Data Mining:
What It Is
and How It Is Related To
Information Retrieval
And
Text Mining

Lori Bowen Ayre
April 2006

Table of Contents

Introduction

Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. More precisely, the term refers to the application of special algorithms in a process built upon sound principles from numerous disciplines including statistics, artificial intelligence, machine learning, database science, and information retrieval (Han & Kamber, 2001).

Data mining algorithms are utilized in the process of pursuits variously called data mining, knowledge mining, data driven discovery, and deductive learning (Dunham, 2003). Data mining techniques can be performed on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data (Frawley, Piatetsky-Shapiro, & Matheus, 1991; Hearst, 1999; Roddick & Spiliopoulou, 1999; Zaïane, O.R., Han, J., Li, Z., & Hou, J, 1998).

Some areas of specialty have a name such as KDD (knowledge discovery in databases), text mining and Web mining. Most of these specialties utilize the same basic toolset and follow the same basic process and (hopefully) yield the same product – useful knowledge that was not explicitly part of the original data set (Benoît, 2002; Han & Kamber, 2001,Fayyed, Piatetsky-Shapiro, & Smyth, 1996).

What is Data Mining?

Data mining refers to the process of finding interesting patterns in data that are not explicitly part of the data (Witten & Frank, 2005, p. xxiii). The interesting patterns can be used to tell us something new and to make predictions. The process of data mining is composed of several steps including selecting data to analyze, preparing the

data, applying the data mining algorithms, and then interpreting and evaluating the results.   Sometimes the term *data mining* refers to the step in which the data mining algorithms are applied. This has created a fair amount of confusion in the literature. But more often the term is used to refer the entire process of finding and using interesting patterns in data (Benoît, 2002).

The application of data mining techniques was first applied to databases.   A better term for this process is KDD (Knowledge Discovery in Databases). Benoît (2002) offers this definition of KDD (which he refers to as data mining):

> Data mining (DM) is a multistaged process of extracting previously
> unanticipated knowledge from large databases, and applying the results to
> decision making. Data mining tools detect patterns from the data and infer
> associations and rules from them. The extracted information may then be
> applied to prediction or classification models by identifying relations
> within the data records or between databases. Those patterns and rules can
> then guide decision making and forecast the effects of those decisions.
> (p. 265)

Today, data mining usually refers to the process broadly described by Benoît (2002) but without the restriction to databases.   It is a "multidisciplinary field drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization. (Han & Kamber, 2001, p. xix).

Data mining techniques can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or Web-based data and complex objects. A more appropriate term for describing the entire process would be *knowledge discovery,* but unfortunately the term *data mining* is what has caught on (Andrássoyá & Paralič, 1999).

Developmental History of Data Mining and Knowledge Discovery

The building blocks of today's data mining techniques date back to the 1950s when the work of mathematicians, logicians, and computer scientists combined to create artificial intelligence (AI) and machine learning (Buchanan, 2006.).

In the 1960s, AI and statistics practitioners developed new algorithms such as regression analysis, maximum likelihood estimates, neural networks, bias reduction, and linear models of classification (Dunham, 2003, p. 13). The term "data mining" was coined during this decade, but the term was pejoratively used to describe the practice of wading through data and finding patterns that had no statistical significance (Fayyad, et al., 1996, p. 40).

Also in the 1960s, the field of information retrieval (IR) made its contribution in the form of clustering techniques and similarity measures. At the time these techniques were applied to text documents, but they would later be utilized when mining data in databases and other large, distributed data sets (Dunham, 2003, p. 13). Database systems focus on query and transaction processing of structured data, whereas information retrieval is concerned with the organization and retrieval of information from a large

number of text-based documents (Han & Kamber, 2001, p. 428). By the end of the 1960s, information retrieval and database systems were developing in parallel.

In 1971, Gerard Salton published his groundbreaking work on the SMART Information Retrieval System. This represented a new approach to information retrieval which utilized the algebra-based vector space model (VSM). VSM models would prove to be a key ingredient in the data mining toolkit (Dunham, 2003, p. 13).

Throughout the 1970s, 1980s, and 1990s, the confluence of disciplines (AI, IR, statistics, and database systems) plus the availability of fast microcomputers opened up a world of possibilities for retrieving and analyzing data. During this time new programming languages were developed and new computing techniques were developed including genetic algorithms, EM algorithms, K-Means clustering, and decision tree algorithms (Dunham, 2003, p. 13).

By the start of the 1990s, the term Knowledge Discovery in Databases (KDD) had been coined and the first KDD workshop held (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 40). The huge volume of data available created the need for new techniques for handling massive quantities of information, much of which was located in huge databases.

The 1990s saw the development of database warehouses, a term used to describe a large database (composed of a single schema), created from the consolidation of operational and transactional database data. Along with the development of data warehouses came online analytical processing (OLAP), decision support systems, data scrubbing/staging (transformation), and association rule algorithms (Dunham, 2003, p. 13, 35-39; Han & Kamber, 2001, p. 3).

During the 1990s, data mining changed from being an interesting new technology to becoming part of standard business practice. This occurred because the cost of computer disk storage went down, processing power went up, and the benefits of data mining became more apparent.   Businesses began using "data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers" (Two Crows, 1999, p. 5).

Data mining is used by a wide variety of industries and sectors including retail, medical, telecommunications, scientific, financial, pharmaceutical, marketing, Internet-based companies, and the government (Fayyad, et al., 1996). In a May, 2004 report on Federal data mining activities, the U.S. General Accounting Office (GAO, 2004) reported there were 199 data mining operations underway or planned in various federal agencies (p. 3), and this list doesn't include the *secret* data mining activities such as MATRIX and the NSA's eavesdropping (Schneier, 2006).

Web mining is an area of much research and development activity. There are many factors that drive this activity including online companies who wish to learn more about their customers and potential customers, governmental agents tasked with locating terrorists and optimizing services, and the user need for filtered information.

Theoretical Principles

The underlying principle of data mining is that there are hidden but useful patterns inside data and these patterns can be used to infer rules that allow for the prediction of future results (GAO, 2004, p. 4).

Data mining as a discipline has developed in response to the human need to make sense of the sea of data that engulfs us. Per Dunham (2003), data doubles each year and

yet the amount of useful information available to us is decreasing (p. xi). The goal of data mining is to identify and make use of the "golden nuggets" (Han & Kamber, 2001, p. 4) floating in the sea of data.

Prior to 1960 and the dawn of the computer age, a data analyst was an individual with expert knowledge (domain expert) and training in statistics. His job was to cull through the raw data and find patterns, make extrapolations, and locate interesting information which he then conveyed via written reports, graphs and charts. But today, the task is too complicated for a single expert (Fayyad, et al., 1996, p. 37). Information is distributed across multiple platforms and stored in a wide variety of formats, some of which are structured and some unstructured. Data repositories are often incomplete. Sometimes the data is continuous and other times discrete. But always the amount of data to be analyzed is enormous.

KDD involves searching large databases, but it distinguishes itself from database *querying* in that it seeks implicit patterns in the data rather than simply extracting selections from the database. Per Benoît (2002), the database query answers the question "what company purchased over $100,000 worth of widgets last year?" (p. 270) whereas data mining answers the question "what company is likely to purchase over $100,000 worth of widgets next year and why?" (p. 270).

All forms of data mining (KDD included) operate on the principle that we can learn something new from the data by applying certain algorithms to it to find patterns and to create models which we then use to make predictions, or to find new data relationships (Benoît, 2002; Fayyad, et al., 1996; Hearst, 2003).

Another important principle of data mining is the importance of presenting the patterns in an understandable way. Recall that the final step in the KDD process is presentation and interpretation. Once patterns have been identified, they must be conveyed to the end user in a way that allows the user to act on them and to provide feedback to the system. Pie charts, decision trees, data cubes, crosstabs, and concept hierarchies are commonly used presentation tools that effectively convey the discovered patterns to a wide variety of users (Han & Kamber, 2001, pp. 157-158).

Technological Elements of Data Mining

Because of the inconsistent use of terminology, data mining can both be called a step in the knowledge discovery process or be generalized to refer to the larger process of knowledge discovery.

Steps in Knowledge Discovery

Table 1 (Andrássoyá & Paralič, 1999, Section 2.2) compares the primary steps in knowledge discovery as presented by different authors. The table helps us understand the basic steps in the knowledge discovery process and where the specific application of data mining fits into the larger picture.

Simoudis [12]	Mannila [10]	Fayyad et al. [5]	Brachman & Anand [1]
	understanding the domain	learning the application domain	task discovery
data selection		creating a target dataset	data discovery
data transformation	preparing the data set	data cleaning and preprocessing	data cleaning
		data reduction and projection	model development
		choosing the function of data mining	
Data mining	discovering patterns (data mining)	choosing the data mining algorithm(s)	data analysis
		data mining	
result interpretation	postprocessing of discovered patterns	interpretation	output generation
	putting the results into use	using discovered knowledge	

Table 1: List of Knowledge Discovery Steps (Andrássoyá & Paralič,1999, Section 2.2)

Step 1: Task Discovery

The goals of the data mining operation must be well understood before the process

begins: The analyst must know what the problem to be solved is and what the questions

that need answers are. Typically, a subject specialist works with the data analyst to refine

the problem to be solved as part of the task discovery step (Benoît, 2002).

Step 2: Data Discovery

In this stage, the analyst and the end user determine what data they need to analyze in order to answer their questions, and then they explore the available data to see if what they need is available (Benoît, 2002).

Step 3: Data Selection and Cleaning

Once data has been selected, it will need to be cleaned up: missing values must be handled in a consistent way such as eliminating incomplete records, manually filling them in, entering a constant for each missing value, or estimating a value. Other data records may be complete but wrong (noisy). These noisy elements must be handled in a consistent way (Benoît, 2002; Fayyad, et al., 1996).

Step 4: Data Transformation

Next, the data will be transformed into a form appropriate for mining. Per Weiss, Indurkhya, Zhang & Damerau (2005), "data mining methods expect a highly structured format for data, necessitating extensive data preparation. Either we have to transform the original data, or the data are supplied in a highly structured format" (p. 1).

The process of data transformation might include smoothing (e.g. using bin means to replace data errors), aggregation (e.g. viewing monthly data rather than daily), generalization (e.g. defining people as young, middle-aged, or old instead of by their exact age), normalization (scaling the data inside a fixed range), and attribute construction (adding new attributes to the data set, Han & Kamber, 2001, p. 114).

Step 5: Data Reduction

The data will probably need to be reduced in order to make the analysis process manageable and cost-efficient. Data reduction techniques include data cube aggregation, dimension reduction (irrelevant or redundant attributes are removed), data compression (data is encoded to reduce the size, numerosity reduction (models or samples are used instead of the actual data), and discretization and concept hierarchy generation (attributes are replaced by some kind of higher level construct, Han & Kamber, 2001, pp. 116-117).

Step 6: Discovering Patterns (aka Data Mining)

In this stage, the data is iteratively run through the data mining algorithms (see Data Mining Methods below) in an effort to find interesting and useful patterns or relationships. Often, classification and clustering algorithms are used first so that association rules can be applied (Benoît, 2002, p. 278).

Some rules yield patterns that are more interesting than others. This "interestingness" is one of the measures used to determine the effectiveness of the particular algorithm (Fayyad, et al.,1996; Freitas, 1999; Han & Kamber, 2001).

Fayyad, et al. (1996) states that interestingness is "usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity" (p. 41). A pattern can be considered knowledge if it exceeds an interestingness threshold. That threshold is defined by the user, is domain specific, and "is determined by whatever functions and thresholds the user chooses" (p. 41).

Step 7: Result Interpretation and Visualization

It is important that the output from the data mining step can be "readily absorbed and accepted by the people who will use the results" (Benoît, p. 272). Tools from computer graphics and graphics design are used to present and visualize the mined output.

Step 8: Putting the Knowledge to Use

Finally, the end user must make use of the output. In addition to solving the original problem, the new knowledge can also be incorporated into new models, and the entire knowledge or data mining cycle can begin again.

Data Mining Methods

Common data mining methods include classification, regression, clustering, summarization, dependency modeling, and change and deviation detection. (Fayyad, et al., 1996, pp. 44-45)

Classification

Classification is composed of two steps: supervised learning of a training set of data to create a model, and then classifying the data according to the model. Some well-known classification algorithms include Bayesian Classification (based on Bayes Theorem), decision trees, neural networks and backpropagation (based on neural networks), k-nearest neighbor classifers (based on learning by analogy), and genetic algorithms. (Benoît, 2002; Dunham, 2003).

Decision trees are a popular top-down approach to classification that divides the data into leaf and node divisions until the entire set has been analyzed. *Neural networks* are nonlinear predictive tools that learn from a prepared data set and are then applied to new, larger sets. *Genetic algorithms* are like neural networks but incorporate natural selection and mutation. *Nearest neighbor* utilizes a training set of data to measure the similarity of a group and then use the resultant information to analyze the test data. (Benoît, 2002, pp. 279-280)

Regression

Regression analysis is used to make predictions based on existing data by applying formulas. Using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions. (Dunham, 2003, p. 6) *Regression trees,* decision trees with averaged values at the leaves, are a common regression technique. (Witten & Frank, 2005, p. 76)

Clustering

Clustering involves identifying a finite set of categories (clusters) to describe the data. The clusters can be mutually exclusive, hierarchical or overlapping. (Fayyad, et al., 1996, p. 44). Each member of a cluster should be very similar to other members in its cluster and dissimilar to other clusters. Techniques for creating clusters include partitioning (often using the k-means algorithm) and hierarchical methods (which group objects into a tree of clusters), as well as grid, model, and density-based methods. (Han & Kamber, 2001, p. 346-348)

Outlier analysis is a form of cluster analysis that focuses on the items that don't fit neatly into other clusters (Han & Kamber, 2001). Sometimes these objects represent errors in the data, and other times they represent the most interesting pattern of all. Freitas (1999) focuses on outliers in his discussion of *attribute surprisingness* and suggests that another criterion for interestingness measures should be surprisingness.

Summarization

Summarization maps data into subsets and then applies a compact description for that subset. Also called *characterization* or *generalization*, it derives summary data from the data or extracts actual portions of the data which "succinctly characterize the contents" (Dunham, 2003, p. 8).

Dependency Modeling (Association Rule Mining)

Dependency or Association Rule Mining involves searching for interesting relationships between items in a data set. Market basket analysis is a good example of this model. An example of an association rule is "customers who buy computers tend to also buy financial software" (Han & Kamber, 2001, pp. 226-117). Since association rules are not always interesting or useful, constraints are applied which specify the type of knowledge to be mined such as specific dates of interest, thresholds on statistical measures (rule interestingness, support, confidence), or other rules applied by end users (Han & Kamber, 2001, pp. 262).

Change and Deviation Detection

Also called *sequential analysis and sequence discovery* (Dunham, 203, p. 9), change and deviation detection focuses on discovering the most significant changes in

data. This involves establishing normative values and then evaluating new data against the baseline (Fayyad, et al., 1996, p. 45). Relationships based on time are discovered in the data.

The above methods form the basis for most data mining activities. Many variations on the basic approaches described above can be found in the literature including algorithms specifically modified to apply to spatial data, temporal data mining, multi-dimensional databases, text databases and the Web (Dunham, 2003; Han & Kamber, 2001).

Related Disciplines: Information Retrieval and Text Mining

Two disciplines closely related to data mining are information retrieval and text mining. The relationship between information retrieval and data mining techniques has been complementary. Text mining, however, represents a new discipline arising from the combination of information retrieval and data mining.

Information Retrieval (IR)

Many of the techniques used in data mining come from Information Retrieval (IR), but data mining goes beyond information retrieval. IR is concerned with the process of searching and retrieving information that exists in text-based collections (Dunham, 2003, p. 26). Data mining, on the other hand, is not concerned with retrieving data that exists in the repository. Instead, data mining is concerned with patterns that can be found that will tell us something new – something that isn't explicitly in the data (Han & Kamber, 2001).

IR techniques are applied to text-based collections (Baeza-Yates & Ribeiro-Neto, 1999). Data mining techniques can be applied to text documents as well as databases (KDD), Web based content and metadata, and complex data such as GIS data and temporal data.

In terms of evaluating effectiveness, IR and data mining system markedly differ. Per Dunham (2003, p. 26), the effectiveness of an IR system is based on precision and recall and can be represented by the following formulas:

$$\text{Precision} = \frac{\text{Relevant and Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant and Retrieved}}{\text{Relevant}}$$

The effectiveness of any knowledge discovery system is whether or not any useful or interesting information (knowledge) has been discovered. Usefulness and interestingness measures are much more subjective than IR measures (precision and recall).

IR Contributions to Data Mining

Many of the techniques developed in IR have been incorporated into data mining methods including Vector Space Models, Term Discrimination Values, Inverse Document Frequency, Term Frequency-Inverse Document Frequency, and Latent Semantic Indexing.

Vector Space Models, or *vector space information retrieval systems*, represent documents as vectors in a vector space (Howland & Park, 2003, p. 3; Kobayashi & Aono, 2003, p. 105). *Term Discrimination Value* posits that a good discriminating term

is one that, when added to the vector space, increases the distances between documents (vectors). Terms that appear in 1%-10% of documents tend to be good discriminators (Senellart & Blondel, 2003, p. 28). *Inverse Document Frequency* (IDF) is used to measure similarity. IDF is used in data mining methods including clustering and classification (Dunham, 2003, pp. 26-27). *Term Frequency-Inverse Document Frequency* (TF-IDF) is an IR algorithm based on the idea that terms that appear often in a document and do not appear in many documents are more important and should be weighted accordingly (Senellart & Blondel, 2003, p. 28). *Latent Semantic Indexing* (LSI) is a dimensional reduction process based on *Singular Value Decomposition* (SVD). It can be used to reduce noise in the database and help overcome synonymy and polysemy problems (Kobayashi & Aono, 2003, p. 107).

Data Mining Contributions to IR

Although IR cannot utilize all the tools developed for data mining because IR is generally limited to unstructured documents, it has nonetheless benefited from advances in data mining. Han and Kamber (2001) describe *Document Classification Analysis* which involves developing models which are then applied to other documents to automatically classify documents. The process includes creating keywords and terms using standard information retrieval techniques such as TF-IDF and then applying association techniques from data mining disciplines to build concept hierarchies and classes of documents which can be used to automatically classify subsequent documents (p. 434).

The data mining idea of creating a model instead of directly searching the original data can be applied to IR. Kobayashi & Aono (2003) describe using Principle

Component Analysis (PCA) and Covariance Matrix Analysis (COV) to map an IR

problem to a "subspace spanned by a subset of the principal components" (p. 108).

Text Mining

Text mining (TM) is related to information retrieval insofar as it is limited to text.

Yet it is related to data mining in that it goes beyond search and retrieval. Witten and

Frank (2005) explain that the information to be extracted in text mining is not hidden;

however, it is unknown because in its text form it is not amenable to automatic

processing. Some of the methods used in text mining are essentially the same methods

used in data mining. However, one of the first steps in text mining is to convert text

documents to numerical representations which then allows for the use of standard data

mining methods (Weiss, Indurkhya, Zhang & Damerau, 2005).

Per Weiss, et al. (2005), "one of the main themes supporting text mining is the

transformation of text into numerical data, so although the initial presentation is different,

at some intermediate stage, the data move into a classical data-mining encoding. The

unstructured data becomes structured" (pp. 3-4).

Weiss, et al (2005) use the spreadsheet analogy as the classical data mining model

for structured data. Each cell contains a numerical value that is one of two types:

ordered numerical or categorical. Income and cost are examples of ordered numerical

attributes. Categorical attributes are codes or true or false. In text mining, the idea is to

convert the text presented as a document to values presented in one row of a spreadsheet

where each row represents a document and the columns contain words found in one or

more documents. The values inside the spreadsheet can then be defined (categorically) as

present (this word is in this document) or absent (this word is not in this document). The spreadsheet represents the entire set of documents or corpus.

The collection of unique words found in the entire document collection represents the dictionary and will likely be a very large set. However, many of the cells in the spreadsheet will be empty (not present). An empty cell in a data mining operation might pose a problem, as it would be interpreted as an incomplete record. However, in text mining, this sparseness of data works to reduce the processing requirements because only cells containing information need to be analyzed. The result is that the size of the spreadsheet is enormous but it is mostly empty. This "allows text mining programs to operate in what would be considered huge dimensions for regular data-mining applications" (Weiss, et al., 2005, p. 5).

Per Weiss, et al. (2005), the process of getting the text ready for text mining is very much like the knowledge discovery steps described earlier. In text mining, the text is usually converted first to XML format for consistency. It is then converted to a series of tokens (sometimes punctuation is interpreted as a token, sometimes as a delimiter). Then, some form of stemming is applied to the tokens to create the standardized dictionary. Familiar IR/data mining processes such as TF-IDF can be applied to assign different weights to the tokens. Once this has been done, classification and clustering algorithms are applied.

Depending on the goal of the text mining operation, it may or may not be important to incorporate linguistic processing in the text mining process. Examples of linguistic processing include marking certain types of words (part-of-speech tagging),

clarifying the meaning of words (disambiguation) and parsing sentences. Per Benoît (2002),

> Text mining brings researchers closer to computational linguistics, as it tends to be highly focused on natural language elements in texts (Knight, 1999). This means TM applications (Church & Rau, 1995) discover knowledge through automatic content summarization (Kan & McKeown, 1999), content searching, document categorization, and lexical, grammatical, semantic, and linguistic analysis (Mattison, 1999). (p. 291)

Conclusion

Data mining is a synonym for knowledge discovery. Data mining also refers to a specific step in the knowledge discovery process, a process that focuses on the application of specific algorithms used to identify interesting patterns in the data repository.   These patterns are then conveyed to an end user who converts these patterns into useful knowledge and makes use of that knowledge.

Data mining has evolved out of the need to make sense of huge quantities of information. Usama M. Fayyad says that stored data is doubling every nine months and the "demand for data mining and reduction tools increase exponentially (Fayyad, Piatetsky-Shapiro, & Uthurusamy, 2003, p. 192)."   In 2006, $6 billion in text and data mining activities are anticipated (Zanasi, Brebbia, & Ebecken, 2005).

The U.S. government is involved in many data mining initiatives aimed at improving services, detecting fraud and waste, and detecting terrorist activities. One such activity, the work of Able Danger, had identified one of the men who would, one year later, participate in the 9/11 attacks (Waterman, 2005). This fact emphasizes the

importance of the final step of the knowledge discovery process: putting the knowledge to use.

The U.S. government's data mining activities have helped stir concerns about data mining and their impact on privacy. Privacy preserving data mining has only recently caught the attention of researchers (Verykios, Bertino, Fovino, Provenza, Saygin & Theodoridis, 2004).

There is much work to done in the area of knowledge discovery and data mining, and its future depends on developing tools and techniques that yield useful knowledge without causing undue threats to individuals' privacy.

References

Andrássoyá, E., & Paralič, J.. (1999, September). Knowledge discovery in databases - a comparison of different views. Presented at the 10th International Conference on Information and Intelligent Systems, Sept. 1999, Varazdin, Croatia.

Baeza-Yates, & R., Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Benoît, Gerald. (2002). Data Mining [Chapter 6, pps 265-310). In Cronin, B. (Ed.), *Annual Review of Information Science and Technology: Vol. 36 (pp. 265-310).* Silver Spring, MD: American Society for Information Science and Technology.

Buchanan, B.G. (2006). Brief History of Artificial Intelligence. Retrieved March 22, 2006 from http://www.aaai.org/AITopics/bbhist.html

Dunham, M.H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

Fayyad, U.M, Piatetsky-Shapiro, G., & Smyth, P. (1996, Fall). From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp. 37-54.

Fayyad, U.M., Piatetsky-Shapiro, G., Uthurusamy, R. (2003). Summary from the KDD-03 Panel – Data mining: The next 10 years. *SIGKDD Explorations, 5*(2). Retrieved March 22, 2006 from ACM Digital Library database.

Frawley, W.J., Piatetsky-Shapiro, G., & Matheus, C.J. (1991). Knowledge Discovery in Databases: An Overview. In Piatetsky-Shapiro, G. & Frawley, W.J. (Eds.), Knowledge discovery in databases (pp. 1-27). Cambridge, MA: AAAI Press/MIT Press.

Freitas, A.A. (1999). On rule interestingness measures. Knowledge-Based Systems, 12, 309-315. Retrieved February 28, 2005 from Elsevier database.

General Accounting Office. (2004, May 4). Data Mining: Federal Efforts Cover a Wide Range of Uses (GAO-04-548). Retrieved February 4, 2006 from http://www.gao.gov/new.items/d04548.pdf

Han, J., & Kamber, M. (2001). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems). San Diego: Academic Press.

Hearst, M. (1999, June). *Untangling Text Data Mining*. Presentation at the 37th Annual Meeting of the Association of Computational Linguistics, University of Maryland, MD.

Hearst, M. (2003). What is text mining? Retrieved from
    http://www.sims.berkeley.edu/~hearst/text-mining.html

Howland, P. & Park, H. (2003) Cluster-Preserving Dimension Reduction Methods for
    Efficient Classification of Text Data. In Survey of text mining: clustering,
    classification, and retrieval. New York: Springer Science+Business Media, Inc.

Kobayashi, M. & Aono, M. (2003) Vector space models for search and cluster mining.
    In Survey of text mining: clustering, classification, and retrieval. New York:
    Springer Science+Business Media, Inc.

Roddick, J.F., & Spiliopoulou, M. (1999, June). A bibliography of temporal, spatial and
    spatio-temporal data mining research. *ACM SIGKDD Explorations Newsletter.*
    Retrieved March 22, 2006 from ACM Digital Library.

Schneier, B. (2005, March 9). Why data mining won't stop terror. *Wired News.*
    Retrieved March 22, 2006 from http://www.wired.com/news/columns/0,70357-
    0.html

Senellart, P.P., & Blondel, V.D. (2003). Automatic Discovery of Similar Words. In
    Berry, M.W. (Ed.). In Survey of text mining: clustering, classification, and
    retrieval. New York: Springer Science+Business Media, Inc.

Two Crows. (1999) About Data Mining [Third Edition]. Retrieved February 7, 2006
    from http://www.twocrows.com/about-dm.htm.

Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., & Theodoridis, Y.
    (2004). State-of-the-art in privacy preserving data mining. *SIGMOD Record,
    33*(1). Retrieved March 22, 2006 from ACM Digital Library.

Waterman, S. (2005, September 20). Probing Able Danger [editorial]. The Washington
    Times [online version]. Retrieved January 20, 2006 from NewsBank database.

Weiss, S.M., Indurkha, N., Zhang, T., Damerau, F.J. (2005). Text mining: Predictive
    methods for analyzing unstructured information. New York: Springer
    Science+Business Media, Inc.

Witten, I.H., Frank, E. (2005). Data mining: practical machine learning tools and
    techniques(2nd ed, Morgan-Kaufman Series of Data Management Systems). San
    Francisco: Elsevier.

Zaïane, O.R., Han, J., Li, Z., Hou, J. (1998). Mining multimedia data. Proceedings of the
    1998 conference of the Centre for Advanced Studies on Collaborative Research,
    Toronto, Ontario, Canada. Retrieved March 22, 2006 from ACM Digital
    Library.

Zanasi, A., Brebbia, C.A., Ebecken, N.F.F. (2005). Preface. In Zanasi, A., Brebbia, C.A., Ebecken, N.F.F.(Eds.), *Sixth International Conference on Data Mining: Data Mining VI*. Southampton, England: WIT Press.