Data Mining

Gerald Benoit

benoit@simmons.edu

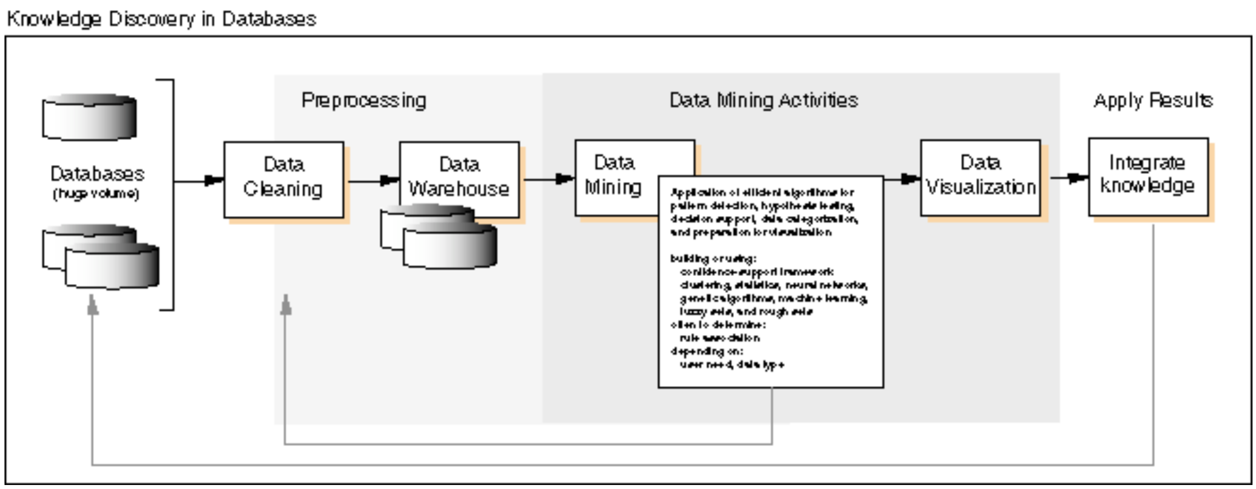
[This is the preprint version of the book chapter from the
Annual Review of Information Science & Technology, 2003]

Introduction

Data mining (DM) is a multi-staged 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.

However, this definition may be applied equally to "knowledge discovery in databases" (KDD). Indeed, in the recent literature of DM and KDD, a confusion emerged making it difficult to determine exactly the parameters of both. KDD is sometimes viewed as the broader discipline, of which data mining is merely a component, specifically pattern extraction, evaluation, and cleansing methods (Raghavan, Deogun and Sever 1998, 397). Thurasingham (1999, 2) remarked that "knowledge discovery", "pattern discovery", "data dredging", "information extraction", and "knowledge mining" are all employed as synonyms for DM. Trybula in the ARIST chapter on text mining observed that the "existing work [in KDD] is confusing because the terminology is inconsistent and poorly defined. Because terms are misapplied even among researchers, it is doubtful that the general public can be expected to understand the topic" (1999, 3). Today the terms are often used interchangeably or without distinction, which as Reinartz (1999, 2) notes, results in a labyrinth.

This review takes the perspective that KDD is the larger view of the entire process, with DM emphasizing the cleaning, warehousing, mining, and interactive visualization of knowledge discovery in databases. Following Brachman et al. (1996, 42), DM in this chapter is considered to be the core function of KDD, whose techniques are used for verification "in which the system is limited to verifying a user's hypothesis" as well as for discovery, in which the system finds new, interesting patterns. The term thus includes the specific processes, computer technology, and algorithms for converting very large databases of structured, semi-structured and full-text sources, into practical, validated knowledge to achieve some user- or application-specific goal. Figure 1 demonstrates the KDD/DM relationship.



Perhaps because of the confusion of the term, DM itself has evolved into an almost independent activity, from one professional meeting in 1995 to over ten in 1998 (Piatetsky-Shapiro 1998). This evolution has sparked considerable investigation into its future (Grossman, Kasif, Moore, Rocke and Ullman 1999; Beeri and Buneman 1999; Gunopulos and Rastogi 2000; Madria, Bhowmick, Ng, and Lim 1999; Raghavan, Deogun and Sever 1998) and research by many academic disciplines into specific DM activities. The impetus comes primarily from the increased volume of data, an expanded user base, and responses by researchers to opportunities in computer technology. For example, in the past decade scientific computing (Fayyad, Haussler and Stolorz 1996), such as genomic, geospatial, and medical research, commonly amass exceptionally large (10^8-10^{12} bytes) stores of high dimensional data (10^2-10^4 data fields) (Ho 2000). Such volume cannot be processed efficiently by most computer environments (Grossman et al. 1999). Therefore questions arise about how to scale and integrate computer systems (Nahm and Mooney 2000; Sarawagi, Thomas and Agrawal 1998; Guo and Grossman 1999) to manage the volume and how to adjust DM algorithms to work efficiently on different system architectures. The volume of minable data surpasses also human capacity to extract meaningful patterns without aid.

DM's evolution is also pressured by a shift in user population from statisticians to individual and domain-specific miners (Ankerst, Ester and Kriegel 2000; Baralis and Psaila 1999). Traditionally, a subject-specialist works with a data analyst in mining closed stores of historical data, suitable to structured, homogeneous databases (Savasere, Omiecinski and Navathe 1995; Toivonen 1996; Agrawal, Imielinski and Swami 1993). Today, end-users increasily are domain-specialists working without an analyst, and may mine structured databases as well as weakly typed, tagged and full-text sources. The emerging form of mixed format mining (Mitchell 1999) overlaps some natural language processing, information retrieval (Robertson and Gaizauskas 1997), and Internet-based records, which further confuses DM's activities in relation to the kin, text mining and information extraction (Wilks, Guthrie and Slator 1996).

Given these developments, this review identifies four as critical challenges to DM's future: data issues, algorithm design, the end-user, and computer architecture. Being an interdisciplinary approach to automating pattern discovery, data mining looks for answers from allied research: machine learning (Michalski, Bratko and Kubat 1998; Mitchell 1999; Langley, Iba and Thompson 1990; Weiss and Kulikowski 1991), artificial intelligence, database design, decision science, high performance computing (Stolorz and Musick 1998; Freitas 1998), inductive logic programming (Thuraisingham 1999), fuzzy logic (Loshin 2000; Pedrycz and Smith 1999), statistics (Bock and Diday 2000; Glymour, Madigan, Pregibon and Smyth 1996), and hybrid investigations.

Scope & Limitations

Readers of *ARIST* are familiar with some of the main DM methods applied to structured data (Trybula 1997) and text mining (Trybula 1999). It is assumed that the reader is familiar with issues in information science, but may not be aware of the variety and depth of activities that overlap DM from other fields. This chapter defines and discusses data mining processes in some detail, perhaps more than is typical in a review article, in order to situate the novelty and currency of some techniques applied to DM problems. The extended introduction to DM processes should sensitize the reader to the many ways they are described in the literature and suggest why research is pursued in cognate fields. Naturally, the breadth of research, practice, and problems facing DM make an exhaustive review of all work and all areas unadvisable. Some important topics, such as continuous vs. discrete data, the missing values, and over-fitting of data, can be considered only briefly. Other issues, such as Kohonan artificial neural networks (Goodacre 2000), deformable Markov model templates (Ge and Smyth 2000), mining high-speed data streams (Domingos and Hulten 2000), vector machines (DeCoste and Wagstaff 2000), temporal (Bettini 2000) and geospatial information data mining (Hyland, Clifton and Holland 1999), must be left aside entirely. Nevertheless, this chapter offers a synoptic review of the major challenges facing DM and the research responses.

Works were examined from artificial intelligence, machine learning, statistics, database theory, professional computing journals as well as subject-specific work applying DM methods, such as medicine. The formats of the materials include primarily English language monographs, serials, conference proceedings, on-line library catalogues, and the Internet. This chapter first offers a synoptic view of DM practice in order to situate the challenges and responses. It then discusses the specific issues related to data, algorithm design, end-users, and data mining architectures, including the Internet and related text-mining activities.

Data Mining Processes

Readers interested in general overviews to DM are fortunate to have in print recently published monographs: Adriaans and Zantinge (1997), Berry and Linoff (1997; 2000), Berson and Smith (1997), Bigus (1996; 1998), Bramer (1999), Cabena, Hadjinian, Stadler, Verhees and Zanasi (1998), Cios, Pedrycz and Swiniarski (1998), Devlin (1996), Groth (1998, 2000), Han and Kamber (2000), Inmon (1996), Kennedy (1997), Pyle (1999), Reinartz (1999), Thuraisingham (1999), Weiss and Indurkhya

(1998), Westphal and Blaxton (1998), and Witten and Frank (2000). On the Internet, Ho (2000) is a thorough introduction to the field.

If our information needs were satisfied only by the discovery of known entities through querying of structured databases, then there would be no need for mining the data. The purpose of data mining is to explore databases for the unknown, by exposing patterns from the data that are novel (or "determining their interestingness" (Freitas 1999, 309)), supporting these patterns through statistical evidence, and presenting these results to the user via a graphic interface that facilitates investigation and interpretation to guide or support actions. To achieve these goals, DM relies on sophisticated mathematical and statistical models, and substantial computing power, to help users convert algorithmic behavior to human-understandable rules for action. For example, a pharmaceutical company develops a new drug that it wants to market. With no information about to whom to market the drug, the company turns to sales records, as evidence of past purchasing behavior, to discover which clients might be interested in the new product. Such data may be stored in a relational database, but standard SQL queries are unproductive. The firm may query the database for "which distributors in the Boston area purchased beta blockers?" but not "which distributors in the Boston area are likely to purchase this new drug and why?". DM assists in the automated discovery of patterns and may establish association rules to be interpreted by the end-user: "If a company distributes beta blocker *x* and has sales of over *$y* per year in the Boston area, the likelihood of that company purchasing the new drug is *z%*".

This same firm may have a research arm that generates technical reports, clinical trial data, and other non-structured records. Searching these types of flat files and weakly typed sources is not possible with SQL queries and full-text retrieval methods may not be useful because the researchers do not have a query (or hypotheses) to answer. Here DM techniques are applied to discover patterns and suggest to the researchers a basis for further investigation.

The mining of data

Brachman and Anand (1996) note that there is no systematized DM methodology, although major steps can be identified:

- “Getting to know the data and the task: this stage is more significant than it sounds, especially when the data is to be pulled from multiple sources and when the analysis will not be done by the business user.
- Acquisition: Bringing the data into the appropriate environment for analysis.
- Integration and checking: confirming the expected form and broad contents of the data and integrating the data into tools as required.
- Data cleaning: Looking for obvious flaws in the data and removing them, and removing records with errors or insignificant outliers.
- Model and hypothesis development: Simple exploration of the data through passive techniques and elaboration by deriving new data attributes where necessary; selection of an appropriate model in which to do analysis; and development of initial hypotheses to test.
- Data mining: Application of the core discovery procedures to reveal patterns and new knowledge or to verify hypotheses developed prior to this step.
- Testing and verification: Assessing the discovered knowledge, including testing predictive models on test sets and analyzing segmentation.
- Interpretation and use: Integration with existing domain knowledge, which may confirm, deny, or challenge the newly discovered patterns.”

Typically a subject specialist, working with a data analyst, refines the problem to be resolved. In what is termed *verification-driven*, or *top-down*, *data mining* (Berry and Linoff 2000), this may be pursued by posing a standard query: e.g., what are the sales in Chicago for 2001. The result of these SQL queries generates a kind of cross-tabs report based on the predetermined structure of the database. The next step is to run appropriate machine learning algorithms (Mitchell 1999; Langley and Simon 1995), or combinations of algorithms. This step may entail repeatedly altering the selection and representation of data. For instance, the miner may segment the data based on a hypothesis that a set of properties (e.g., median age, income, and zip code) form an appropriate group for a direct mail campaign and alter the selection of properties if nothing interpretable is generated.

Alternatively, the miner may not have an hypothesis (Nakhaeizadeh, Reinartz, and Wirth 1997) and so asks the system to create one (called *predictive*, *discovery-driven*, or *bottom-up data mining* (Weiss and Indurkhya 1998; Berry and Linoff 2000)), such as "do sales for beta blockers in the Chicago-area outpace those in the Los Angeles area." The DM system either proves or disproves it through statistical regression. But to achieve this end, the data must have been previously selected and cleaned, deciding on the granularity of each data type (Cabena et al. 1998). For instance, does the "Chicago area" include the geographic limits of that city or all markets served from Chicago area zip codes? Will a distributor's sales be represented by a category (e.g., $1 - $2 million sales/annum) or a value (e.g., $1,400,000).

In both situations a DM application may first classify or cluster (Jain, Murty and Flynn 1999) the data, through some artificial intelligence algorithms (of which artificial neural networks are most common), into a self-organizing map from which cause-effect association rules can be established. For instance, by clustering credit card purchasing histories of high-fashion clothing from a six-month period, it is possible to determine which customers are likely to purchase related adult-luxury products. By altering the underlying statistical model, it is also possible to have neural networks build non-linear predictive models. An illustration is determining which graduate school marketing campaign is likely to draw which types of applicants, regardless of the candidates' past academic performance.

The generated association rules also include probabilities. In Date's example (2000, 722) of a customer buying shoes, the association rule suggests that socks will be purchased, too: for all transactions tx (Shoes $\in tx \rightarrow$ socks $\in tx$) where "shoes $\in tx$" is the rule antecedent and "socks $\in tx$" is the rule consequent, and tx ranges over all sales transactions, the probability of both purchases occurring in the same sale is x%.

Association rules provide the user with two additional statistics: support and confidence. Support is the fraction of the population that satisfies the rule; confidence is that set of the population in which the antecedent is satisfied in which the consequent is also satisfied. In Date's socks and shoes example, the population is 4, the support is 50% and the confidence is 66.6%. The end-user is fairly confident in interpreting the association as "If a customer buys shoes, he is likely to buy socks as well, though not necessarily in the same transaction." The decision-making knowledge (or heuristic) of the domain-specialist helps in avoiding derived correlations that, for a specific data mining activity, may be useless. These include the "known but trivial" (people who buy shoes will buy socks), "chance" (the shoes and a shirt were on the same sale), "unknown but trivial" (brown shoes were purchased with black ones).

Association rules may be time-dependent or sequential. To illustrate, the purchases of a customer base may be grouped into sales periods (e.g., the "Spring Sale", "Summer White Sale", "Pre-School Fall Sale") and sequential algorithms may determine that if children's beach wear is purchased in the Spring Sale, there's an 80% chance that school clothing will be purchased during the Fall sale.

Besides association and sequencing, other main processes include classification and clustering, which are performed by specific computing algorithms. These techniques group based on how they treat the data: by correlating or finding relationships between the records (e.g., neural networks, link-analysis), those that partition the data (e.g., decision trees and rules, example-based nearest-neighbor classification, case-based reasoning, decision trellises (Frasconi, Gori and Soda 1999)), those that record deviations (deviation detection (Arning 1996), non-linear regression) and others (inductive logic, hybrid multi-strategy techniques, such as combining rule-induction and case-based reasoning (Coenen, Swinnen, Vanhoof and Wets 2000)).

Finally, to profit from data mining activities, the human analyst, a domain expert, must be able to interpret the results of the model in a manner appropriate for that field: "Each industry has evolved salient and customary ways of presenting analyses. The data mining output has to fit into this framework to be readily absorbed and accepted by the people who will use the results" (IBM 1999). The results of the calculations are visualized on screen, displaying complex multi-dimensional data, often in 3-dimensional renderings. Such visualization software is intended to give the user a mental framework for interpreting the data. (See Keim 1999 for a comprehensive review of visualization techniques).

The basic DM processes described above incorporate several assumptions about the size and quality of the data, the knowledge of the end-user, and the computing environment. These assumptions cannot be taken for granted as DM evolves.

Data

All data mining activities are founded on the properties and representations of the data. As DM tools have no built-in semantic model of the data (Moxon 1996; Spaccapietra and Maryanski 1997), users must take necessary precautions to insure that the data are "cleansed" or in a state that minimizes errors based on the data. Addressing the issue of missing values (Ragel and Crémilleux 1999), inconsistent, noisy, and redundant data is part of the data cleaning process. In situations where the nuisance data cannot be eliminated or probabilistically determined, DM requires more sophisticated statistical strategies to compensate by identifying variables and dependencies. However, data that are mined using compensating mathematical methods risk over-fitting the data to the model , that is by accidentally selecting the best parameters for one particular model. Thus, preparing data for mining entails certain amount of risk and so must be carefully performed.

Miners must determine record usability (Wright 1996) and preprocess data to a fixed form, such as binary (Tsukimoto 1999) or ordered variables. However, there are times when data may not be mapped to a standard form, such as when processing free text where replicated fields may be missed. Similarly many DM methods are designed around ordered numerical values and cannot easily process categorical data. Users who attempt to standardize their data through any number of methods (normalization, decimal scaling, standard deviation normalization, and data smoothing) may be able to improve feature selection but accidentally introduce new errors (Liu and Motoda 1998a, b). For instance, when measures are small, neural networks often train well, but if not normalized, distance measures for nearest-neighbor calculations overweight those features. Moreover, the miner must ensure that the normalization applied to the training set also be applied to mined data. Some methods, such as neural networks and regression trees, have smoothers implicit in their representation and perform well for prediction. Smoothing also may reduce the search space by discretizing continuous features into a set of discrete ones, covering a fixed range of values.

This section discusses some responses to data mining of continuous, missing and reduced data sets.

Continuous variables are often discretized (Dougherty, Kohavi and Sahami 1995; Zhong and Ohsuga 1994; Fayyad and Irani 1993), although it may entail a loss of information value. In the pharmaceutical company example, the marketing group may convert the volume of sales into discrete groups of "high" and "low" volume. This may help the sales force conceptualize the question, although it may degrade the DM process. In neural networks, for instance, input parameters that are not scaled or coded appropriately impact its learning ability. The simplest method is to divide the range into equal width units. Miners must be aware of the risk of losing information about the relationship between pre-assigned classes and interval boundaries when discretizing (Ching, Wong and Chan 1995).

One solution to missing data is to predict the values from other data. Such surrogate techniques are possible, such as when using decision trees, but the answer is not simple. Some missing values may be null, but they may also be inapplicable for the task. This situation arises where heterogeneous databases are mined because the relational model requires all types in a relation to have the same number of attributes (Deogun, Raghavan and Sever 1995). For example, in selecting a patient group for possible inclusion in a clinical trial, some missing data attributes may be estimated based on other examples for which the value is known. In growing a decision tree, the miner assigns a common value to a missing attribute, calculated from the entire set or projected from other members within a cluster. The missing data may also be assigned a probability of all possible values and then re-estimated based on observed frequencies of other values within the training set (Quinlan 1986; 1993).

Similar to the case of continuous values, missing data in neural networks is difficult to detect and prevents the network from converging. This, as Ho (2000, 48) notes, is a situation where both domain expert and analyst should work together and where most DM applications fail to provide more interactive opportunities for users.

Another avenue to resolving missing data, or addressing uncertainty, by prediction comes from fuzzy sets and rough sets (Raghavan, Sever and Deogun 1994; Deogun, Raghavan, Sarkar and Sever 1996; Lingras and Yao 1998; Lin and Cercone 1997). Rough sets expose hidden deterministic rules in databases and can be expanded to extract probabilistic rules (Luba and Lasocki 1994). The generalized rough set model can be applied where data are missing, or when users provide a range for the data, which addresses a great challenge for DM. Zhong, Skowron and Ohsuga (1999) outline the interaction between rough sets, data mining and granular-soft computing. Finally, Hirota and Pedrycz (1999) outline the potential of fuzzy computing for data mining. Chiang, Chow and Wang (2000) examine fuzzy sets for time-dependent linguistic systems, something that might first suggest using hidden Markov models.

Another data-centered technique to improve computer efficiency minimizes the size of the data set before processing. Data reduction is performed to reduce the size of the search space or to remove fields that do not contribute to increasing the efficiency of the mining algorithm (Agrawal, Mannila, Srikant, Toivonen and Inkeri Verkamo 1996) or which contribute in insignificant ways. Reducing the data requires careful, validated selection of properties which are redundant (and so do not contribute to increasing effectiveness or information gain (Furtado and Madeira 1999)) and which do not accidentally make the new data to be mined unrecognizable when compared to the training set.

One method is *feature selection*, a pre-pruning, inductive learning process. Feature selection both improves computation speed and the quality of classification (Deogun, Raghavan and Sever 1995; Kira and Rendell 1992).

Users may want to select the "best" features of their data when there are a large number of features, or when calculating standard errors is computationally intensive. Simplification improves computer time but users may tend to select the features to best suit their model (Elder 2000) instead of working more with the data. For example, in decision tree learning, methods are developed that stop growing the tree earlier, before it reaches the point where it classifies perfectly the data set, and approaches that allow the tree to over-fit the data and then prune the resulting rule set (Ho 2000). The later case may be preferable because the rule set is more interpretable to the end-user.

Additionally, smaller sets increase the system's ability to test hypotheses. If the smaller set yields good results, it may be possible to bypass other tests on the entire dataset. Inexperienced miners may mistake good-looking sets for actually valid results and skip confirmatory tests (Elder 2000). A reduction method based on smaller sets, on the other hand, can and should be subjected to confirmatory algorithms because the set can be efficiently manipulated. This also suggests that small sets may be appropriate for distributed systems, which later can take the aggregate for a final output (Provost and Kolluri 1999).

Data reduction techniques vary depending on the learning algorithm. For example, when reducing data for neural networks, the inputs must be fitted to a range, usually 0-1. The transformation choice will impact the training of the system. Inappropriate reduction introduces outliers, which in turn skew distributions and consequently the network will perform poorly. Caruana and Freitag (1994) demonstrate a system that outperforms on the subset compared to the full set. This suggests that subsets can generate information about optimal values for testing against the entire dataset.

Algorithms

Algorithm design stressing computational efficiency (Joshi, Han, Karypis and Kumar 2000; Joshi 2000) has become a critical issue for DM for several reasons. One is that most "first-generation algorithms" (Mitchell 1999, 30) assume certain properties of the data, such as fitting into a single computer's memory (Grossman, Kasif, Mon, Ramu, and Malhi 1999) or dealing only with numeric and symbolic data (Mitchell 1999, 3). Another reason is the difficulty of learning rules from extremely large databases (Mitchell 1999; Agrawal, Imielinski and Swami 1993; Gray, Bosworth, Layman and Pirahesh 1995). DM algorithms also assume that the data have been carefully prepared before being subjected to largely automated rule production systems, minimizing the human end-user's interactive role (Fayyad 1998). To illustrate, algorithms designed for small search spaces may generate spurious associations when applied to large, distributed or parallel sources (Imasaki 2000), which might be handled more effectively if the user's knowledge were incorporated at key stages (Talavera and Bejar 1999; Mitchell 1999). The task in algorithm design, then, is how to accommodate diverse sources of data, increases in the number of records and attributes per observation, derived rule sets used to analyze the collection, and how to increase the user's participation. Some of the developments are outlined below.

Agent-based approaches (Mattox 1998) are software applications programmed to investigate and collect data on their own. These intelligent agents prowl the Internet relying on user profiles (Joshi 1999; Joshi, Joshi, Yesha and Krishnapuram 1999) and user-supplied information about the subject (e.g., medical data, Kargupta, Hamzaoglu and Stafford 1997b) and document types of interest. PADMA (Kargupta, Hamzaoglu, Stafford, Hanagandi and Buescher 1996), Harvest, ParaSite (Spertus 1997), OCCAM and FAQ-Finder systems are examples. More interactive agents such as the Internet shopping tool ShopBot interact with and learn from unfamiliar information sources.

Association or *rule induction* procedures come originally from the retail industry, such as analyzing customers' account portfolios, to express item affinities in terms of confidence-rated rules, and have been adapted to many situations. Indeed, a most active area in DM research is improving the efficiency and removing redundancy of association and classification rules. Association rule production is not efficient with continuous classes, or when there are many intervals in the data (Fayyad and Irani 1993). In response, fuzzy techniques (Kuok, Fu, and Wong 1998) improve predictions, but degrade the end-user's ability to comprehend the generated rules.

Ankerst, Ester, and Kriegel (2000) examine how to improve the user's participation in semi-automatic classification. Some efficiency-oriented research examines the influence on processing speed versus set size (Shen, Shen and Chen 1999) and set type (Pasquier, Bastide, Taouil and Lakhal 1999a). Other work considers the impact of the data type on rule production: numeric (Fukuda, Morimoto, Shinichi, and Takeshi 1999) or quantitative (Hong, Kuo and Chi 1999). Liu, Hsu and Ma (1998) generalize association rules to classify high dimensional data.

Clustering, often the first step in DM, assigns database records with many shared attributes into smaller segments, or clusters. DM systems automatically identify distinguishing characteristics and assign records to an *n*-dimensional space. It is common in demographic-based market analysis. In image databases, "clustering can be used to detect interesting spatial patterns and support content based retrievals of images and videos using low-level features such as texture, color histogram, shape descriptions, etc." (Aggarwal and Yu 1999, 14). Good clustering techniques maximize the cluster membership while minimizing accidental membership, by applying either supervised or unsupervised artificial intelligence techniques.

The algorithms used in clustering must examine all data points, determine potential clustering features, and refine cluster membership, or "classifier generation and classifier application" (Reinartz 1999, 32). As the size of the database grows, the likelihood of outliers also grows, requiring some means of removing irrelevant dimensions, such as feature selection or pruning (Kohavi and Sommerfield 1995). A popular technique is *k-means* (Zaki and Ho 2000,12), which randomly picks *k* data points as cluster centers and assigns new points to clusters in terms of squared error or Euclidean distance. The challenge, as Farnstrom, Lewis and Elkan (2000) note, is scaling *k-means* clustering. Through multiple additive regression, scaled *k-means* clustering offers secondary validation and may be applied to parallel and distributed DM environments. For large datasets, Joshi, Han, Karypis and Kumar (2000) describe a method of creating "candidate *k*-itemsets", a minimized frequent itemset, such as those used in market based analysis. In a similar vein, Jagadish, Madar and Ng (1999) suggests using "fascicles" to create association rules with small sets of entities that share a great number of properties rather than seeking larger sets of items with less commonality.

Classification of data is arguably the most important function of data mining (Reinartz 1999) and the most commonly applied technique (Moxon 1996). Classification employs a set of pre-determined examples to develop a model to categorize population of records to predefine the similarity of neighbors before machine learning techniques are employed (Datta 1997; Wilson and Martinez 1997). A typical use is fraud detection (Bonchi, Giannoti, Mainetto and Pedreschi 1999) and credit risk applications. Classification employs some form of supervised learning method such as decision trees, neural networks, DNF rules, Bayesian classifiers (Langley, Iba, and Thompson 1990), and genetic algorithms (Fu 1999) to predict the membership of new records.

Another typical technique is nearest neighbor classifiers, which use a training set to measure the similarity (or distance function) of all tuples and then attempts an analysis on the test data. Variations include *k* nearest neighbors (which classifies each record based on a combination of classes of *k* records that are most similar to it in the data set), weighted voting of nearest neighbors (Cost and Salzberg 1993), and edited nearest neighbor (Dasarathy 1991). Mining of heterogeneous sources requires updated distance measurement functions (Wilson and Martinez 1997).

Decision trees are a popular top-down approach to classification that divides the data set into leaf and node divisions (a "recursive partitioning approach" (Zaki and Ho 2000)) until the entire set has been analyzed (Reinartz 1999). Growing the tree usually employs CART (classification and regression) and CHAID (chi squared automatic interaction detection). Each interval node in the tree represents a decision on an attribute, which splits the database into two or more children. Decision trees are popular because they process both qualitative and quantitative data in an efficient and accurate manner. For qualitative attributes, the "set of outcomes is the set of different values in the respective attribute domain"; quantitative attributes rely upon a specific threshold value be assigned by the user to generate different branches. This greedy search over the entire search space for all possible trees is computationally very intense and in light of the huge size of databases becoming impossible to perform. There are other related techniques which seek the "best" test attribute, such as nearest neighbor classifiers that handle only a few hundred tuples, entropy, and information gain, which are mentioned in passing for completeness' sake but cannot be addressed here.

Note that other techniques are useful. Each of the following have mature literature, too vast to include in this review, although these techniques are important in DM. Very popular in business and classification (Smith and Gupta 2000), a*rtificial neural networks* are non-linear predictive models that learn from a prepared data set and then are applied to new, larger sets. Zhang and Zhang (1999) describe a novel approach based on a geometrical interpretation of McCulloch-Pitts neural model. *Genetic algorithms* (GAs), like neural networks, are based on biological functions. GAs work by incorporating mutation and natural selection and have been applied in scalable data mining (Kargupta, Riva

Sanseverino, Johnson and Agrawal 1998). An offspring of genetic-based mining, genetic programming, is also employed (Wong 2000). *Sequence-based* analysis is time-dependent such as when the purchase of one item might predict subsequent purchases. *Graphic models* and *hierarchical probabilistic representations* are directed graph, generalized Markov models and hidden Markov models. These techniques are usually employed in conjunction, with others, among them case-based reasoning, fuzzy logic, fractal-based transforms, lattice, rough sets (Lin and Cercone 1997).

Software applications implement the algorithms.

The computing platform that stores, manipulates, examines, and presents the data must be sufficiently powerful or be provided with efficiently designed software. This is an important issue in DM because it often involves considerable computing overhead to perform iterative data analyses (Savasere, Omiecinski and Navathe 1995) and complex, interactive visualization.

The software used in data mining may be categorized based on the application's operation (Simoudis 1995): generic, single-task; generic, multi-task; and application-specific.

Generic, single-task applications emphasize classification (decision trees, neural networks, example-based, rule-discovery). These applications require significant pre- and post-processing by the user, typically a developer who integrates these approaches as part of a complete application.

Generic, multi-task systems support a variety of discovery tasks, typically combining several classification techniques, query/retrieval, clustering, and visualization. Multi-task DM systems are designed primarily for users who understand data manipulation. See www.kdnuggets.com for a complete list of software applications.

Application-specific tools, on the other hand, are employed by domain-specialists, people trained in a field, say bioinformatics (Bourne 2000), but who know little about the process of analysis. Such miners rely therefore more heavily upon the software to validate patterns detected in the data and to guide in the interpretation of results.

Users

The users of data mining traditionally were people who worked within a subject domain, such as business, assisted by trained statistical analysts. Despite complex visualization tools to represent the results of mining, interpretation of association rules may still overwhelm the end-user. Independent application of DM techniques introduces new user-centered concerns. For instance, some algorithms confuse the enduser because they do not map easily to human terms (such as "if-then" rules) or may not use the original data's named attributes (Moxon 1996).

DM supports the end-user by automating as much as possible hypothesis discovery and testing. Recently, however, researchers and applications developers feel that the purpose of data mining is better served by integrating more of the user's knowledge and heuristics through the interface (Moxon 1996). Many methods are not interactive and so cannot incorporate the user's prior knowledge, except in simple ways, where the domain knowledge of the user could improve the choice of algorithm and interpretation of results. This suggests work in graphic representations and natural language generation to improve the understandability of data mining results.

Grossman, Kasif, Moore, Rocket and Ullman (1999) note that the explosion of digital data has outpaced the ability of domain-specific users to process it. They suggest that the number of doctorates awarded in statistics has remained constant while the need for statistical analysts has grown, forcing subject-specialists to depend more upon the software's guidance. The increased use of data mining technology by non-data analysts and the need for more human-oriented interactivity (queries and display) should spawn research in improving the user interface, casual browsing and developing techniques to manage the meta-data required for data mining.

Elder (2000) outlines ten concerns of the inexperienced applications-oriented data miner: lacking data, lack of training, relying on a single technique, asking the "wrong" question of the data, listening only to the data, accepting over-fitted data, discounting the difficult cases, premature extrapolation, improper sampling, and looking too much for an interpretable answer. For example, inexperienced data miners may believe the first presentation of results and not see that variables in the data may accidentally "lift" the conclusions, that is, exert a causality that distorts the true behavior of the data, while an experienced miner, or more sophisticated applications, may bundle several techniques for greater validation and present a multifaceted analysis.

Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, and Simoudis (1996, 44) also sound the insufficient training alarm: "graduates of business schools are familiar with verification-driven analysis techniques, occasionally with predictive modeling, but rarely with other discovery techniques." Because of this users may opt for tools which support models with which the user is comfortable. New data miners may also ignore the problems associated with missing data. Although in some domains, such as finance (Kovalerchuk 2000), data warehousing minimizes the impact of dirty data, this is particularly a concern for users who emphasize interactive queries. New users also are subject to formulating poor or inappropriate hypotheses and so are faced with an overabundance of patterns (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro and Simoudis 1996, 44).

Domain-Specific Applications

Domain-specific data mining now plays a broader, more influential role because of the dearth of analysts and the expanded interest in applying DM techniques to serve domain-specific knowledge needs. Fountain, Dietterich and Sudyka (2000), for example, turn integrated circuit tests into a method for optimizing VLSI design. Gavrilov, Anguelov, Indyk and Motwani (2000) use stock market data to determine which evaluative measures are best for that field.

Astronomy, for example, employs time-dependent and image data (Ng and Huang 1999). Formerly relying on visual inspection of photographs to find new astronomical phenomena, DM applications for this field are tailored to classify properties unique to astronomy (e.g., brightness, area and morphology). Work is underway to define the best models for astronomy (Schade, Dowler, Zingle, Durand, Gaudet, Hill, Jaeger, and Bohlender 2000) and to apply some of these models to digital surveys of the skies (Odewahn 1999). Brown and Mielke (2000) demonstrate the relationship of statistical mining with visualization for atmospheric sciences.

The biological sciences, medicine (Luvrac, Keravnou and Blaz 1996) and chemistry (Hemmer and Gasteiger 2000) are particularly interested in adopting DM techniques. The trend within medical data mining is to focus on specific diseases or on processing the particular data objects generated in medical practice, such as term domain distribution for medical text mining (Goldman, Chu, Parker and Goldman 1999). Hsu, Lee, Liu and Ling's work (2000) on diabetic patients, and Pendharkar, Rodger, Yaverbaum, Herman and Benner (1999) for breast cancer, Holmes, Durbin and Winston (2000) for epidemiology are representative.

The nature, complexity and volume of the data – such as genome expressions and sequence data – make biology a natural domain for exploitation by data mining techniques. Brazma (1999, ¶1) describes a yeast problem which suggests to the reader just how much computerized efforts have influenced the thinking of scientists: "First genomic scale data about gene expression have recently started to become available in addition to complete genome sequence data and annotations. For instance DeRisi et al. have measured relative changes in the expression levels of almost all yeast genes during the diauxic shift at seven time points at 2 hour intervals. The amounts of such data will be increasing rapidly, thus providing researchers with new challenges of finding ways to transform this data into knowledge, on one hand, while opening new possibilities of pure *in silico* studies of various aspects of genome functioning, on the other hand."

Genomic sequencing and mapping research have generated many web-based databases. Along with other on-line sources, there is untapped potential for mining these systems for gene identification, cellular function and relationships to diseases. Indeed, through scaled algorithms, it is possible to compare entire genomes. Other biological DM is highly specific: King, Karwath, Clare and Dehaspe (2000) demonstrate the predictive uses of DM in biotechnology; Zweiger (1999) explains using biotechnical information to generate new metadata. Some work is underway to integrate DM full-text biomedical sources and linking the results to web-based database sites, such as SwissProt and GratefulMed, with interactive visualization (Stapley and Benoît 2000; Benoît and Andrews 2000). Advances in medical research on the Internet (genomic and other disease, cellular function, drug data) and locally-housed full-text holdings notwithstanding, finding the relationships between these data sources remains largely unexplored.

Data Mining Architecture

Trends in incorporating increasingly large databases and the integration of DM into non-business endeavors suggest that data mining is moving away from back-end technical offices with trained analysts to the front office or lab computer, impacting computer system architecture (Nestorov and Tsur 1999; Skillicorn 1999; de Sousa, Mattoso and Ebecken 1999). More powerful networked desktop and micro computers suggest opportunities to resolve DM problems with distributed, parallel and client/server architectures. For example, as Moxon (1996, ¶9) noted, although multiprocessing systems able to compute over 10,000 transactions per second are routine, "low-end four- and eight-way Pentium-based SMPs [symmetric multi-processing] and the commoditization of clustering technology promise to make this high transaction-rate technology more affordable and easier to integrate into business…". Newer network architectures, such as SMP workstations, MMP (massively parallel processing) (Kargupta and Chan 2000), high performance workstation clusters, and distributed DM are favorable paths. The hardware-oriented responses may be based on high performance computers, such as the ACSys Project (Williams et al. 1999), or networks of high performance workstations.

In addition, the Internet as a form of distributed computing encourages data mining mixed media and heterogeneous databases, and introduces concerns associated with distributed processing. As Grossman, Kasif, Moore, Rocke and Ullman (1999, 5) state, the next generation Internet will increase throughput to "OC-3 (155 Mbytes/second) and higher, over 100 times faster than the connectivity provided by current networks." This will impact scientific research, such as the Human Genome Project and space exploration data (Zaki and Ho 2000), which in days generate petabytes (Fayyad, Haussler, and Stolorz 1996) of high dimension data and which increasingly make databases available via the Internet. This section examines architecture-centered responses to very large data sets, through distributed, parallel, and client/server methods.

Distributed data mining partitions the data store and computing load across a network is one avenue to handling very large datasets (Chattratichat, Darlington, Guo, Hedvall, Köler and Syed 1999).

The JAM (Stolfo, Prodromidis and Chan 1997) and BODHI (Kargupta, Hamzaoglu and Stafford 1997b) models are examples that use local learning techniques to build the model at each site and then integrate the models at a centralized location. Distributing data across a network for DM purposes requires tight integration of the communication protocols and the workstations (e.g., Id-Vis (Subramonian and Parthasarathy 1998) and the Papyrus system (Grossman, Baily, Kasif, Mon, Ramu and Malhi 1998)). Distributed DM is not limited to high performance computers. Shintani and Kitsuregawa (2000) describe how to generalize association rule mining on large-scale clusters of personal computers. This approach to load balancing combines the power of interconnected PCs, in a computer network that a large, data-rich organization might have.

Integrating distributed data for mining (Sarawagi, Thomas and Agrawal 1998; Lavington, Dewhurt, Wilkins and Freitas 1999) resolves memory and storage issues but introduces new ones. The heterogeneity of the data may increase (El-Khatib, Williams, MacKinnon and Marwick 2000), requiring more attention on the data cleaning stage and addressing local data variance. On the other hand, awareness of the data structure of distributed databases, or the metadata of tables in distributed systems, can be mined to generate a new information source, from which patterns across the structure of databases might be established (Tsechansky, Pliskin, Rabinowitz and Porath 1999).

Alternatively, mining may be performed on parallel architectures. Mining in parallel inherits many local-database issues, such as preparing a good data mart and indexes, and also requires careful choice of model. For example, allocating data in parallel systems risks skewing the results and may occasion shifting data across the network which is not always feasible because of limited network bandwidth, security concerns, and scalability problems (Kargupta and Chan 2000). The basic approach to parallelization is partitioning of the data, processing, and queries. One method assigns parts of the programming to different processors. This kind of "inter-model parallelism" increases execution without reducing throughput (Small and Edelstein 2000) such as might be found in a neural net application on which different nodes or hidden layers run simultaneously on each processor. Alternatively "intra-model parallelism" distributes the load among processors and then recombines the results for an answer. In all parallel data mining (PDM), some means of inter-node communication is needed to coordinate the independent activities.

Parallelization of data mining raises also some data modeling questions (Agrawal and Shafer 1996; Parthasarathy, Zaki and Li 1998). For example, even with good data distribution, parallel data mining algorithms need to reduce I/O, to minimize competition for shared system busses (Brown 2000). Brin, Motwani and Silverstein (1997, 265), for instance, propose a method for "large itemset generation to reduce the number of passes over the transaction database by counting some (k + 1)-itemsets in parallel with counting k-itemsets."

Queries, too, may be parsed and relayed to individual CPUs (called "inter-query parallelism") or parts of the query distributed ("intra-query parallelism"). The actual data mining may be performed through "partitioned parallelism", with individual threads processing subsets of the data. Data may be partitioned by "approximate concepts" (partitioned by rows), or "partial concepts" (partition by columns) (Skillicorn 1999; see also Zhong and Ohsuga (1994)).

Algorithms require adjustment to work in parallel: those designed to work on one system may fail on parallel systems unless refitted to be aware of dependencies or time sequential data. For instance, Quinlan's decision tree algorithm C4.5 (Quinlan 1993; 1986) has been adapted for parallel computing, PC4.5 (Li 1997). Glymour, Madigan, Pregibon and Smyth (1996) explore the similarities and differences between data mining and statistics and suggest opportunities, such as applying linear algebra, to solve problems of parallel algorithms.

The standard sequential algorithm for parallel DM is *Apriori* (Agrawal, Imielinski and Swami 1993; Agrawal and Srikant 1994). Apriori assumes that patterns in the data are short. This may not be the case in large databases: in an example where the longest itemset is 40, 2^{40} subsets would be generated (Aggarawal and Yu 1999, 16), each of which needs to be validated against the database. The problem of size requires some solution, such as "look-ahead" techniques to locate long patterns before processing (Bayardo 1998). Other algorithms, such as *count distribution* , minimize communications overhead and have been tested to 32-node configurations. The *Eclat* system has been shown to obtain speeds as much as 11 times faster than count distribution . PDM is not perfected: even the best algorithms will suffer from load-balancing problems when run in MPP-type environments.

Internet and Data Mining

As the Internet matures, it will play an increasingly important and diverse role for data mining. The Internet has already influenced DM itself in the sense that all web sites and accessible back-end databases offer a tremendous collection to be mined (Florescu, Levy and Mendelzon 1998). Today it is used primarily to deliver text, weakly-typed documents, and mixed media but offers great potential for association analysis by mining the web site content, document structure, site relations, and user behavior. Because of the text orientation of most web documents, web mining is closely linked to text mining and information retrieval. However, the Net also delivers images, sound, as well as accessing structured databases in a client/server (Fong 1997) environment, which means mining the Internet is made especially difficult because of the heterogeneity of formats, questions of document structure, and the lack of quality control. For instance, a single research web site may host published technical reports, lab notes, structured databases of unknown quality, chat and e-mail archives, as well as non-textual source data. Using Internet-specific techniques, such as creating documents with HTML, CSS, and XML tags, provides some semantic framework that can be analyzed. Wong and Fu (2000) suggest parsing web documents to form associations among text data. Joshi, Joshi, Yesha, and Krishnapuram (1999) see stores of web documents as a mine for analyzing the behavior of the user for system optimization or to profile the user for mass-personalization (Joshi 1999). This "webhousing" (Mattison 1999) or web-based analysis (Greening 2000; Kimball and Merz 2000; Kimball 1999; Paliouras, Papaheodorou, Karkaletsis, Spyropolous and Tzitziras 1999; Pinter and Tsur 1999; Pravica 2000a, 2000b; Smith 1999; Winter 1999) is expected to influence how web sites and business decisions are planned. In fact, webhousing already impacts issues of mass commercialization, such as how commercial graphics are selected for real-time web-based advertising, or as part of e-commerce (Meña 1999), and market modeling (Loshin 2000; Auditore 1999; Chou and Chou 1999). Mining the Net for clickstreams and combining the use behavior with commercial personal database information is controversial. The literature suggests this alters the relationship between marketing and customers (Biafore 1999; Gardner 1996) and raises privacy issues (Meña 1999; Agrawal and Srikant 2000; Berry 1999b).

As the Internet came to be incorporated as a mine, it is not surprising that the early views were database-biased. SQL was extended to create Web-oriented query languages. One, WebSQL (Mendelzon, Milhaila and Milo 1997), combines structured queries based on the hyperlinks of the documents, and content analysis based on information retrieval techniques (Frakes and Baeza-Yates 1992). Other database-oriented methods have appeared: WebSite (Beeri, Elber, Milo et al. 1998), WebOQL (Arocena et al. 1997) and WebLog (Lakshmanan, Sadri and Subramanian 1996). Similarly, programs like TSIMMIS (Chawathe et al. 1994) correlate data extracted from heterogeneous and semi-structured sources to create a database representation of the information. As will be discussed below, text mining and information extraction turn to mining data from the semi-structured sources on the Internet.

Some applications, such as the ARANEUS system, focus on Internet-unique phenomena, such as hyperlinks (Merialdo, Atzeni and Mecca 1997), others call for a Web-wide schema for metadata mining (Khosla, Kuhn and Soparkar 1996) and, responding to the dynamic nature of web sites, incremental integration of schema for individual sites (Chen and Rundensteiner 1999), or mining SGML's derivatives (XML, HTML, ODA (Thurisingham 1999)).

Application of data mining to web-based data has also impacted DM theory (Chaudhuri 1998; Beeri and Buneman 1999; Dasarathy 2000; Cooley, Srivsatava, and Mobasher 1997; Chen, Han and Yu 1996). Traditionally, DM required a closed, very large, historical-oriented database, optionally supported by data warehouses and data marts. Web-based data mining introduces the notion of the "clickstream" (Kimball and Merz 2000). A clickstream is the trail of mouse and hyperlink actions taken by an end-user. These actions are recorded into a transaction log that is parsed almost continuously with analysis sent to the administrator in near real time. The immediacy of data gathering and the volume of Internet-based data traffic raise questions of data granularity (Pedrycz and Smith 1999) and algorithms for time series . Client/server transactions offers interesting research possibilities into artificial intelligence and belief systems (Xiang and Chu 1999), the nature of implicit facts in rule construction (Chou and Chou 1999), and the categorization of data (Bleyberg 1999).

Data mining as a management information system or web-master practice has evolved also to integrate web-based methods, such as discovering document structures from web pages (Ahonen, Mannila, and Nikunen 1994), Java code (Witten and Frank 2000; Bose and Sugumaran 1999) and *n*-tier client/server architecture (Chattratichat, Darlington, Guo, Hedvall, Köler and Syed 1999).

The architecture of digital libraries (DLs) are often thought of as part of the Internet. DLs are digitized information distributed across several sites and perhaps consists of text, images, voice and video. Grossman (1996) notes that DLs while text-oriented, they also consist of tabular data and suggests that mathematical methods can be applied to DLs. For example, tabular data describing "new homes in a region to the number of violent crimes per 100,000 per count" (Grossman 1996, 2) might be mined fruitfully for prediction, classification, clustering, and anomalies.

Moreover, DLs often have keywords or other attributes. This suggests concept clustering (Grossman 1996) by term or latent semantic indexing (Jiang, Berry, Donato, Ostouchov and Grady 1999), or association queries for attribute-based associations (Abiteboul 1997).

Data Mining and Text Mining

DM combined with the Internet's current emphasis on textual data questions the relationship between data mining and text mining (Ahonen, Heinonen, Klemettinen and Inkeri Verkamo 1997). Text mining (TM) is a fairly independent research domain with its own literature (Trybula 1999). It is related to digital libraries, information retrieval, and computational linguistics (Lee and Yang 2000) in the sense that it aims to discover knowledge from semi-structured or unstructured text in text collections. Hearst (1999, ¶5), however, interprets data mining, text data mining and information retrieval as different phenomena, because "the goal of data mining is to discover or derive new information from data, finding patterns across data sets, and/or separating signal from noise" while information retrieval is "query-centric" and text data mining is closer to corpus-based computational linguistics and EDA. A large store of web-based semi- and unstructured documents may be thought of as much as a data warehouse as the highly structured database is to typical DM. Text documents can be used for data discovery and analysis, and when prepared, can be used for predictive purposes using regression, forecasting techniques, CHAID, decision trees and neural networks, and so on, just as DM does (Mattison 1999) and so is included here.

Text mining, like IR, may use any number of text tokenization and manipulation algorithms, such as singular value decomposition (Tan, Blau, Harp and Goldman 2000) and latent semantic indexing . Like data mining, TM requires techniques to analyze data and extract relationships from large collections of “weakly typed” (Beeri et al. 1998), usually local-area network or web-based documents. For a chapter-length treatment of text mining, see (Trybula 1999).

Web-based data are difficult to process because the formats are often missing some identification of the sections. Similarly, there are often many sources of data on a topic, but locations differ making web-based text mining a distributed computing and redundant data issue. Reminiscent of data mining’s need for cleansed data, web documents may be only partial and there are no guarantees that the documents do not contain complementary, similar, or contradictory data (Beeri et al. 1998), which suggests data integration from web sources will be difficult (Atzeni, Mendelzon and Mecca 1999). Compounding the difficulty of integration is a text mine of mixed script, multilingual documents (Lee and Yang 2000).

Advances in text mining algorithms alleviate some of these concerns. Using Reuters newswire sources, researchers (Feldman, Dagan and Hirsch 1998; Fedlman Klosgen and Zilberstein 1997; Feldman, Aumann, Fresko, Liphstat, Rosenfeld and Schler 1999), for instance, analyze text category labels to find unexpected patterns among text articles. As the following will demonstrate, there are many similarities between evolving DM and web-based text- and data mining.

Three approaches to web-based TM are mining the metadata (the data about the documents), such as performing knowledge discovery operations on labels associated with documents (Feldman, Aumann, Fresko, Liphstat, Rosenfeld and Schler 2000)), itemsets (groups of named entities that commonly occurred together; Hafaz, Deogun and Raghavan 1999) and word-level mining. Mining the terms in text corpuses is aimed at automatic concept creation (Feldman et al. 1998), topic identification (Clifton and

Couley 2000) and in discovering phrases and word co-occurrence (Ahonen et al. 1997). Others, such as Kryskiewicz (2000) and Pasquier, Bastide, Taouil and Lakhal (1999b) describe a method of discovering frequent closed itemsets for association rules. Holt and Chung (2000) expand on this by minimizing the search space through inverted hashing and pruning. Makris, Tsakalidis and Vassiliadis (2000) apply these techniques specifically to net-based searching and filtering for e-commerce.

It is interesting to note that text mining brings researchers closer to computational linguistics as it pays more attention to natural language elements in texts (Knight 1999). This means TM applications (Church and Rau 1995) discover knowledge through automating content summarization (Kan and McKeown 1999), content searching, document categorization, lexical, grammatical, semantic, and linguistic analysis (Mattison 1999). So standard DM techniques, such as self-organizing maps (Kohonen 1998; Kaski, Honkela, Lagus and Kohonen 1996; Honkela, Kaski, Lagus and Kohonen 1996), can be adjusted to integrate linguistic information from the texts in the form of self-organizing *semantic* maps (Ritter and Kohonen 1989) as a preprocessing stage for documents. Once prepared the text documents can be subjected to other DM techniques such as clustering and automatic extraction of rules.

The semi-structured format of web-based text documents, while presenting interesting opportunities, such as semantic network analysis (Papadimitriou 1999), is a crucial question in expanding the use of text mining for knowledge discovery (Lenzerini 1999; Beeri and Buneman 1999). Without a "common data model" of semi-structured data and a common schema model, it will be difficult to develop web- and text-oriented DM models to develop translation and integration systems to support user tasks, such as query formulation and system query decomposition (Beeri and Milo 1999).

Data Mining and Information Extraction

Another form of mining that merges textual and structured databases is *information extraction* (IE). The function of IE, write Gaizauskas and Wilks (1998, 17), is "to extract information about a pre-specified set of entities, relations or events from natural language texts and to record this information in structured representations called templates." Unlike text mining and information retrieval, both of which may extract terms from free text and establish relationships between them (Baeza-Yates and Ribeiro-Nero

1999), primarily to answer a query, IE is a complementary technique that populates a structured information source (the template), which then is analyzed using conventional queries or DM methods or to generate rules (Soderland 1999). Text mining concerns applying data mining techniques to unstructured text; IE "is a form of shallow text understanding that locates specific pieces of data in natural language documents, transforming unstructured text into a structured database" (Nahm and Mooney 2000, 627). Unlike IR, IE must maintain linguistic information about the extracted text: "'Carnegie hired Mellon' is not the same as 'Mellon hired Carnegie' which differs again from 'Mellon was hired by Carnegie'" (Gaizauskas and Wilks 1998, 18).

IE was originally related to story comprehension and message understanding, based on the communications theory notion of scripts (Schank and Abelson 1977) in which the role played by participants provided a predictive structure. IE quickly became associated with newswire analysis and online news (Jacobs and Rau 1990) for business trend analysis.

Finally, IE's relationship with other knowledge extraction fields is not yet settled. Wilks, Guthrie and Slator (1996) see IE as the next step beyond document retrieval while Robertson and Gaizauskas (1997) foresee a union of the two. IE has found special acceptance applied to domain-specific documents, such as medicine (Lehnert, Soderland, Aronow, Feng, and Shmueli 1994) and law (Pietrosanti and Graziadio 1997).

Summary and Conclusions

This chapter presented a synoptic view of DM's responses to its recent evolution, as evidenced by the literature published between 1997-2000. This review follows the lead of several independent assessments in identifying four grand challenges: data-centered issues, data mining architecture, algorithm design, and the user. Within the framework of those four themes, the review presented a sample of specific research questions and activities, along with a brief description of the associated data mining process to guide readers in understanding the application of those activities. The conclusion one draws is that DM has reached a level of maturity, expanding its role in business (see Bergeron's chapter in this ARIST) and to other areas, such as science.

Nevertheless, DM is at a crossroad. DM's unfolding results in a field too broad to be easily analyzed: the level of sophistication of constitutive research is advanced by several disciplines. As a result, DM's purview is not clearly defined by researchers or users, signaling that DM is at a crucial stage. Increasingly, DM responds to pressures arising from its growth, by adopting cognate research, such as investigations into the efficiency of very large databases. In the same vein, DM practice moves to integrate mixed media formats, influencing, and be influenced by, explorations in text mining, information extraction, and multimedia databases.

The review concludes that critical challenges remain in many areas of DM, including fundamentals of DM theory and the physical components of DM practice, the particulars of networked mining environments, and data reduction techniques. Additionally, DM practice is stressed by greater participation of independent miners who work without the aid of statistical analysts. These movements suggest opportunity for specially design interactive interfaces (Nguyen, Ho and Himodaira 2000) and query support (Konopnicki and Shmueli 1999), suitable to DM's increased access to local, distributed, and heterogeneous information resources. Moreover, increased professional activities, such as the European Symposia (Zytkow and Quafafou 1998; Zytkow and Rauch 1999), may help stabilize DM's boundaries. Whatever DM's future, the question put to DM investigators is whether more robust, more powerful algorithms can be provided that are computationally efficient and able to return results to the user that are both interpretable and valid.

Citations:

Bibliography

Abiteboul, S. (1997). Object database support for digital libraries. [On-line]. Available: http://rocq.inria.fr/~abitebou/pub/dl97.ps

Adriaans, P. D., & Zantinge, D. (1997). *Data Mining*. Reading, MA: Addison-Wesley.

Aggarwal, C. C., & Yu, P. S. (1999). Data mining techniques for associations, clustering and classification. In N. Zhong & L. Zhou (Eds.), *Methodologies for Knowledge Discovery and Data Mining, PAKDD-99* (pp. 13-23). Berlin: Springer.

Agrawal, R., & Shafer, J. C. (1996). Parallel mining of association rules: design, implementation and experience. IBM Research Report RJ 10004. *IEEE Transactions on Knowledge and Data Engineering, 8*(1), 962-969.

Agrawal, R., & Srikant, R. (2000, June). Privacy-Preserving Data Mining. In W. Chen, J. F. Naughton, P. A. Bernstein (Eds.), *Proceedings of the ACM-SIGMOD 2000 Conference on Management of Data. SIGMOD Record 29*(2), 439-450.

Agrawal, R., & Srikant, R. Fast algorithms for mining association rules. In M. Jarke & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB'94* (pp. 487-499). San Francisco: Morgan Kaufmann.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: a performance perspective. *IEEE Transactions on Knowledge Data Engineering, 5*(6), 914-925.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Inkeri Verkamo, A. (1996). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press.

Ahonen, H. Mannila, H. & E. Nikunen: Generating grammars for SGML tagged texts lacking DTD. In M. Murata & H. Gallaire (Eds.), *Proceedings of the Workshop on Principles of Document Processing (PODP) '94* . Darmstadt, 1994. [On-line]. Available: http://www.cs.helsinki.fi/u/hahonen/ahonen_podp94.ps

Ahonen, H., Heinonen, O., Klemettinen, M., & Inkeri Verkamo, A. (1997). Applying data mining techniques in text analysis. (Report C-1997-23). Helsinki: University of Helsinki, Dept. of Computer Science. [On-line]. Available: http://www.cs.helsinki.fi/u/hahonen/publications.html

Ankerst, M., Ester, M., & Kriegel, H-P. (2000). Towards an effective cooperation of the user and the computer for classification. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2000* (pp. 179-188). New York: ACM.

Arning, A. R., (1996). A linear method for deviation detection in large databases. *Proceedings of the 2nd International Conference on KD and DM* (pp. 164 ff). Menlo Park: AAAI Press/ MIT.

Arocena, G., Mendelzon, A., & Mihaila, G. (1997). Applications of a Web query language. In G. O. Arocena, A. O. Mendelzon, & G. A. Mihaila (Eds.), Proceedings of the 6th International WWW Conference. [On-line]. Available: http://www.scope.gmd.de/info/www6/technical/paper267/paper267.html

Atzeni, P., Mendelzon, A., & Mecca, G. (Eds.). (1999). The World Wide Web and databases : EDBT Workshop, WebDB'98. New York: Springer.

Auditore, P. J. (1999, Oct.). Turning up the software evolution. *Enterprise Systems Journal, 14*(10).

Baeza-Yates, R., & Ribeiro-Nero, B. (1999). *Modern information retrieval.* New York: ACM Press.

Baralis, E., & Psaila, G. (1999). Incremental refinement of mining queries. In M. Mohania, & A. M. Tjoa (Eds.), *Data Warehousing and Knowledge Discovery* (pp. 173-182). Berlin: Springer.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. In L. M. Haas & A. Tiwary (Eds.), *SIGMOD 1998, Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 85-93). New York: ACM.

Beeri, C., & Buneman, P. (Eds.). (1999). *Database Theory – ICDT '99.* (Lecture notes in computer science 1540). Berlin: Springer.

Beeri, C., & Milo, T. (1999). Schemas for interpretation and translation of structured and semi-structured data. In C. Beeri & P. Buneman (Eds.), *Database Theory – ICDT'99* (pp. 296-313). Berlin: Springer.

Beeri, C., Elber, G., Milo, T. Sagio, Y., Shmueli, O., Tishby, N., Kogan, Y., Konopnicki, D., Mogilevski, P. & Slonim, N. (1998). WebSite – a tool suite for harnessing web data. In P. Atzeni, A. Mendelzon & G. Mecca (Eds.), *International Workshop on World Wide Web and Databases, WebDB* (pp. 152-171). Berlin: Springer.

Benoit, G., & Andrews, J. E. (2000). Data discretization for novel resource discovery in large medical data sets. *Proceedings of the AMIA Symposium*, 61-65. Available: http://www.amia.org/pubs/symposia/D200636.pdf

Berry, M. J. A. & Linoff, G. (2000). *Mastering Data Mining: the art and science of customer relationship management*. New York: Wiley.

Berry, M. J. A., (1999a, June 22) "Mining the Wallet", *Decision Support* [On-line]. Available: http://www.intelligententerprise.com/992206/decision.shtml

Berry, M. J. A. (1999b, Oct. 26). The Privacy Backlash. *Intelligent Enterprise, 2*(15). Available: http://www.intelligententerprirse.com/992610/decision.shtml

Berry, M. J. A., & Linoff, G. (1997) *Data Mining Techniques for Marketing, Sales, and Customer Support*. New York: Wiley.

Berson, A., & Smith, S. J. (1997). *Data Warehousing, Data Mining, and OLAP*. New York: McGraw-Hill.

Berson, A., Smith, S., & Thearling, K. (2000). *Building Data Mining Applications for CRM*. New York: McGraw Hill.

Bettini, C. (2000). *Time Granularities in Databases, Data Mining, and Temporal Reasoning*. Berlin: Springer.

Biafore, S. (1999, Nov.). Predictive solutions bring more power to decision makers. *Health Management Technology, 20*(10) [On-line]. Available: http://healthmgttech.com

Bigus, J. P. (1996). *Data mining with neural networks: solving business problems - from application development to decision support.* New York: McGraw-Hill.

Bigus, J. P., & Bigus, J. (1998). *Constructing intelligent agents with Java: a programmer's guide to smarter applications.* New York: Wiley.

Bleyberg, M. Z. (1999). Preserving text categorization through translation. *IEEE SMC'99 Conference Proceedings. International Conference on Systems, Man, and Cybernetics.* (pp. 912-17). Piscataway, NJ: IEEE.

Bock, H.-H., & Diday, E. (Eds.). (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. New York: Springer.

Bonchi, F., Giannoti, F., Mainetto, G., & Pedreschi, D. (1999). Using data mining techniques in fiscal fraud detection. In M. K. Mohania, & A. Min Tjoa (Eds.), *Data Warehousing and Knowledge Discovery, DaWaK'99*. (Lecture notes in computer science 1676) (pp. 369-376). Berlin: Springer.

Bose, R., & Sugumaran, V. (1999). Application of intelligent agent technology for managerial data analysis and mining. *Data Base for Advances in Information Systems*, *Association for Computing Machinery, 30*(1), 77-94.

Bourne, P. E. (2000, June). Bioinformatics meets data mining: time to dance? *Trends in Biotechnology, 18*(6), 228-230.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communications of the ACM, 39*(11), 42-48.

Brachman, R., & Anand, T. (1996). The process of knowledge discovery in databases: a human-centered approach. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 35-57). Cambridge: AAAI Press/MIT Press.

Bramer, M. A. (Ed.) (1999). *Knowledge Discovery and Data Mining.* London: Institution of Electrical Engineers.

Brazma, A. (1999). Mining the yeast genome expression and sequence data. *Bioinformer* [On-line]. Available: http://bioinformer.ebi.ac.uk/newsletter/archives/4/lead_article.html

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. *Proceedings of the ACM SIGMOD*, 265-276.

Brown, A. D. (2000). High-bandwidth, low-latency, and scalable storage systems. [On-line]. Available: http://www.pdl.cs.cmu.edu/NASD/

Brown, T. J., & Mielke, P. W. (2000). *Statistical Mining and Data Visualization in Atmospheric Sciences*. Boston: Kluwer.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining: from concept to implementation.* Upper Saddle River, NJ: Prentice-Hall.

Caruana, R. & Freitag, D. (1994), *Greedy attribute selection.* In W. W. Cohen & H. Hirsh (Eds.), *Machine Learning: Proceedings of the Eleventh International Conference (pp.* 28-36). San Francisco: Morgan Kaufmann.

Chattratichat, J., Darlington, J., Guo, Y., Hedvall, S., Köler, M., Syed, J. (1999). An architecture for distributed enterprise data mining. *HPCN Europe 1999*, 573-582.

Chaudhuri, S. (1998, March). Data mining and database systems: where is the intersection? *Bulletin of the Technical Committee on Data Engineering, 21*(1), 4-8.

Chaudhuri, S., & Madigan, D. (Eds.). (1999). *The 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD-99*. New York: Association for Computing Machinery.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., Widom, J. (1994). The Tsimmis Project: integration of heterogeneous information sources. *Proceedings of the 100th Anniversary Meeting of the Information Processing Society of Japan*. Tokyo: IPS, 7-18.

Chen, L., & Rundensteiner, E. A. (1999). Aggregation path index for incremental web view maintenance. Available online: http://wpi.edu/pub/techreports/9933.ps.gz

Chen, M.-S, Han, J., & Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883. [On-line]. Available: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html

Chiang, D-A., Chow, L. R., & Wang, Y-F. (2000, June). Mining time series data by a fuzzy linguistic summary system. *Fuzzy Sets and Systems, 112*(3), 419-432.

Ching, J., Wong, A., & Chan, K. (1995). Class-dependent discretization for inductive learning from continuous and mixed mode data. *IEEE Transactions on Knowledge and Data Engineering, 17*(7), 641-651.

Chou, D. C., & Chou, A. Y. (1999). A Manager's Guide to Data Mining. *Information Systems Management, 16*(14), 33-42.

Church, K. W., & Rau, L. F. (1995). Commercial applications of natural language processing. *Communications of the ACM, 38*(11), 71-79.

Cios, K., Pedrycz, W., & Swiniarski, R. (1998). *Data Mining Methods for Knowledge Discovery*. Boston: Kluwer.

Clifton, C., & Couley, R. (2000) TopCat: data mining for topic identification in a text corpus. In M. J. Zaki and C.-T. Ho (Eds.), *Large-scale Parallel Data Mining*. Berlin: Springer, 174-183.

Coenen, F., Swinnen, G., Vanhoof, K., & Wets, G. (2000). The improvement of response modeling: combining rule-induction and case-based reasoning. *Expert Systems with Applications, 18*, 307-313.

Cooley, R., Srivastava, J., & Mobasher, B. (1997). Web mining: information and pattern discovery on the world wide web. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, ICTAI'97*. Menlo Park: IEEE.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning, 10*, 37-78.

Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos: IEEE.

Dasarathy, B. V. (2000). *Data Mining and Knowledge Discovery: theory, tools, and technology II*. Bellingham, WA: SPIE-International Society for Optical Engineering.

Date, C. J. (2000). *An Introduction to Database Systems*. (7th ed.). Reading, MA: Addison-Wesley.

Datta, P. (1997). Applying clustering to the classification problem. *Proceedings of the 14th National Conference on Artificial Intelligence* (pp. 82-87). Menlo Park: AAAI Press.

de Sousa, M. S. R., Mattoso, M., & Ebecken, N. F. F. (1999, December). Mining a large database with a parallel database server. *Intelligent Data Analysis, 3*(6), 437-451.

DeCoste, D., & Wagstaff, K. (2000). Alpha seeding for support vector machines. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 345-349.

Deogun, J. S., Raghavan, V. V., & Sever, H. (1995). Exploiting upper approximations in the rough set methodology. In U. Fayyad, & R. Uthurusamy (Eds.), *First International Conference on Knowledge Discovery and Data Mining* (pp. 69-74). Berlin: Springer.

Deogun, J. S., Raghavan, V. V., Sarkar, A., & Sever, H. (1996). Data mining. Trends in research and development. In T. Y. Lin, & N. Cercone (Eds.), *Rough Sets and Data Mining: Analysis of Imprecise Data* (pp. 9-46). Boston: Kluwer.

Devlin, B., (1996). *Data Warehouse, from architecture to implementation*. Reading, MA: Addison-Wesley.

Domingos, P. & Hulten. G. (2000). Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 71-80). New York: ACM Press. [On-line]. Available: http://www.kric.ac.kr:8080/pubs/contents/proceedings/ai/347090/

Dougherty, N., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In A. Prieditis & S. Russell (Eds.), *Machine Learning: Proceedings of the Twelfth International Conference* (pp. 194-202). San Francisco, CA, Morgan Kaufmann.

Elder, J. F. (2000). *Top 10 Data Mining Mistakes*. [On-line]. Available: http://www.dataminininglab.com.

El-Khatib, H. T., Williams, M. H., MacKinnon, L. M., & Marwick, D. H. (2000). A framework and test-suite for assessing approaches to resolving heterogeneity in distributed databases. *Information and Software Technology, 42*, 505-515.

Farnstrom, F., Lewis, J., & Elkan, C. (2000). Scalability for clustering algorithms revisited. *SIGKDD Explorations, 2*(1), 51-57.

Fayyad, U. (1998, March). Mining databases: towards algorithms for knowledge discovery. *Bulletin of the Technical Committee on Data Engineering, 21*(1) [On-line]. Available: http://www.research.microsoft.com/research/db/debull/98mar/fayyad6.ps

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous attributes for classification learning. In R. Bajcsy (Ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pp. 1022-1027). San Francisco: Morgan Kauffman.

Fayyad, U., & Uthurusamy, R. (1999, November). Data mining and knowledge discovery in data bases: introduction to the special issue. *Communications of the ACM, 39*(11), 24-26.

Fayyad, U., Haussler, D., & Stolorz, P. (1996, November). Mining scientific data. *Communications of the ACM, 39*(11), 51-57.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery. An overview. *Advances in Knowledge discovery and data mining*. Menlo Park: AAAI Press/MIT.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: toward a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD-96 (pp.* 82-88). Menlo Park: AAAI Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., (1996). KD and DM: Towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining.* Menlo Park: IAAA.

Feldman, R., Dagan, I., & Hirsch, H. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems, 10*, 281-300.

Feldman, R., Klosgen, W., Zilberstein, A. (1997). Visualization Techniques to Explore Data Mining Results for Document Collections. In D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy (Eds.) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, KDD 97* (pp. 16-23). Menlo Park: AAAI Press.

Feldman, R., Aumann, Y., Fresko, M., Liphstat, O., Rosenfeld, B., & Schler, Y. (1999). Text mining via information extraction. In M. J. Zaki and C.-T. Ho (Eds.), *Large-scale Parallel Data Mining* (pp. 165-173). Berlin: Springer.

Florescu, D., Levy, D., & Mendelzon, A. (1998). Database techniques for the world-wide web: a survey. *SIGMOD Record, 27*(3), 59-74.

Fong, J. (Ed.) (1997). *Data Mining, Data Warehousing & Client/Server Databases: Proceedings of the 8th International Database Workshop.* New York: Springer.

Fountain, T., Dietterich, T., & Sudyka, B. (2000). Mining IC test data to optimize VLSI testing. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2000*, 18-25.

Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: data structures & algorithms.* Englewood Cliffs, NJ: Prentice Hall.

Franstrom, F., Lewis, J., & Elkan, C. (2000). Scalability for cluster algorithm revisited. *SIGKDD Exploration, 2*(1), 51-57.

Frasconi, P., Gori, M., & Soda, G. (1999). Data categorization using decision trellises. *IEEE Transactions on Knowledge and Data Engineering, 11*(5), 697-712.

Freitas, A. A. (1998). *Mining Very Large Databases with Parallel Processing.* Boston: Kluwer.

Freitas, A. A. (1999, October). On rule interestingness measures. *Knowledge-Based Systems, 12*(5-6), 309-315.

Fu, Z. (1999, September). Dimensionality optimization by heuristic greedy learning vs. genetic algorithms in knowledge discovery and data mining. *Intelligent Data Analysis, 3*(3), 211-225.

Fukuda, T., Morimoto, Y., Shinichi, M., & Takeshi, T. (1999, February). Mining optimized association rules for numeric attributes. *Journal of Computer and System Sciences, 58*(1), 1-12.

Furtado, P., & Madeira, H. (1999). Analysis of accuracy of data reduction techniques. In M. K. Mohania & A. Min Tjoa (Eds.), *Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99.* (pp. 377-388). New York: Springer.

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: beyond document retrieval. *Computational Linguistics and Chinese Language Processing, 3*(2), 17-60.

Gardner, C., (1996). *IBM Data Mining Technology* [On-line] Available: http://booksrv2.raleigh.ibm.com/cgi-bin/bookmgr/bookmgr.exe/NOFRAMES/datamine/CCONTENTS

Gavrilov, M., Anguelov, D., Indyu, P., & Motwani, R. (2000). Mining the stock market: which measure is best? *KDD 2000*, 487-496.

Ge, X., & Smyth, P. (2000). Deformable Markov model templates for time-series pattern matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 81-90). Menlo Park: ACM.

Glymour, C., & Cooper, G. F. (1999). *Computation, Causation, and Discovery.* Menlo Park, CA: AAAI Press.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1996, November). Statistical inference and data mining. *Communications of the ACM, 39*(11), 35-41.

Goldman, J. A. Chu, W. W., Parker, D.S., & Goldman, R. M. (1999). Term domain distribution analysis: a data mining tool for text databases. *Methods of Information in Medicine, 38*(2), 96-101.

Goodacre, R. (2000). Kohonan artificial neural networks. [On-line]. Available: http://gepasi.dbs.aber.ac.uk/roy/koho/kohonen.html

Gray, J., Bosworth, A., Layman, A., & Pirahesh, H. (1995). *Data Cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals.* (Microsoft Technical Report MSR-TR-95-22). Redmond, WA: Microsoft.

Greening, D. R., (2000, Jan). Data mining on the Web. *Web Techniques, 5*(1) [On-line]. Available: http://www.webtechniques.com/archives/2000/01/greening/

Grossman, R. L., Baily, S., Kasif, S., Mon, D., Ramu, A. & Malhi, B. (1998). The preliminary design of Papyrus: a system for high performance, distributed data mining over clusters, meta-clusters and super-clusters. In P. Chan & H. Kargupta (Eds.), *Proceedings of the KDD-98 Workshop on Distributed Data Mining*, pp. 37-43. Menlo Park: AAAI. [On-line]. Available: http://www.lac.uic.edu/~grossman/papers/kdd-98-ddm.pdf

Grossman, R. L. (1996, December). Data mining challenges for digital libraries. *ACM Computing Surveys, 28*(4es). [On-line]. Available http://www.acm.org/pubs/citations/journals/surveys/1996-28-4es/a108-grossman/

Grossman, R., Kasif, S., Moore, R., Rocke, D., & Ullman, J. (1999). *Data Mining Research: Opportunities and Challenges. A report of three NSF workshops on mining large, massive, and distributed data.* [On-line]. Available: http://www.ncdm.uic.edu/M3D-final-report.htm

Groth, R. (1998). *Data Mining: a hands-on approach for business professionals.* Upper Saddle River, NJ: Prentice Hall.

Groth, R. (2000). *Data Mining: building competitive advantage.* Upper Saddle River, NJ: Prentice Hall.

Gunopulos, D., & Rastogi, R. (2000). Workshop Report: 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. *SIGKDD Explorations, 2*(1), 83-84.

Guo, Y., & Grossman, R. (1999). *High Performance Data Mining: scaling algorithms, applications, and systems.* Boston: Kluwer.

Hafaz, A., Deogun, J., & Raghavan, V. V. (1999). The item-set tree: a data structure for data mining. In M. Mohania & A. M. Tjoa (Eds.), *Data Warehousing and Knowledge Discovery* (pp. 183-192). Berlin: Springer.

Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann.

Hearst, M. (1999). Untangling text data mining. In *Proceedings of ACL'99.* [On-line]. Available: http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Hemmer, M. C., & Gasteiger, J. Data mining in chemistry. Available online: http://www.terena.nl/tnc2000/proceedings/10B/10b5.html

Hirota, K., & Pedrycz, W. (1999). Fuzzy computing for data mining. *Proceedings of the IEEE, 87*(9), 1575-1600.

Ho, T. B. (2000). *Introduction to Knowledge Discovery and Data Mining.* [On-line]. Available: http://203.162.7.85/unescocourse/knowledge/AllChapters.doc

Holmes, J. H., Durbin, D. R., Winston, F.K. (2000). The learning classifier system: an evolutionary computational approach to knowledge discovery in epidemiologic surveillance. *Artificial Intelligence in Medicine, 19*, 53-74.

Holsheimer, M., Kersten, M., Mannila, H., & Toivonen, H. (1995). A Perspective on databases and data mining. *First International Conference on Knowledge Discovery and Data Mining* (pp. 150-155). Menlo Park: AAAI Press.

Holt, J. D., & Chung, S. M. (2000). Mining of association rules in text databases using inverted hashing and pruning. In Y. Kambayashi, M. Mohania, and A. M. Tjoa (Eds.), *Data Mining and Knowledge Discovery* (Lecture notes in computer science 1874) (pp. 290-300). Berlin: Springer.

Hong, T-P., Kuo, C.-S., & Chi, S-C. (1999, November). Mining association rules from quantitative data. *Intelligent Data Analysis, 3*(5), 363-376.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. (Technical Report A32). Helsinki: Univ. of Technology, Laboratory of Computer and Information Science.

Hsu, W., Lee, M. L., Liu, B., & Ling, T. W. (2000). Exploration mining in diabetic patients databases: findings and conclusions. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2000*, 430-436.

Hurwitz Group. (1997, July 16). The Changing Role of Data Warehousing. *ENT, 11*(2) [On-line]. Available: http://www.entmag.com/papers.asp

Hyland, R., Clifton, C., & Holland, R. E. (1999). GeoNODE: Visualising news in geospatial contexts. [On-line]. Available: http://www.mitre.org/resources/centers/it/g061/geonode/AFCEA_GeoNODE_paper.html

IBM (1999). Data Mining: extending the information warehouse framework. [On-line]. Available: http://www.almaden.ibm.com/cs/quest/papers/whitepaper.html#data-mining

Imasaki, K. (2000). *A Survey of Parallel Data Mining.* [On-line]. Available: http://www.scs.carleton.ca/~kimasaki/DataMining/summary.htm

Imielinski, T., & Virmani, A. (1999). MSQL: a query language for database mining. *Data Mining and Knowledge Discovery, 3*(4), 373-408.

Imielinski, T., Virmani, A., & Abdulghani, A. (1999). DMajor – application programming interface for database mining. *Data Mining and Knowledge Discovery, 3*(4), pp. 347-372. Boston: Kluwer.

Inmon, W. H. (1996). *Building the data warehouse.* (2nd ed.). New York: Wiley.

Jacobs, P. S., & Rau, L. F. (1990). SCISOR: extracting information from on-line news. *Communications of the ACM, 33*(11), 88-97.

Jagadish, H. V., & Ng, R. T. (2000). Incompleteness in Data Mining. *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1-10.

Jagadish, H. V., Madar, J., & Ng, R. T. (1999). Semantic compression and pattern extraction with fascicles. In M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik & M. L. Brodie (Eds.), *VLDB'99 Proceedings of the 25th International Conference on Very Large Data Bases* (pp. 186-198). San Francisco: Morgan Kauffman.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999, September). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264-323.

Jiang, J., Berry, M., Donato, J. M., Ostouchov, G. and Grady, N. W. (1999, November). Mining consumer product data via latent semantic indexing. *Intelligent Data Analysis, 3*(5), 377-398.

Joshi, J. (1999). Mining server logs for user profiles. [On-line]. Available: http://www.cs.umbc.edu/~ajoshi/web-mine/tr1.ps.gz

Joshi, K. P. (2000). Analysis of data mining algorithms. http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm

Joshi, K. P., Joshi, A., Yesha, Y, & Krishnapuram. R. (1999). Warehousing and mining web logs. In *Proceedings of the second international workshop on Web information and data management.* (pp. 63-68). Menlo Park: AAAI Press. [On-line] Available: http://www.acm.org/pubs/citations/proceedings/cikm/319759/p63-joshi/

Joshi, M. V., Han, E-H, Karypis, G., & Kumar, V. (2000). Efficient parallel algorithms for mining associations. In M. J. Zaki and C.-T. Ho (Eds.), *Large-Scale Parallel Data Mining* (pp 83-126). Berlin: Springer.

Ju, P. (1997). *Databases on the Web: designing and programming for network access.* New York: M&T Books.

Kan, M-Y., & McKeown, K. R. (1999). Information extraction and summarization: domain independence through focus types. [On-line]. Available: http://www.cs.columbia.edu/~min/papers/sds/sds.html

Kargupta, H. & Chan, P. (2000). *Advances in Distributed and Parallel Knowledge Discovery*. AAAI Press.

Kargupta, H., Hamzaoglu, I., & Stafford, B. (1997a). Web based parallel/distributed medical data mining using software agents. *1997 Fall Symposium, American Medical Informatics Association*.

Kargupta, H., Hamzaoglu, I., & Stafford, B. (1997b). Scalable, Distributed Data Mining using An Agent Based Architecture. In D. Heckerman, H. Mannila & D. Pregibon (Eds.), *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, KDD-97*, pp. 211-214. Menlo Park: AAAI Press. [On-line]. Available: http://diadic1.eecs.wsu.edu/pubs.html

Kargupta, H., Hamzaoglu, I., Stafford, B., Hanagandi, V., & Buescher, K. (1996). PADMA: parallel data mining agents for scalable text classification. *Proceedings of the High Performance Computing on the Information Superhighway HPC-Asia '97*. IEEE, pp. 290-295.

Kargupta, H., Riva Sanseverino, E., Johnson, E., & Agrawal, S. (1998). The genetic algorithm, linkage learning, and scalable data mining. In H. Cartwright (Ed.), *Intelligent Data Analysis in Science: A Handbook*. Oxford: Oxford University Press.

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1996). Newsgroup extraction with websom method and browsing interface. [On-line]. Available: http://websom.hut.fi/websom/doc/websom.ps.gz

Keim, D. A. (1999). *Visual Techniques for Exploring Databases*. (Tutorial Notes) [On-line]. Available: http://www.dbs.informatik.uni-muenchen.de/~daniel/publication.html

Kennedy, R. L. (1997). *Solving Data Mining Problems through Pattern Recognition*. Upper Saddle River, NJ: Prentice Hall.

Khosla, I., Kuhn, B., & Soparkar, N. (1996). Internet searching using information mining. *Workshop on Research Issues on Data Mining and Knowledge Discovery. Proceedings of the 1996 ACM-SIGMOD International Conference on the Management of Data*. New York: ACM.

Kimball, R. (1999, Dec. 7). The Matrix. *Intelligent Enterprise, 2*(17). [On-line]. Available: http://www.intelligententerprise.com/

Kimball, R., & Merz, R. (2000). *The Data Webhouse Toolkit*. New York: Wiley.

King, R.D., Karwath, A., Clare, A., & Dehaspe, L. (2000) Genome scale prediction of protein functional class from sequence using data mining. In S. R. Ramakrishnan (Ed.) *The Sixth International Conference on Knowledge Discovery and Data Mining, KDD 2000* (pp. 384-389). New York: ACM Press.

Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. *Proceedings of the 10th National Conference on Artificial Intelligence*, 129-134.

Knight, K. (1999, November). Mining online text. *Communications of the ACM, 42*(11), 58-61.

Kohavi, R., & Sommerfield, D. (1995). Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In U.M. Fayyad & R. Uthurusamy (Eds.), *Knowledge Discovery And Data Mining (KDD-95) Proceedings* (pp. 192-197). Menlo Park: AAAI Press.

Kohonen, T. (1998). Self-organization of very large document collections: state of the art. In: L. Niklasson, M. Boden, & T. Ziemke, (Eds.), *Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks* (vol. 1, pp. 65-74). London: Springer.

Konopnicki, R., & Shmueli, O. (1999). WWW exploration queries. In R. Y. Pinter, & S. Tsur, (Eds.), *Next Generation Information Technologies and Systems. 4th International workshop, NGITS'99* (pp. 20-39). (Lecture Notes in Computer Science 1649). Berlin: Springer.

Kovalerchuk, B. (2000). *Data Mining in Finance: advances in relational and hybrid methods*. Boston: Kluwer.

Kryszkiewicz, M. (2000). Mining around association and representation rules. In J. Stuller, J. Pokorny, B. Thalheim, & Y. Masunaga (Eds.), *Current Issues in Databases and Information Systems* (pp. 117-127). Berlin: Springer.

Kuok, C. M., Fu, A., & Wong, M. H. (1998). Mining fuzzy association rules in databases. *ACM SIGMOD, 27*(1), 1-12.

Lakshmanan, L., Sadri, F., & Subramanian, I. N. (1996). A declarative language for querying and restructuring the Web. *Proceedings of the 6th International Workshop on Research Issues in Data Engineering: Interoperability of Nontraditional Database Systems (RIDE-NDS'96).* Menlo Park: IEEE.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM, 38*(11), 55-64.

Langley, P., Iba, W., & Thompson, K. (1990). An analysis of Bayesian classifiers. *Proceedings of the 8th National Conference on Artificial Intelligence, (AAAI-90)*, 223-228. Menlo Park: AIII; Cambridge: MIT Press, 1990.

Lattig, M. (1998, Nov. 8). The latest data warehouse buzz may be a bust. *InfoWorld, 21*(45).

Lavington, S., Dewhurt, N., Wilkins, E., & Freitas, A. A. (1999, June). Interfacing knowledge discovery algorithms to large database management systems. *Information and Software Technology, 41*(9), 605-617.

Lavrac, N. (Ed.). (1997). *Intelligent Data Analysis in Medicine and Pharmacology*. Boston: Kluwer.

Lee, C.-H., & Yang, H.-C. (2000). Towards multilingual information discovery through a SOM based text mining approach. In A.-H. Tan & P. Yu (Eds.), *PRICAI 2000 Proceedings of the International Workshop on Text and Web Mining*, 80-87. [On-line]. Available: http://textmining.krdl.org.sg/PRICAI2000/TWMproceedings.html

Lehnert, W., Soderland, S., Aronow, D., Feng, F., & Shmueli, O. (1994). An inductive text classification for medical applications. *Journal for Experimental and Theoretical Artificial Intelligence, 7*(1), 49-80 [On-line]. Available: http://www-nlp.cs.umass.edu/ciir-pubs/scamc_st.pdf

Lenzerini, M. (1999). Description logics and their relationship with databases. In C. Beeri & P. Buneman, (Eds.), *Database Theory ICDT '99* (pp. 32-38). Berlin: Springer.

Li, B. (1997). Parallel C4.5 (PC4.5). [On-line]. Available: http://www.cs.nyu.edu/~binli/pc4.5/

Liao, W. (1999, May). Data mining on the Internet. [On-line]. Available: http://www.mcs.kent.edu/~wliao

Lin, T. Y., & Cercone, N. (1997). *Rough Sets and Data Mining: analysis of imprecise data*. Boston: Kluwer.

Lingras, P. J., & Yao, Y. Y. (1998). Data mining using roughness theory. *Journal of the American Society for Information Science, 49*(5), 415-422.

Liu, H., & Motoda, H. (1998a). *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Boston: Kluwer.

Liu, H., & Motoda, H. (1998b). *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In R. Agrawal & P. Stolorz (Eds.), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 80-86). Menlo Park: AAAI.

Loshin, D. (2000, Feb. 9). Value-added data: merge ahead. *Intelligent Enterprise, 3*(3), 46-51.

Luba, T., & Lasocki, R. (1994). On unknown attribute values in functional dependencies. *Proceedings of the International Workshop on Rough Sets and Soft Computing* (pp. 490-497). Los Alamitos: IEEE.

Luvrac, N., Keravnou, E. T., & Blaz, Z. (1996). *Intelligent Data Analysis in Medicine and Pharmacology*. Boston: Kluwer.

Madria, S. K., Bhowmick, S. S., Ng, W.-K., & Lim. E. P. (1999). Research issues in web data mining. In M. Mohania & A. M. Tjoa (Eds.), *Data Warehousing and Knowledge Discovery* (pp. 303-312). Berlin: Springer.

Makris, C., Tsakalidis, A. K., Vassiliadis, B. (2000). Towards Intelligent Information Retrieval Engines: A Multi-agent Approach. In J. Stuller, J. Pokorny, B. Thalheim & Y. Masunaga, (Eds.) *East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications ADBIS-DASFAA* (pp. 157-170). Berlin: Springer.

Mattison, R. (1999). *Web Warehousing and Knowledge Management*. New York: McGraw Hill.

Mattox, D. (1998). Software agents for data management. In B. Thuraisingham, (Ed.), *Handbook of Data Management.* New York: Auerbach.

Meña, J. (1999). *Data Mining your Website*. Boston: Digital Press.

Mendelzon, A. O., Mihaila, G. A., & Milo, T. (1997). Querying the world wide web. *International Journal on Digital Libraries, 1*(1), 54-67.

Merialdo, P., Atzeni, P., & Mecca, G. (1997). Semistructured and structured data in the Web: going back and forth. *Proceedings of the Workshop on the Management of Semistructured Data ACM SIGMOD '97*. In W. Chen, J. F. Naughton, P. A. Bernstein (Eds.), *SIGMOD Record, 29*(2), 1-9.

Michalski, R. S., Bratko, I., & Kubat, M. (1998). *Machine Learning and Data Mining*. New York: Wiley.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM, 42*(11), 30-36.

Mobasher, B., Jain, N., Han, E., & Srivastava, J. (1996). Web mining: pattern discovery from world wide web transactions. (Technical report TR-96050). Minneapolis: University of Minnesota, Dept. of Computer Science.

Mohania, M., Tjoa, A. M. (Eds.) (1999). *Data Warehousing and Knowledge Discovery, DaWaK'99*. Berlin: Springer.

Moxon, B. (1996, Aug.) Defining Data Mining. *DBMS Online* [On-line serial.] Available: http://www.dbmsmag.com/9608d53.html.

Nahm, U. Y., & Mooney, R. J. (2000). A mutually beneficial integration of data mining and information extraction. *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 627-632.

Nakhaeizadeh, G., Reinartz, T., & Wirth, R. (1997). Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In G. Nakhaeizadeh (Ed.), *Data Mining: Theoretische Aspekte und Anwendungen* (pp. 1-33). Heidelberg: Physica.

Nestorov, S., & Tsur, S. (1999). Integrating data mining with relational DBMS: a tightly-coupled approach. In R. Y. Pinter, & S. Tsur, (Eds.), *Next Generation Information Technologies and Systems, 4th International Workship, NGITS'99* (pp. 295-311). Berlin: Springer.

Ng, M. K., & Huang, Z. (1999, June). Data-mining massive time series astronomical data: challenges, problems and solutions. *Information and Software Technology, 41*(9), 545-556.

Nguyen, T. D., Ho, T. B., Himodaira, H. (2000). Interactive visualization in mining large decision trees. In T. Terano, H. Liu, & A. L. P. Chen (Eds.), *Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2000* (pp. 345-348). Berlin: Springer.

Odewahn, S. (1999). Science efforts underway with DPOSS. Available online: http://www.astro.caltech.edu/~sco/sco1/dposs/science.html

Paliouras, G., Papatheodorous, C., Karkaletsis, V., Spyropoulos, C., & Tzitziras, P. (1999). From web usage statistics to Web usage analysis. *IEEE SMC'99 Conference Proceedings. International Conference on Systems, Man, and Cybernetics* (vol. 2, pp. 159-64). Piscataway, NJ: IEEE.

Papadimitriou, C. H. (1999). Novel computational approaches to information retrieval and data mining. In C. Beeri & P. Buneman, (Eds.), *Database Theory ICDT '99* (p. 31). Berlin: Springer.

Parthasarathy, S., Zaki, M., & Li, W. (1998). Memory placement techniques for parallel association mining. *Fourth International Conference on Knowledge Discovery in Databases* (pp. 304-308). New York: Springer.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999a). Efficient mining of association rules using closed itemset lattices. *Information Systems, 24*(1), 25-46.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999b). Discovery of frequent closed itemsets for association rules. In C. Beeri & P. Buneman, (Eds.), *Database Theory ICDT '99* (pp. 398-416). Berlin: Springer.

Pedrycz, W., & Smith, M. H. (1999). Linguistic selectors and their optimization. *IEEE SMC'99 Conference Proceedings. International Conference on Systems, Man, and Cybernetics.* (pp. 906-11). Piscataway: IEEE.

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, H. & Benner, M. (1999, October). Association, statistical, mathematical and neural approaches for mining breast cancer patients. *Expert Systems with Applications, 17*, 223-232.

Piatetsky-Shapiro, G. (1998). Data mining and knowledge discovery tools: the next generation. Boston: Knowledge Stream. Available online http://www.kdnuggets.com/gpspubs/dama-nextgen-98/sld001.htm

Pietrosanti, E., & Graziadio, B. (1997). Artificial intelligence and legal text management: tools and techniquees for intelligent document processing and retrieval. In *Natural Language Processing: Extracting Information for Business Needs* (pp. 277-291). London: Unicom Seminars.

Pinter, R. Y., & Tsur, S., (Eds.). (1999). *Next Generation Information Technologies and Systems. 4th International Workshop, NGITS' 99.* (Lecture Notes in Computer Science, 1649). Berlin: Springer.

Pravica, D. (2000a, Jan.) "Who do you want to know today? *Computing Canada, 26*(1).

Pravica, D. (2000b, March 17). Tracking the transactions is key to results – because all e-commerce relationships are mediated by computing applications, it's possible to develop data for every Web-based transaction. *Computing Canada* [On-line]. Available: http://www.findarticles.com/m0CGC/6_26/61888010/p1/article.jhtml

Provost, F. J., & Kolluri, V. (1999). A survey of methods for scaling up inductive learning algorithms. *Data Mining and Knowledge Discovery Journal, 3*(2), 131-169.

Pyle, D. (1999). *Data Preparation for Data Mining.* San Francisco: Morgan Kaufmann.

Quinlan, J. R., (1993). *Q4.5 Programs for Machine Learning.* San Francisco: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Ragel, A., & Crémilleux, B. (1999, October). MVC – a preprocessing method to deal with missing values. *Knowledge-Based Systems, 12*(5-6), 285-291.

Raghavan, V. V., Doegun, J. S., & Sever, H. (1998). Introduction to the special issue on DM. *JASIS,49*(5), 397-402.

Raghavan, V. V., Sever, H., & Doegun, J. S. (1994). A system architecture for database mining applications. In W. P. Ziarco (Ed.), *Rough Sets, Fuzzy Sets and Knowledge Discovery* (pp. 82-89). Berlin: Springer.

Reinartz, T. (1999). *Focusing solutions for data mining: analytical studies and experimental results in real-world domains.* (Lecture notes in artificial intelligence 1623). Berlin: Springer.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics, 61*, 241-254.

Robertson, A. M., & Gaizauskas, R. (1997). On the marriage of information retrieval and information extraction. In: J. Furner and D. J. Harper (Eds.), *Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research* (pp. 60-67). London: British Computer Society. [Online]. Available: http://lorca.compapp.dcu.ie/BCS_IRSG-97/sand1.html

Sarawagi, S., Thomas, S., & Agrawal, R. (1998). Integrating association rule mining with relational database systems: alternatives and implications. *SIGMOD'98*, 343-354. [On-line]. Available: http:/dev.acm.org/pubs/contents/proceedings/mod/276304/

Savasere, A., Omiecinski, E., Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases. In U. Dayal, P. M. D. Gray, & S. Nishio (Eds), *Proceedings of 21st International Conference on Very Large Data Bases VLDB95* (pp. 432-444). San Francisco: Morgan Kaufmann.

Schade, D., Dowler, P., Zingle, R., Durand, D., Gaudet, S., Hill, N., Jaeger, S., & Bohlender, D. (2000). A data mining model for astronomy. In N. Manset, C. Veillet, D. Crabtree (Eds.), *ASP Conference Series, Vol. 216, Astronomical Data Analysis Software and Systems IX* (p. 25). San Francisco: ASP.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, NJ: Lawrence Erlbaum.

Shen, L., Shen, H., & Cheng, L. (1999, September). New algorithms for efficient mining of association rules. *Information Sciences, 118*(1-4), 251-268.

Shintani, T., & Kitsuregawa, M. (2000). Parallel generalized association rule mining on large scale PC cluster. In M. J. Zaki and C.-T. Ho (Eds.), *Large-Scale Parallel Data Mining* (pp. 145-160). Berlin: Springer.

Simon, A. R. (1997). *Data Warehouse for Dummies.* Foster City, CA: IDG.

Simoudis, E. (1995). Reality check for datamining. IBM Almaden Research Center. Available: http://www.almaden.ibm.com/stss/papers/reality/

Skillicorn, D. (1999). Strategies for parallel data mining. *IEEE Concurrency, 7*(4), 25-35.

Small, R. D., & Edelstein, H. (2000). Scalable Data Mining. [On-line]. Available: http://www.twocrows.com/whitep.htm

Smith, G., (1999, Dec. 13). Improved information retrieval. *Information Week.* Available: http://www.informationweek.com

Smith, K. A., & Gupta, J. N. D. (2000). Neural networks in business: techniques and applications for the operations researcher. *Computers & Operations Research, 27*, 1023-1044.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free-text. *Machine Learning, 34*, 1-44.

Spaccapietra, S., & Maryanski, F. (Eds.) (1997). *Data Mining and Reverse Engineering: searching for semantics.* New York: Chapman & Hall.

Spertus, E. (1997). ParaSite: mining structural information on the Web. [On-line]. Available: http://www.mills.edu/ACAD_INFO/MCS/SPERTUS/Parasite/parasite.html

Stapley, B., & Benoît, G. (2000). BioBibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium on Biocomputing, 5*, 526-537. Available: http://www-smi.stanford.edu/projects/helix/psb-online/

Stolfo, S., Prodromidis, A. L., & Chan, P. K. (1997). JAM: Java Agents for Meta-Learning over Distributed Databases, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, (pp. 74-81). Menlo Park: AAAI Press.

Stolorz, P., & Musick, R. (1998). *Scalable High Performance Computing for Knowledge Discovery and Data Mining.* Boston: Kluwer.

Subramonian, R., & Parthasarathy, S. (1998). A framework for distributed data mining. In R. Agrawal & P. Stolorz (Eds.), *Proceedings of the Workshop on Distributed Data Mining, KDD98* [23].

Talavera, L., & Bejar, J. (1999). Integrating declarative knowledge in hierarchical clustering tasks. *Advances in Intelligent Data Analysis. Third International Symposium, IDA-99* (pp. 211-22). (Lecture Notes in Computer Science 1642). Berlin: Springer.

Tan, P.-N., Blau, H., Harp, S., & Goldman, R. (2000). Textual data mining of service center call records. *Knowledge Discovery in Databases, 2000*. Available: http://www-users.cs.umn.edu/~ptan/public.html

Terano, T., Liu, H., & Chen, A. L. P. (Eds.) (2000). *Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2000*. New York: Springer.

Thuraisingham, B. (1999). *Data mining: technologies, techniques, tools, and trends*. Boca Raton: CRC.

Toivonen, H. (1996). Sampling large databases for association rules. In T. M. Vijayaraman, A. P. Buchmann, C. Mohan & N. L. Sarda (Eds.), *VLDB'96 Proceedings of the 22th International Conference on Very Large Databases*, pp. 134-145. San Francisco: Morgan Kaufmann.

Trybula, W. J. (1997). Data mining and knowledge discovery. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology, 32*, 197-229.

Trybula, W. J. (1999). *Text Mining*. In M. E. Williams, (Ed.), *Annual Review of Information Science and Technology, 34*, 3-38.

Tsechansky, M. S., Pliskin, N., Rabinowitz, G. & Porath, A. (1999). Mining relational patterns from multiple relational tables. *Decision Support Systems, 27*(1-2), 117-195.

Tsukimoto, H. (1999). Rule extraction from prediction models. In N. Zhong, & L. Zhou (Eds.), *Methodologies for Knowledge Discovery and Data Mining, PAKDD-99* (pp. 34-43). Berlin: Springer.

Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining: a practical guide*. San Francisco: Morgan Kaufmann.

Weiss, S. M. & Kulikowski, C. A. (1991). *Computer Systems that Learn*. San Mateo: Morgan Kaufmann.

Westphal, C. R. & Blaxton, T. (1998). *Data Mining Solutions: methods and tools for solving real-world problems*. New York: Wiley.

Wilks, Y., Guthrie, L., & Slator, B. (1996). Information extraction: beyond document retrieval. In *Electric Words*. Cambridge: MIT Press.

Williams, G., Atlas, I., Bakin, S., Christen, P., Hegland, M., Marquez, A., Milne, P., Nagappan, R., & Roberts, S. (1999). Large-scale parallel and distributed data mining. (Lecture nodes in computer science, vol. 1759). Berlin: Springer.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*, 1-34.

Winter, R. (1999, Dec. 21). The E-Scalability Challenge. *Intelligent Enterprise, 2*(18) [On-line]. Available: http://www.intelligententerprise.com/992112/scalable.shtml

Witten, I. H., & Frank, E. (2000). *Data Mining: practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.

Wong, M. L. (2000). *Data Mining using Grammar-based Genetic Programming and Applications*. Boston: Kluwer.

Wong, W-C., & Fu, A. W-C. (2000). Incremental document clustering for web page classification. [On-line]. Available: http://www.cs.cuhk.hk/~adafu/Pub/IS2000.ps

Wright, P. (1996). Knowledge discovery preprocessing: determining record usability. In U. M. Fayyad, G. Piatetsky-Shaprio, & P. Smyth (Eds.), "From data mining to knowledge discovery: an overview" *Advances in Knowledge Discovery* (pp. 10-11). Menlo Park, CA: AAAI Press/MIT Press.

Xiang, Y, & Chu, T. (1999). Parallel learning of belief networks in large and difficult domains. *Data Mining and Knowledge Discovery* (vol. 3, pp. 315-39). Boston: Kluwer Academic.

Yang, J., Parekh, R., Honavar, V., & Dobbs, D. (1999). Data-driven theory refinement algorithms for bioinformatics. In *IEEE-INNS Proceedings of the International Joint Conference on Neural Networks*. Available: http://www.cs.iastate.edu/~parekh/papers/ijcnn99.ps

Zaki, M. J., & Ho, C.-T. (2000). *Large-scale Parallel Data Mining*. New York: Springer.

Zhang, L., & Zhang, B. (1999). Neural network based classifiers for a vast amount of data. In N. Zhong, & L. Zhou (Eds.), *Methodologies for Knowledge Discovery and Data Mining, PAKDD-99*. (pp. 238-246). Berlin: Springer.

Zhong, N., & Ohsuga, S. (1994). Discovering concept clusters by decomposing databases. *Data & Knowledge Engineering, 12*, 223-244.

Zhong, N., Skowron, A., & Ohsuga, S. (1999). *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'99*. Berlin: Springer.

Zweiger, G. (1999, November). Knowledge discovery in gene-expression-microassay data: mining the information output of the genome. *Trends in Biotechnology, 17*(11), 429-436.

□ ytkow, J. M., & Quafafou, M. (Eds.). (1998). *Principles of data mining and knowledge discovery: Second European Symposium, PKDD '98*. Berlin: Springer.

□ ytkow, J. M., & Rauch, J. (Eds.). (1999). *Principles of data mining and knowledge discovery: Third European Conference, PKDD'99*. Berlin: Springer.