authors have sometime noticed the tendency to "hoard" data with no pre-planning for how the data should be arranged. This has resulted in a potpourri of data files that have not been properly indexed and may not relate to each other properly or that they have both third (3NF) and second norm form (2NF) co-existing – all are hidden risks severely impact accurate and efficient analysis. If the high salaries paid to DBA (database administrator) are any indication of the complexity of this task, the authors recommend that researchers working with databases study database theory and design.

Data Mining

Data mining is a knowledge discovery process to reveal patterns and relationships in large and complex data sets (De Veaux 2000). It should not be confused with data “fishing”, which is a dubious post hoc search for spuriously significant findings. Data mining uses well-defined algorithms from the disciplines of machine learning and artificial intelligence to discern rules, associations, and likelihoods of events by looking backward at data in its raw and longitudinal form. The power delivered by data warehousing to data mining software has broken the mold of the traditional statistical methodologies revered by generations of analysts (Mena 1998). Trained in the theories of sampling and the sutra of hypothesizing, statisticians have been practicing their trade with pre-conceived notions, or a-priori hypotheses, tested on data sets collected ideally after the hypothesizing and specifically for testing the hypothesis. Conversely, data mining discerns an optimal model based upon existing data sets created by routine and often continuous data collection, such as MIS data. If the traditional methods can be viewed as top down, data mining is truly bottom up. Research questions do not begin with “what if”, instead, they begin with “what is”.

Data mining has been recently discovered by academia but was first put to full use by the Fortune 500 who have since benefited tremendously. Data mining was behind numerous successful market campaigns and quality assurance (QA). Below in Table 1 is a comparison chart depicting some of the core questions most often used in the business world and their analogs in higher education:

Table 1. Comparing data mining questions in education and the corporate world:

Bottom-line Questions in the Business World	Counter-part Questions in Higher Education
Who are my most profitable customers?	Who are the students taking most credit hours?
Who are my repeat website visitors?	Who are the ones likely to return for more classes?
What clients are likely to defect to my rivals?	What type of courses can we offer to attract more students?

Exploring Data Mining for Higher Education

In a traditional approach to science, the researcher would make specific a-priori hypotheses and then test them using data collected specifically to test these hypotheses. For instance, a researcher might propose that the probability of a transfer student graduating can be predicted from GPA at time of transfer, GPA at the transfer institution, ethnicity, and field of study. Logistic regression or some other analysis would test this model for significance. In a strict analysis, we would be left with our hypothesis either being supported or not. Further testing after the fact would encourage the discovery of spurious findings unless care was taken to protect against alpha inflation and even then one could find grounds to object to such post-hoc “fishing”.

In an exploratory setting where we are searching for models rather than testing hypotheses, the researcher would retain variables that predicted a significant proportion of the variance and discard those that did not. The researcher could choose to alter the model by adding previously ignored variables such as mean unit load per term or make other adjustments to the analysis. We might also wish to test for interactions or control for confounding variables. Perhaps a variable excluded at one step becomes important in conjunction with a variable added later in our discovery process. A researcher in such a procedure would have to be highly adept at statistics and data manipulation and spend a great deal of time and effort in this search for models and patterns and still might miss a parsimonious model.

The researcher in this example is data mining but is doing so “by hand”. It is the introduction of the computer that makes even this manual level of data mining possible. One might even joke that it was the thought of performing repeated analyses literally by hand that motivated the tradition of the a-priori hypothesis during the pre-computer development of statistics. The advent of the computer and its exponential increase of processing and storage capabilities and reduction in price revolutionized numeric analysis and to date culminates in programs such as data mining software. Rather than laboriously test a few variables at a time and alter iteration parameters, the researcher assigns variables independent (“in”) or dependent (“out”) status, chooses a mining method, and lets the program come up with a set of rules that governs the situation. The set of rules is clear even to those not familiar with statistics. For instance (using imaginary data), a simple model might say:

If transfer GPA > 3.0 , then graduation = yes (confidence = 0.87)

If transfer GPA ≤ 3.0 and

If university GPA > 3.5 , then graduation = yes (confidence = 0.97)

If university GPA ≤ 3.5 , then graduation = no (confidence = 0.77)

This result is easy to obtain and has a clear interpretation. The data mining program has processed a large number of variables and considered them all in the analysis and has tested its own predictions to improve its model and provide empirically based confidence levels.

An analyst might look with some suspicion on such numeric magic (Elder and Pregibon 1996). Many would be wary to trust an analysis that we couldn't replicate by hand or at least by a proven statistical program. However, the mining program is based upon machine learning algorithms similar to our own mind's discovery process of trial and error and derives its mystique from the ability of the computer to calculate more quickly than most humans.

Hosking et al. (1997) point out that in statistics, we often work with fixed conceptual and hypotheses spaces where we select a set of parameters and a model to test our hypotheses. In data mining, we also have a fixed conceptual space but the hypothesis space is left to the learning algorithm, which attempts to create a model with a minimum of prediction error. Data mining also relies less on assumptions about data distributions

and generally results in more complex models judged not on how well they support theory but on how well the model generalizes to new data (Bengio et al. 2000). In general, statistics has been developed for testing simple models where variables of importance are selected by educated judgment or by theory (Table 2). Data mining has been designed to operate on large data sets containing numerous variables with unknown or complex relations. These approaches can also be combined. For example, a linear combination derived from a discriminant analysis can be included as a variable in a data mining effort. Exploratory data analysis (EDA) techniques such as factor analysis blur the distinction between statistics and data mining and help illustrate how these approaches are not by any means mutually exclusive but can help confirm results in some cases. A major difference between EDA and data mining is that EDA, coming from a statistical tradition, often relies on specific distributions and co-variances of variables while data mining techniques, originating in a machine learning and artificial intelligence context, use algorithms designed to seek patterns.

Table 2. Non-exhaustive general comparison of statistics with data mining from the perspective of a social scientist.

Inferential Statistics	Exploratory Statistics	Data Mining
t-test, ANOVA, regression, Chi-square	Factor analysis	neural networks, Kohonen networks, C5.0 rule induction
a priori hypotheses	factors	generated rules
simple models	simple to complex models	complex models
small data sets	large data sets	very large data sets
variables of known importance	variables with unknown relevance	variables with unknown relevance
"familiar" formulae	less "familiar" formulae	"Black Box"
distributional assumptions	distributional assumptions	learning algorithms
significance	significance	accuracy and generalizability

Just like many statistical methods, such as regression, data mining models can be immediately applied to new data for prediction or profiling purposes. It is interesting to note that while statistical methods apply a fixed set of variables to all cases, data mining models can apply different variables to different cases. This is another strength of data mining, for the ultimate reason to carry out pattern identification or rule setting is for the

knowledge gained from this exercise to tell the policy makers what the future holds down to each individual case.

De Veaux (2000) listed five data mining models: Descriptions, Classifications, Regressions, Clustering, and Association. He argued that OLAP (On-line Analytical Processing) and query software, such as BrioQuery, fall in the category of Descriptive. Classifications work with sets of categorical or discrete values while Regression analyzes continuous values. Clustering discovers groupings within data sets and Association elucidates hidden relations among data elements. Some data mining techniques touch on more than one model. Examples of clustering techniques include Kohonen Networks (Kohonen 1984), Kmeans, and Nearest Neighbor. For classification and association through decision trees, some of the most often used are: CART (Classification and Regression Trees) (Breiman et al. 1994), CHAID (Chi-square-Automatic-Interaction-Detection), and GRI (Generalized Rule Induction), and C5.0.

Data mining is a growing field expanding into many disciplines with large and complex data sets including business, ecology, education, cladistics, atmospheric science, astrophysics, and genetic research. New techniques are being developed continuously in response to demands from particular research areas and opportunities offered by improved processing technology. Data mining is also in the process of being harmonized with statistics to provide researchers with a richer and more integrated palate of analysis tools.

Exploring Data Mining in a Higher Education Setting: Three Case Studies

Case Study 1: Early Intervention

To help improve student success and goal attainment, it would be useful to predict those students who will be placed on academic probation. Such a prediction, if sufficiently accurate, would allow for an early intervention strategy to provide additional support services to students at risk. One simple method would be to intervene based upon a student's GPA. A more refined method would include other variables such as demographic information, high school origin, educational background, major, enrollment status, planned work hours per week, number of units attempted and completed, and grade points.

Using GPA alone as a predictor on Spring 1999 data with those below 2.0 receiving intervention and those at or above not receiving intervention resulted in a prediction accuracy of 90% with just less than half being false positives and missing about half of all who should have received intervention. An intervention strategy based on this model would be misspent on half of the students and would miss half who may benefit from intervention. Our goal was to develop a more accurate model based upon data obtained solely from the application and units attempted, units earned, grade points, and GPA. Data from Summer 1992 to Fall 1998 provided a training set for three algorithms, CART, Neural Network, and C5.0. Each model predicted probation status for Spring 1999 data and resulted in accuracies of 92% for CART, 91% for the Neural Network, 92% for C5.0,

91% using Logistic Regression on Spring 1999 data (Table 3). Compared to the GPA trigger, all the machine learning and statistical models were better at reducing false positives but not as good at reducing false negative. By adding a “misclassification cost” for false positives or negative to the C5.0 model, we can adjust somewhat which type of error we favor (Table 4).

Using more complex models to predict probation resulted in an early warning model superior to using GPA alone. The C5.0 model appeared best at capturing those who be placed on probation. The fact that all methods miss about half of those who will be on probation indicates that other unmeasured factors have a strong influence, such as sudden life crises, which are inherently unpredictable. Further, all models had around a quarter of predicted probations as false positive. This suggests that early intervention strategies should be discrete and devoid of any implication student has been predicted to be on probation to avoid a self-fulfilling prophecy occurring for those who would otherwise not fall into probation. The intervention could be as innocuous as a letter from the school that generically describes student support services or more directive such as an requesting that the student speak with a counselor. There seems to be an unavoidable trade off between false positives and false negatives, with the first C5.0 model minimizing false negatives and the second model minimizing false positives. The choice between these trade offs may depend on the intensity of the planned intervention strategy.

Table 3. Accuracy comparison of 5 different prediction models

	Accuracy	False Positive	False Negative
GPA Trigger	90%	45%	49%
CART	92%	23%	57%
Neural Network	91%	31%	57%
C5.0	92%	26%	52%
Logistic Regression, on Spring 1999	91%	26%	66%

Table 4. Effects of adding misclassification costs to the C5.0 model.

C5.0	Accuracy	False Positive	False Negative
No Cost	92%	26%	52%
Cost to False Positive	91%	2%	73%
Cost to False Negative	89%	48%	51%

Case Study 2: Understanding Transfers

Transfer is one of the three missions for community colleges. For many years, transfer reporting has been predominantly a numeration of the total numbers of students moving from community colleges to universities. No meaningful research can be conducted on these transfer students because of lack of unitary data. In 1998, three separate joint partnership agreements resulted in course level transfer student data being provided to Cabrillo College's Office of Planning and Research (PRO). University of California Santa Cruz (UCSC), San Jose State University (SJSU), and California State University Monterey Bay (CSUMB) have provided longitudinal data on the students who have transferred from Cabrillo to their institutions in the past five years. This partnership represents $\frac{2}{3}$ to $\frac{3}{4}$ of our transfers in any given year. The files contained information on their course enrollment, majors, and their graduation information for the purpose of identifying the education outcomes former Cabrillo College students obtained after they left Cabrillo. This type of data has gone beyond simple head counting and can answer many questions that faculty have been asking, such as what are the potential factors influencing transfer outcomes.

A policy question on transfer can be "what measures can a community college take to increase transfer rates?" This policy question is translated into the following research question: what factors to what degree are influencing the student in his/her transfer behavior and based upon this what is the profile of a transfer student regarding the student's course taking pattern, course outcomes, demographics, and his/her educational outcomes in the transfer institution that can, in turn, be used to modify the community college curriculum?

A test set was created using data from all Cabrillo College students enrolled in the 1997-1998 academic year (please refer to the addendum for the five essential steps in building data sets for data mining). The study chose Clementine as the data mining software (please also refer to the addendum for data mining tools). The main data mining technique was C5.0, a decision tree model. The data file is a single occurrence file (one record per student) restricted by term, which was a product of a one to many and many to many relational database querying. No restrictions existed on how many variables there

were to enter into the data mining data file. As an example, some of the variables are listed as follows:

- Total Units Attempted, Units Earned, and Grade Points
- Grade Point Average
- Total Transfer, Vocational, and Basic Skills Courses Taken
- Educational Goal
- Demographics: Age, Gender, Ethnicity, Disability, Education, City, Major
- Financial aid and AFDC status
- Transfer Status in Fall 1998

This is not a complete list of the variables, nor is it considered a large pool of variables, as most data mining software is capable of handling very large numbers of variables at one time.

The research explored what factors are associated with students who either transferred or did not in Fall 1998. Decision Tree C5.0 in Clementine uncovered a long set of rules. Some are highlighted below with confidence measures in parentheses.

For those with less than 71 units attempted, will transfer if

Educational goal is a vocational AA/AS (56%)

If educational goal is AA/AS and receive financial aid (67%)

If educational goal undecided and more than 23 transfer courses (79%)

If educational goal is transfer with AA/AS:

Unknown majors over 20 years old (87%)

If educational goal is to form career interests and more than 2 basic skills courses and receive financial aid (100%)

If GPA is greater than 3.2 (90%)

For those with more than 71 units attempted will transfer if:

Not Native American (100%)

White and either

High school diploma or not and more than 9 transfer courses (100%)

AA/AS or higher and no vocational courses (100%)

The findings point out that a student's stated educational goal was important in predicting transfers, as was the presence of receiving financial aid in some cases. Ethnicity appeared important only for students with a large number of units attempted. This profile indicates to campus marketers, policy makers, and creators of curricula who is currently being served, who needs outreach services, and what factors appear critical for successful transfer outcomes. A college could use such findings to guide counseling and other matriculation efforts to certain groups and to justify financial assistance to students.

We can also compare results from data mining to statistical analyses (Brieman 1994; De'ath and Fabricus 2000; Lim, Loh, & Shih 2000). For example, a Neural Network algorithm identified the following five variables as being most important in predicting transfer status: Major, educational goal, number of transfer courses taken, number of units attempted, and age with major being by far the most important. A stepwise logistic regression entered the variables number of transfer courses, educational goal, units attempted, vocational courses, residency, GPA, age, major, and education in that order. There is much overlap in variables but differences occur in the influence attributed to each possibly due to the interaction of the nature of the variable and the model characteristics.

Note that some variables are noteworthy by their exclusion. Neither gender nor disability appeared in any of these models nor were there significant differences between groups in proportions of transfer status using Chi-square analysis (with the exception that females with transferring as an educational goal were significantly more likely to transfer than males but with a small effect size). This suggests that matriculation efforts to give equal opportunity to transfer to gender and disability groups do not appear to need major adjustment based upon our data set. At this point, it is prudent to remember that these rules and models do not necessarily generalize to other academic years or institutions.

Case Study 3: Determining Sampling Adequacy in a Transportation Survey

In Fall 2000, Cabrillo conducted a transportation survey of staff, faculty, and students. The student sample was randomly selected while an attempt was made to census the employee community. About 58% of full-time faculty and about 24% of all staff responded. To address concerns about the representativeness of staff and faculty responses, demographic comparisons between the respondents and the population were conducted. The case for a representative sample was stronger with staff than with faculty based upon just the demographic data. These differences between respondent and population demographics would be especially of concern if there is some relation between responses and demographic variables. A Kohonen neural network was trained to see if responses could be effectively clustered by gender, ethnicity, or age. None of these demographics appeared to separate respondents into distinct groups. A confirmatory logistic regression showed that neither gender, ethnicity, nor age could significantly predict the outcomes of several key responses. These findings indicate that the shortfall in demographic similarity of respondents to the population at least did not appear to relate strongly to survey responses, increasing the confidence in the generalizability of the results.

In this case, the advantage of the Kohonen network was that it can simultaneously incorporate all responses and demographic variables whether numeric or categorical. Statistical clustering techniques such as K means or Principle Components can only use numeric variables. Repeated logistic regressions to see if demographics could significantly predicted outcomes results in an uncomfortable probability of a spurious result. Multivariate techniques such as Canonical Correlation have assumptions of

multivariate normality that must be accounted for and are unreliable if there are fewer than 5 cases per variable. The disadvantage is that the decision of whether clusters are “distinct enough” and relate to a particular variable is performed by “eyeballing” the resulting cluster map. There is no number to report and little established tradition of how to report such an assertion.

From Data Mining to Knowledge Management

The birth of data mining appears to have completed the road map for research in higher education. From this moment on, researchers have an entire set of tools for their trade on their desktop. However, in a fast changing and technology driven world, one can easily be overwhelmed by a number of factors. For example, in the late ‘80s, terms like Decision Support Systems (DSS) and Executive Information Systems (EIS) became fashionable to describe the process of using data to guide our decisions. However, higher education is not about making decisions as much as knowledge creation and transmission and Decision Support is too closely associated with administration and management, unfriendly terms to some faculty and students. Since higher education is about the creation, transformation and transmission of knowledge (Laudon & Laudon 1999), an appropriate and increasingly widely accepted term is now “Knowledge Management” - the process for which is called “Knowledge Discovery”. With the development in datawarehousing and data mining, the landscape for knowledge management has greatly changed. It is now time to take a bird’s-eye view of the playing field. After extensive research and based on actual experience, a model for managing knowledge for research and planning is proposed to be the Tiered Knowledge Management Model (TKMM) (Figure One). As with any model, it is assumed that the model itself will remain robust and stable, while the specific techniques and technologies may change over time, as they should.

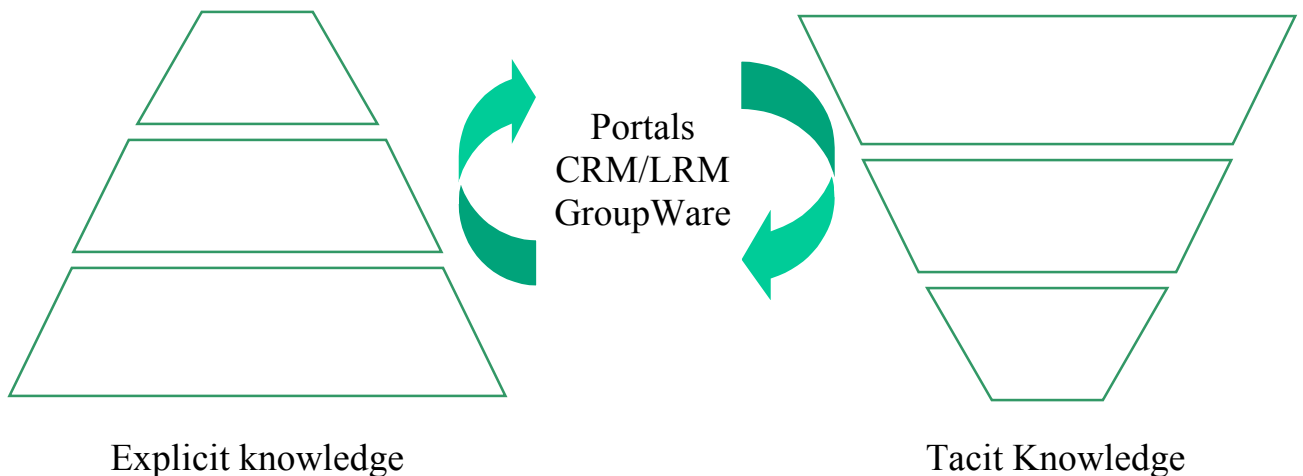


Figure One. Tiered Knowledge Management Model (TKMM):

According to Crowley (2000), explicit knowledge is codified knowledge that is transmittable in formal, systematic language and tacit knowledge is personal, context-

specific and not yet formalized and codified. Tacit knowledge is assumed to be personal in origin, job specific, related to context, difficulty to fully articulate, poorly documented, but highly operational in the minds of the possessor.

Tacit knowledge may be intertwined with explicit knowledge, such as data from surveys where information is cut and diced into categories and ranks. The good news about tacit knowledge is that it can be converted to explicit knowledge. In the book “Knowledge Engineering and Management”, Schreiber et al. (2000), established the technical base for codifying tacit knowledge down to the code (programming) level.

According to Figure One, Knowledge Management is linking explicit knowledge with tacit knowledge via techniques such as portals, CRM/LRM, and groupware. A portal is a gateway, a starting site for users when they get connected to the Web or a place customized by the user that may include a directory of Web sites, a facility to search for other sites, news, weather information, e-mail, stock quotes, phone and map information, and sometimes a community forum.

CRM (customer relationship management) is an information industry term for methodologies, software, and usually Internet capabilities that help an enterprise manage customer relationships in an organized way. For example, an enterprise might build a database about its customers that described relationships in sufficient detail so that management, salespeople, people providing service, and perhaps the customer directly could access information, match customer needs with product plans and offerings, remind customers of service requirements, know what other products a customer had purchased, and so forth (WhatIs.com 2001). We would call a corresponding concept applicable in higher education as Learner Relationship Management (LRM).

First pioneered by major consulting firms, groupware was developed to help consultants work together while located remotely from one another. Lotus Notes dominated the market until the advent of MS Exchange. Groupware services can include the sharing of calendars, collective writing, e-mail handling, shared database access, electronic meetings with each person able to see and display information to others, and other activities.

Knowledge management usually is mentioned in the same sentence with data warehouses and data analysis tools, which are specifically designed for handling explicit knowledge. Explicit knowledge is what exists in codified environments. Therefore, explicit knowledge management is a breakthrough point for KM. Figure Two helps illustrate methods currently available to information professionals, IR in particular.

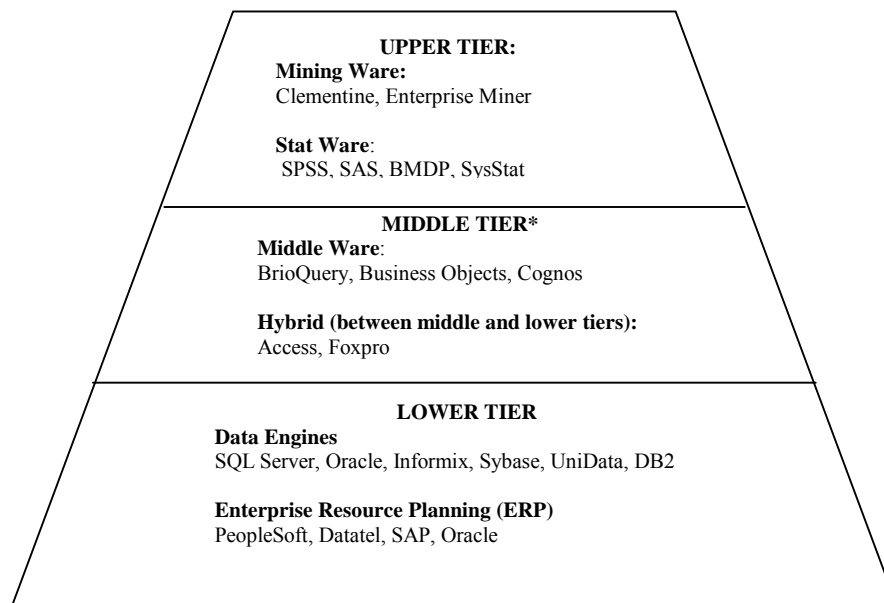


Figure Two. Topography of Tiered Knowledge Management Model (TKMM) for explicit knowledge

* Web based client side (end user, thin client) OLAP software now includes JSP, Dreamweaver, Coldfusion, etc.

Note: this typography for TKMM is for illustrative purpose only. Not all products by all vendors are displayed nor are any products shown necessarily endorsed.

At the bottom of the Lower Tier, Enterprise Resource Planning (ERP) systems, such as Peoplesoft or SAP, are the source for most of the data that researchers need, short of surveys and other supplemental sources. These systems are on-line transaction processing (OLTP) applications that maintain the most scattered and fragmented relational data files. OLTP data should first be prepared by methods referred to as staging and denormalizing before moving into a datawarehouse where SQL servers and their ilk unleash their power. The Middle Tier houses programs that act between the servers and the clients for data querying and reporting. Some call them Middleware, as in a distributed network-computing environment. The Upper Tier of this model points to traditional statistical packages as well as data mining software programs.

What's the Significance of TKMM?

Coming together as a knowledge management model and using the TKMM for Explicit Knowledge alone, TKMM holds significant implications to researchers in the areas of securing funding, updating knowledge, managing the office, and understanding the relationships between research and other technology intensive departments on campuses. A road map like TKMM may help guide the efforts for researchers to update their skills and choose the right tool. Specifically, the implications are in the following areas:

- **Project Management**-This model explains what tool is best for which project, e.g., for real-time query as in census FTES reporting, you need OLAP tools; for building data sets for Statware or Miningware, you need relational database querying tools.
- **Skills Update**-This model describes the relationship of the software programs in each tier and the level of knowledge needed for each, e.g., to be comfortable with the Middle Tier, you should know Javascripting for Brio, web based querying, and SQL (Structured Query Language)
- **Managing the Office**-The model helps you identify on which tier you have the strength and for which tier you need to work with other departments. It helps you determine your SOP (Standard Operating Procedure), e.g., understanding how data is processed into datawarehouses and by whom, and what you should expect from your data processing department.
- **Resource Planning**-This model guides the planning and allocation of resources for data and for research, i.e., using the Total Cost Ownership model (TCO), you can successfully argue what software and hardware to invest in and when to upgrade them.

For research and planning to be useful and for all of us to be successful, the current knowledge base must be updated with new technology and TKMM is a good starting point.

Conclusion

Insights from data sets and variable lists previously seen as unwieldy and chaotic can be drawn out with data mining and developed into the foundations for program planning or to help focus research efforts. For e-commerce, data mining helps with individualized marketing; for the insurance industry, fraud detection; for manufacturing, most efficient approach; and for education, enrollment management. The use of data mining is not just limited to these few examples, but what can be uncovered via a comprehensive and automated data mining process can mean millions of dollars for any given user. A one-percentage point (1 %) change may mean \$500,000 for a typical college of 20,000 students. Data mining conducted for alumni donations may correctly pinpoint the right donors and the right target amount. This both saves campaign costs and increases the campaign achievement. Data mining conducted to predict the likelihood of an applicant's enrollment following their initial application may allow the college to send the right kind of materials to the potential student and prepare the right counseling for him or her. The potential of data mining in education cannot be underestimated.

To successfully engage in knowledge discovery, the TKMM model can help researchers determine what tools best fit for a particular project (Project Management), what skills are a must for surviving in a fast changing world of technology (Skills Update), which tier in the model contains most of the human capital of the research office and on which tier the office needs cooperate with other departments (Managing the Office), and what to buy and when to buy (Resource Planning).

Addendum

Data Preparation and Validation

Data mining software is not built for, or has not evolved into, working directly with relational databases or conducting SQL queries. Referring to the TKMM (Figure One), data mining relies on the work from lower tiers in TKMM. The authors of this paper propose the following five steps to successful data mining. The first four are all related to the steps before data is brought into a data mining environment.

Step One – Move data into a datamart. A researcher may need to first build a datamart to house the unitary relational data files. A datamart will help avoid interfacing with a datawarehouse. Depending on the size of the datawarehouse and the technical skills of the researcher, this may or may not be totally necessary.

Step Two – Querying data for building a flat file. This step calls for skills in SQL commands and familiarity with the four types of joins: basic join, left-outer join, right-outer join, and full-join. There needs to be one occurrence per record with multiple occurrences converted into fields (i.e., recode all courses taken by a students into types with each type occupying a field that holds the number of courses taken within type), or multiple values aggregated (i.e., units counts for courses collapsed in one value). It is strongly recommended that the researcher save the SQL in case there is a need to redo the flat file or slightly modify the SQL for other uses. BrioQuery (*.bqy files), Foxpro (*.qpr files) and even SPSS (*.sps scripts) can be useful, although SPSS works with data transforming only within a flat file.

Step Three – Data visualization. This means both examining frequency counts as well as generating scatter plots, histograms, and other graphics. A graph is the best spokesperson for a correlation estimate. This step gives the researcher the first impression of what each of the data fields contains and how they may play out in the analysis. More often than not, the researcher needs to be very familiar with his or her data elements and this step is a sure way to guarantee that.

Step Four – Data validation. This step may take place simultaneously with Step Three. Depending on the type of query software, some data elements may have been extracted erroneously. For example, Foxpro does not like multiple joins for two files. The second join condition may not run or may bring back a completely wrong set of data all displayed in the correct field with the correct values.

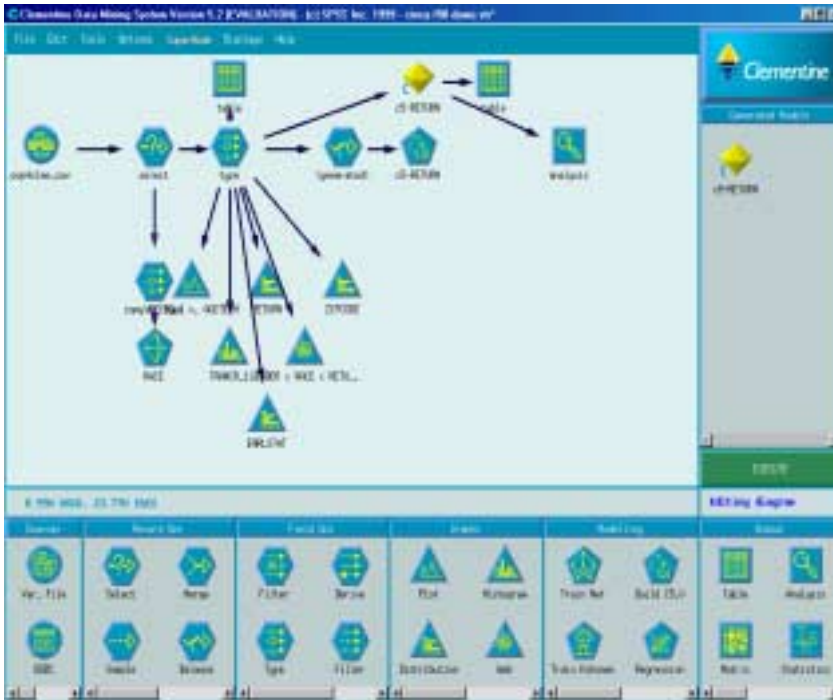
Step Five – Mine your data!

In a simple data mining task for the case study in this primer, a number of issues confronted the researcher even before the data could be brought into Clementine. These issues helped to demonstrate the importance of data preparation. The issues manifested themselves in several areas. They were (1) compiling a data set from one to many and many to many relational databases, (2) converting variables that were not meaningful or far too inclusive, i.e., 27 different ethnicity categories, (3) delineating courses into

transfer, vocational, and basic skill course count from course names, (4) resolving missing values (data imputation), (5) deleting data entry errors or system determined values. Frequency counts also uncovered other minor problems that would have remained unnoticed until outrageously alarming findings came out of the analysis. It took several days to generate the final file while it took only an hour to complete the first run using C5.0. There is truth in the statement that researchers should spend 95% of their time preparing the data. Other lessons learned are the need for carefully defining the research question and the benefit of knowing the data element dictionary (DED) well.

Data Mining Software

Figure Three. Clementine Working Model:



Some in the industry call data mining software packages “toolboxes”, and the modeling techniques “tools”. Others view data mining software as an end-to-end solution due to some software packages ability to interface directly with databases. Clementine (Figure Two), for example, is a data mining package developed by SPSS in the mid ‘90s. In addition, SAS has developed their data

mining toolbox called Enterprise Miner and SGI offers MineSet. Many other packages are available and being currently developed. Some focus on offering machine learning algorithms while others provide novel data visualization tools.

Clementine was used to analyze the example data in this paper. Designed in England, its GUI interface is laid out differently then the typical Windows product. The lower part of the window contains various toolboxes with graphing and analysis functions. The larger space, called a pane, is the workbench. The data analysis is represented by a visual basic programming flow chart called a stream. The data typically flows from a data source node to a data manipulation node to a modeling node. Clementine also can be programmed using scripting language. Algorithms in Clementine include CART (Classification and Regression Trees), C5.0, Kohonen Neural Networks, and Kmeans as well as some traditional statistics such as logistic regression.

References

- Bengio, Y., Buhmann, J.M., Embrechts, M., Zurada, J.M. (2000) Introduction to the special issue on neural networks for data mining and knowledge discovery. *Transactions on Neural Networks*. 11.3: 545-549.
- Breiman L. (1994) Comment. *Statistical Science*, Vol 9, Issue 1, 38-42
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1994) *Classification and regression trees*. Chapman and Hall: New York, New York.
- Corey, M., Abbey, M. et al. (1999) *SQL Server 7 Data Warehousing*. Microsoft: Seattle, WA.
- Crowley, Bill (2000) *Knowledge Management for the Information Professional*. In Srikantaiah and Koenig, eds. *Tacit Knowledge and Quality Assurance: Bridging the Theory-Practice Divide*. Chapter 12. Informational Today Inc.: Medford, New Jersey.
- De'ath, G., & Fabricus, K. E. (2000) *Classification and regression trees: A powerful yet simple technique for ecological data analysis*. *Ecology*. 81.11: 3178-3192.
- De Veaux, R. (2000). *Data Mining What's New, What's Not*. Presentation at the Data Mining Workshop, Long Beach, California.
- Elder IV, J.F. & Pregibon, D. (1996) *A statistical perspective on knowledge discovery in databases*. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, eds. *Advances in knowledge discovery and data mining*. MIT Press: Menlo Park, CA. 83-116.
- Hosking, J.R.M., Pednault, E.P.D., & Sudan, M. (2000) *A statistical perspective on data mining*. *Future Generation Computer Systems*. 13: 117-134.
- Kohonen, T. (1984) *Self-Organization and Associative Memory*. Springer-Verlag: New York, New York.
- Laudon, K., Laudon, J. (1999) *Management Information Systems-organization and technology in the networked enterprise*. Prentice Hall: New Jersey.
- Lim, T., Loh, W., & Shih, Y. (2000) *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*. *Machine Learning*. 40: 203-228.
- Mena, J. (1998) *Data-mining FAQs*. *DM Review*, January 1998.
- Nowlin C., Bliss G., Williams E., (2001) <http://www.WhatIs.com>

Further Reading

Hand, D. (1999) Statistics and Data Mining: Intersecting Disciplines. SIGKDD Explorations. ACM SIGKDD. Volume 1, Issue 1 (page 16).

Luan, J., Holbert A., & Satren M. et al. (1996) Using A Datawarehouse to Conduct Longitudinal Tracking of Watsonville Students. Proceedings at the 35 Annual RP Conference, Berkeley, CA.

Perry, P. (1999) Microsoft SQL Server 7.0 and Brio Enterprise Fuel California Community Colleges' "Student Right-to-Know" Program. Industry Solutions Vol. 4. Microsoft, Redmond, WA

Pickering, C. (2000) They're Watching You. Business 2.0 (page 135 – 136) Feb, 2000.

Schreiber G., Akkermans H., Anjewierden A., et al. (2000) Knowledge Engineering and Management. MIT. The MIT Press. Cambridge, MA.

For statistical, machine learning, and data mining terms:

<http://www.statsoft.com/textbook/stathome.html>

http://www.cs.sfu.ca/people/GradStudents/melli/MD_Terms.html

Cabrillo College Planning and Research Office (PRO)

<http://www.cabrillo.cc.ca.us/oir>