6-1-1999

Knowledge Discovery in Databases

Melanie J. Norton
University of Southern Mississippi, Melanie.Norton@usm.edu

Knowledge Discovery in Databases

M. Jay Norton

Abstract

Knowledge discovery in databases (KDD) revolves around the investigation and creation of knowledge, processes, algorithms, and the mechanisms for retrieving potential knowledge from data collections. Related issues include data collection, database design, the description of entries in the database using the most appropriate representation, and data quality. This article is an introductory overview of knowledge discovery in databases. The rationale and environment of its development and applications are discussed. Issues related to database design and collection are reviewed.

Introduction

Development of techniques to investigate databases, or the contents of databases, is of significant interest. As data storage space becomes less expensive, data collection as a tool has become more accessible and more used. Organizations are literally stockpiling data in warehouses for future investigation. Research is being done to ascertain if there are patterns, not just within databases but within documents and disciplines, that contribute to knowledge retrieval.

Every discipline has borders that expand and contract with the practical and intellectual adventurism of its members. As the collective knowledge base has grown, it is apparent that aspects of one field cross into many other fields. The evolution of information technology also provides a bridge across disciplines—in its theories and applications to various

M. Jay Norton, School of Library and Information Science, The University of Southern Mississippi, Box 5146, Hattiesburg, MS 39406-5146
LIBRARY TRENDS, Volume 48, Number 1, Summer 1999, pp. 9-21

disciplines. Knowledge discovery in databases (KDD) is another manifestation of the expansion of investigative tools across fields of interest and applications.

Many disciplines contribute to the undertaking of KDD. Some are more cognizant than others of the many factors involved with data collection. This article is an overview of knowledge discovery in databases. Discussion of recurring concerns from different perspectives about the collection, classification, and quality of data related to applications of KDD is presented.

Databases and Knowledge Discovery

Dramatic improvements in information technology have encouraged the massive collection and storage of data in all areas from commerce to research. From operational databases where personnel data are kept; to transactional systems that track sales, inventory and patron data; to full-text document databases and more; databases are growing in size, number, and application. The enormous increase in databases of all sizes and designs is evidence of our ability to collect data, but it also creates the necessity for better methods to access and analyze data. Human capacity to handle the data available in these databases is not adequate for timely examination and analysis. Technology presents opportunities to maximize the use of these data in an economical and timely fashion. Attempts to improve the search and discovery processes when dealing with databases have generated significant interest across many fields resulting in a multidisciplinary approach. Knowledge discovery in databases employs diverse fields of interest including statistics, computer science, and business, as well as an array of methodologies, many still evolving: machine learning, pattern recognition, artificial intelligence, knowledge acquisition for expert systems, and more. Knowledge discovery in databases revolves around the investigation and creation of knowledge, processes, algorithms, and mechanisms for addressing the retrieval of potential knowledge. An important component of this activity is identification of patterns or trends, from metadata through, and including, the semantic level, which suggest an entity's relationships. KDD techniques have been successful with large-scale scientific databases, notably in astronomy to classify sky objects. In addition, techniques have been used in medical, environmental, political, and census research. Other applications have been made with industrial and business-oriented databases in marketing, finance, manufacturing, and Internet agents (Fayyad et al., 1996, pp. 37-38; Vickery, 1997, pp. 107-08).

The phrase "knowledge discovery in databases" is attributed to a 1989 workshop on KDD (Fayyad, 1996). The phrase was intended to clarify that the end result of investigating data should be the discovery of usable knowledge and to differentiate KDD as a whole process, not just one of its

components—i.e., data mining (Fayyad et al., 1996, p. 39). Knowledge discovery in databases encompasses all the processes, both automated and nonautomated, that enhance or enable the exploration of databases, large and small, to extract potential knowledge. The most commonly referenced component of these processes has been data mining which involves activities oriented toward identifying patterns or models in data representation, classification, semantics, rules application, and so on (Fayyad et al., 1996).

Emphasis that KDD is a whole process is intended to clarify that knowledge seeking in data collections involves intellectual and technological undertakings designed to seek useful knowledge and not merely stir data. Certain basic premises underlie these efforts: (1) knowledge is a relevant term rooted in individual information bases and needs; (2) finding patterns in data is not equivalent to discovering information; (3) data mining, to be effective, must be structured; (4) results of any discovery activity have to be evaluated within a context; (5) search mechanisms of this type of inquiry may require substantial iteration; and (6) many aspects of KDD are dynamic and interactive in application. While some facets of KDD are best served by technology, the ultimate evaluators and discoverers are the human agents generating the initial queries and directing the process (Fayyad, 1996; Fayyad et al., 1996).

Legacy and Design

Though one mission of KDD is to automate as many of the basic processes as possible, several factors impede progress in this sector. Intelligent data analysis techniques are still not sophisticated enough to resolve some data problems without intervention. The methods for identifying information appropriate to include in a database, adding it to the classification and organizational scheme of the database, and providing access points for retrieval are neither trivial nor uniform. Design and implementation of databases has relied on the purpose, scope, data characteristics, and technical limitations of the organization sponsoring the enterprise. The vitality of these databases has been dependent on the imposition of appropriate criteria for inclusion, characterization, and maintenance. Legacy databases designed for specific organizational tasks are rarely uniform in structure within a given enterprise, nor is there consistent data quality, representation, or depth. This diminishes the possibility of generalizing even tasks which are common to each discovery effort as the description of the database has to be customized, and variations in the construction and quality of the data accommodated (Raghavan et al., 1998; Deogun & Sever, 1998; Fayyad, 1996).

Databases are organized collections of data. They can typically be separated into reference or source databases. According to Rowley (1992): "*Reference databases* refer or point the users to another source (such as a

document, an organization or an individual for additional information or the full text of the document" (p. 14). These databases may contain citations, abstracts, addresses, and directory type information that allows the user to locate other resources. "*Source databases* contain the original source data" (p. 15) and may include a combination of numeric and text data such as corporate reports, stock information, pure numeric data such as statistics, or full-text documents (Rowley, 1992). It is common to use surrogates, which allow for locating information about an entity without having to interact directly with the primary entity or full-text data as a method to identify and manipulate the data. Such surrogates could be a title, a citation, an abstract, or any attribute that may be identified and associated with a specific entity. The surrogates may be what is to be manipulated in order to understand or react to the entities. An inventory database could be the collection of information that reflects the holdings of the business, the movement of the inventory, the stock, or the vendors. Sorting these data can yield information about inventory levels or the speed in which a vendor responds to orders. The data may be a surrogate or a representation for activities. Properly configured, it may be possible to use the database to model activities. For example, if a vendor takes longer to fulfill an order than another, it might be advisable to have an earlier reorder date attached to the stock of that vendor. Full-text databases may also contain full and complete documents and may or may not have a metadata descriptor set that includes subject fields, though most will have minimal fields such as author, title, and publication data. Databases are collections based on some relationship—maybe as basic as membership in the collection—that causes them to be placed in common or related files. The attributes that describe the entity are portions of an overall structure that should optimize the collection of data relevant to, and descriptive of, the entities. Consider a checkbook page; there is a column for a date, check number, item description, transaction amount, and transaction. Each column is an attribute and is intended to contain information that describes the checking transaction. Each row forms a single record—i.e., the fields or attributes that describe one entity. The entity is the checking transaction associated with one check number. When attributes appropriately and adequately describe an entity, it provides a better understanding of the entity and may reveal information about one entity's relationship to another. Information limitations, as well as constraints of space and money, impact database design. Collection of data may be based on weighing cost and need versus alternative resources and attempting to serve the most critical information needs. Designing databases takes into account what information resources the organization might require as well as the costs involved in time and technology in acquiring the data. Costs may be related to whether it is incidental to other activities, such as purchasing history collected at the checkout counter as part

of the inventory control program, or full text of documents acquired as part of the publishing process versus directed collection such as surveying. Designing database systems usually involves modeling the information environment and information mission for which the database is being implemented. Information limitations and costs are related to what is known and not known about the users, the environment, and the corpus of resources they might require now and in the future. Information needs, environments, and cost factors change over time.

A consideration of KDD database design and cost is data quality. The accuracy of the data's representation of the entity and environment from which it originated, as well as its currency, are factors of data quality. Orr (1998) uses the theoretical framework of the feedback-control system (FCS) to define data quality as "the measure of the agreement between the data views presented by an information system and the same data in the real world" (p. 67). His position is that data entered into databases and left unused for periods of time, without feedback, may become stagnant in comparison to what occurs to the entity the data originally represented. The lack of, or the failure to apply, feedback to data creates a discontinuity between static data and the continually changing world. For example, if the age of a person is recorded in a database but not the birth date or an aging algorithm, the age data remain the same and quickly become inaccurate in reflecting the age characteristic of the entity. If the attribute of age is not used, the error may go unnoticed, another aspect of Orr's contention that data that are unused may lose representativeness. Not utilizing data may result in not recognizing it has been erroneous or that it may never have been useful initially. Further, if data are used but do not specify criteria surrounding its acquisition, use, and maintenance, the value of the data will be decreased. Failing to include attributes to handle name changes, or failing to update a record when a name change is reported, reduces the accuracy of the record, may impede the location of the remainder of the record at a future time, and may miss data related to the name change pertinent to the record. Lack of rules governing the maintenance of data elements may cause reassignment of a field application without any overt documentation or clear history. For example, when it was discovered that there was no field to capture name changes, another field that did not seem much used might be informally redesignated as the name change field. At a later date, possibly using KDD techniques, it becomes apparent that the field in question was being used for recording the names of beneficiaries for specific insurance plans that were not grandfathered into the new plan. Data quality problems of this type will tend to multiply over time until the entire database's quality and usefulness is questionable.

What can happen to data in these situations can also happen to metadata. This perspective of data quality may be an issue of significant

importance in light of the trend toward data warehouses—i.e., if data are collected but unused, how accurate will it be by the time it is used (Orr, 1998)? A converse concern might be whether metadata records for documentary entities that contain subject descriptions should have the subject words modified to reflect new nomenclature or preferred subject terms? That is, what should happen if the environment of the information changes but the entity must remain the same? Should metadata change?

Intentional collection of data prior to identification of any specific purpose for it results from the recognition that information needs change over time and the data may be an unrecognized reservoir of knowledge. The emergence of data warehouses as a means to capitalize on an organization's data collection activities has potential as an advantage for KDD activities. "Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business" (Gardner, 1998, p. 54). If data warehousing is undertaken in a planned and logical manner, according to Fayyad et al. (1996), it could improve KDD opportunities and applications. With future KDD in mind, initial determination of how a data warehouse is designed, what attributes will be included, how the structures will be related or not will require more attention. Something to be considered is what information will be contained in the warehouse and how it can best be represented to require the least manipulation to access. If the warehousing organization will invest in uniform representation—methods for covering missing data and correcting errors—it will significantly decrease the preparation of data for KDD. Whether discussing databases or data warehouses, the underlying requirements to improve access are planning the collection, organizing, and rationally characterizing the structure for the best handling of the data with as much flexibility as possible. Some anticipation of what information will serve users in the future and how to provide access to the data without knowing what might be relevant in the future is the challenge (Sen & Jacob, 1998).

Currently, databases having KDD techniques applied to them were not necessarily built with this exploratory methodology in mind. Indeed, some of the techniques developed for KDD are in response to the lack of uniformity in database construction and omissions in retrieval capacity. Selection for inclusion in a database is based on user needs as they are identified in the construction process. If the construction process has not taken into account potential changes in information requirements, managerial decision path modifications, or new product data considerations, the database may have limited future value. The combination of currently serving users and forecasting what future services will be needed is a significant collection and design problem. Experience with legacy database data quality emphasizes the necessity of reviewing and improving planning and construction of databases and warehouses.

CLASSIFICATION

Historically, methods to provide access to the collected corpus of information have resulted in the imposition of artificial or controlled classification structures and languages. An example would be the development of classification systems such as Dewey Decimal or Library of Congress; both attempt to organize knowledge. Attached to these are subject heading or descriptor manuals—e.g., Library of Congress Subject Headings, MeSH (Medical Subject Headings), or specific thesauri. These tools suggest the classification and position in the hierarchy of knowledge of materials, permitting both assignment of subject and retrieval of subject by conformity to structured headings. Use of a controlled language in describing entities entered into a collection provides parameters to be employed in both building and searching. The subject headings and controlled language attempt to address the multiple layers of meaning which are part of language. The labeling of entities and meaning of words may change dramatically from discipline to discipline but also within subsets of disciplines and even over time. The imposed structure allows for information retrieval in relative proportion to the searchers' ability to manipulate the system and how well the information entered fits the structure. Use of controlled languages has resulted in using intermediaries to decode the systems imposed as the searchers were rarely those who classified the objects. The controlled structure also lent itself to application of information technology as uniform constructs are more easily manipulated by machines than natural language. Much machine searching currently relies upon matching input to some aspect of the database record. This may be simple and effective if the correct terms are entered into the searching algorithm—very resource intensive if the terms do not match and no internal algorithms allow for variations in the matching. When controlled language and related tools, such as indexes and thesauri, are implemented, it is possible to maximize the effectiveness of searches by using the controlled language. This naturally assumes that the language has been appropriately applied. When databases do not have controlled languages, the resources to search are more intensively expended, with varying results, dependent upon the searcher's ability to identify what terminology has been used to describe what they seek. The combination of machine limitations and the advantages of classification schemes impacted the design of early databases from both input and retrieval perspectives. Another method for organizing data for databases is embedded in the architecture of the database. By using a consistent structure exploiting the common attributes of the entities that are being entered into the database, it is possible to use the attribute structure for searching. For example, if the entity is an employee, then using attributes such as ID number, name, department of employment, supervisor, pension plan, or pay scale, could, if the searchware is properly designed, permit the searcher

to retrieve all employees from a given department and examine only their records or only the records of those on a specific pension plan. Early space and memory shortcomings restricted the amount and manner in which data could be stored. Data were "abbreviated" and arranged to maximize space savings. This resulted in using codes in the attribute fields and should have involved the application of value range rules. For example, the pension plan mentioned above may have been noted by a numeric code tied to a specific pension plan. In this way only a few bytes of space would be required to retain the information as long as somewhere there was a list (paper or electronic) of the code number associated with the pension. In a database of customer data, it might involve recording the inventory number of materials purchased rather than a textual description, or using a zip code as a region identifier rather than a street address. Applying rules to the content of the fields (attributes) would include specifying whether the whole name was in a single field or in two fields, and if there was a field size limit such that long names would be truncated and, if so, how; what date representation will be used—year first or last—and how many digits for the year? These concerns, coupled with the characteristics of the schema used, perpetuated the need for intermediaries. Now KDD is part of the intermediary force that can maximize the usefulness of such databases.

Classification and organization schemes are critical to any retrieval activity. To date these have been limited by technology, economics, knowledge, and tradition to selected access points usually identified by people who are not experts in the given discipline. Developing classification schemes to accommodate all knowledge has proven to be an evolutionary process. As understanding is gained as to the interrelatedness of our world, restrictive class structures have to be modified. Classification occurs at the database design level. Determining what attributes will describe an entity, the governing criteria, and the detail of description will affect what retrieval is possible. Seeking additional patterns that may be hidden within databases to generate new classification criteria via KDD is complex but less so than attempting to expand attribute descriptions to be complete classification structures, especially when some characteristics are not apparent without KDD. The ideal KDD evaluates data for trends or patterns that might be otherwise overlooked and, if statistical relevance is found, these may indicate subclasses or relationships. Such relationships might be used to further clarify and expand a database's value. Recognizing that this trend exists when it was previously unknown can provide new information value that poses new questions requiring further examination. Knowledge arises from prior knowledge.

Full-text searching algorithms for documents and textual databases are options but are still technologically cumbersome when working with any sizable database. Even when full-text searching becomes more reli-

able and economical, representations of documents will continue to be employed. Despite the power of web crawlers in recording and tracking pages, there is significant discourse about the use of metadata elements to represent pages. The bandwidth and time economy of identifying potentially desirable documents via surrogates, such as metadata or bibliographic record notations, is a significant savings. Full-text searching can be resource intensive and currently not particularly more effective than surrogate searches. Despite the advances in information technology, there are still difficulties with searching efficiently in large-scale databases. Seeking patterns in data is compromised if only portions of the data can be evaluated at a time, something that many searching algorithms overlook relative to large-scale databases. If the assumption is that all necessary data to consider is in memory, when it cannot physically be so, means detected trends or patterns are false results. Continued research and development to provide better algorithms for manipulating what may be tetrabytes of data in some rational manner is proceeding (Fayyad et al., 1996). Devising more complex sampling or modeling approaches using KDD techniques may yield some advantage. Certainly as the technology continues to advance, the application of pure brute processing power may be the answer.

The Processes

The actual processes are much more involved and complex than presented here. The following is a limited overview of what is a formidable undertaking. The application of KDD techniques requires substantial researcher involvement in determining the problem to explore and framing it within a meaningful context. Further, the investigator will, by necessity, be engaged in repetitive examination of the processes and results to direct or redirect the exploration. The discovery processes may dramatically influence the paths that research must follow.

Similar to any research endeavor, KDD requires defining the problem domain and acquiring underlying information relevant to the inquiry to identify the research path to follow. Establishing the parameters of the problem and determining the potential goal of the research is followed by the selection and possible extraction of the data set or subset to explore. Depending on the problem, a test set may be necessary to identify the best methodology. In fact, ascertaining the appropriate data to examine may in itself be a series of tasks and tests. The data set must then be prepared, and rules for dealing with missing data, erroneous data, redundant attributes, data corruption, and such must be determined and implemented. The problem of legacy databases—diverse platforms, errors introduced over time, changes in data entry procedures, poorly organized data, or technological limitations—must be adjusted for to enable statistical manipulation. (Perhaps the amount of preparation of data involved

in the data mining phase of knowledge discovery is why there has been so much focus on the data mining process.) The data set is then reduced or transformed, if appropriate, and/or standardized in representation and structure to enable manipulation, analysis, or modeling. There are a large number of possible algorithms to apply depending on the problem or the theorized pattern and the goal of the exploration. If patterns emerge, they must be analyzed, evaluated, and retested. Any or all of these processes may require repetition or modification along the way in response to difficulties encountered and findings that might influence the original theory. Selection of the methods to apply in a given situation is related to the intention of the research, amount and construct of the available data, as well as the quality of the data. There are no hard and fast rules governing the application of techniques beyond the appropriateness of one method to a particular domain or problem. Like any research effort, posing the correct or best inquiry will provide the best results (Fayyad et al., 1996). KDD is usually invoked to verify or discover; just pulling out data patterns is not sufficient. Some additional measures or tests are necessary to determine if data patterns have any value to the investigator, whether the pattern is an experimental phenomena or an actual recurring pattern. The expertise at this level of decision, whether there is any valuable meaning to derive from any detected patterns, has to come from the human investigator. By its nature, research requires human investment; curiosity is the fountain of knowledge discovery.

Reality Check

It would be a disservice to overlook two key points about knowledge discovery in databases. First, in the context presented, this is an emerging field, an evolving study, and not a finished product. Second, it is not a panacea for all the research interests or ills of the database universe. It does present momentous potential in its future incarnations in conjunction with evolving information technology, especially artificial intelligence areas. KDD is already demonstrating its value in its current state. Ongoing efforts to address the shortcomings of the data it examines and the technology employed will result in rapid advances. Many of the challenges facing KDD are the same as those which confront the entire information community:

- The multiple layers of data quality problems in databases resulting from design and implementation shortcomings present serious difficulty. When both modeling the information environment and the information mission have failed or have been incomplete, the database structure is inadequate to properly represent the data or to ensure consistency—e.g., a database containing student records that has no data field for name changes. In such case, a name change may not

be recorded or the new information lost, and any requests for the student with the new name will fail. When there are no rules for selecting data, no attempt to provide meaningful classification for the attributes, they do not describe the entities. For example, if customer names are to be kept in the database, but there is no rule for inclusion—such that sometimes names are entered as one field, sometime as two—it is possible for more than one record per customer to be created and for the search engine to be unable to match any of the records. If an inventory database has no timing attributes—that is, no date of inventory receipt or inventory decreases—then it is no more than a list of inventory which may or may not be present. Problems of data currency and accuracy in legacy databases, where the original collection of data may have been some time in the distant past and there have been no updates to the data, can damage the accuracy of the data. Lack of uniformity in the collection and loss of consistency as different entry systems changed field context can make a database unusable.

- Large databases with many attribute fields and variables pose complex and, as yet, unresolved computing and search difficulties. Memory management of these huge databases makes it difficult to analyze whole data sets at one pass, requiring different algorithms to perform analysis over smaller sets and still produce valid and reliable results. The sheer number of fields (attributes) in some databases make analysis extremely complex. Determining the influencing factors and fields to evaluate becomes a more sophisticated statistical problem as well as an information management problem.
- Increased complexity of relationships within databases requires more sophisticated search algorithms and more rigorous inspection techniques. This is a problem related to the depth of fields but also to the types of data collected. How datum fits into the database and its role in the information environment impacts our ability to analyze it. Why is datum included, and why is another not?
- Insufficient tools to incorporate prior knowledge into systems in more meaningful ways present special problems. What is the best way to make the available domain knowledge accessible to the search systems? Training systems to be expert systems is one approach, but this is still an evolving field and dependent upon human expertise. Further, as more is known, how should the system be adapted? Can we develop algorithms with sufficient robustness to adapt?
- Lack of historic platform integration and proprietary software restrictions contribute to the confusion and frustration of dealing with legacy databases. When databases are restricted to a specific platform and a specific software interface or program manipulation, it requires investment to adapt or overcome the barriers. Sometimes it means

re-entering the data in a less restrictive system, which increases the likelihood of data corruption.

- Certainly not the last nor the least challenge to KDD is the lack of information and knowledge about the human factors and roles in the construction, design, collection, classification, and retrieval related to databases (Fayyad et al., 1996). The construction of databases is a series of complex problems that include modeling the information environment as it is perceived and forecast at the time; identifying the appropriate information to incorporate into a database; and selecting the most representative attributes and the value limitations, to mention but a few. Each of these decisions is limited by technology—both the availability and the designers' ability to use it. Characterizing environments is a human activity that is embedded in belief systems, political perspectives, social concerns, and business acumen. The designers' background, comprehension of the information problem, the needs of the users, and the demands of the future all will impact the outcome of the database. How the construction is undertaken, and how the rules for validation and inclusion are composed and conveyed to the data collectors all affect the database contents. If the database design is exceptional and the interface used to enter data is poor, the result will be a poor database. The presence of thorough planning, construction, and implementation documentation is critical for those attempting KDD. Though often not available, such documentation could provide context to significantly improve the investigation. Even though computers collect data, the human actors design the database, create intellectual models for its construction and implementation and, ultimately, for its reinterpretation. Leaders in the KDD effort stress the role of human involvement in the retrieval processes. It is also a critical and little understood factor in the creation processes. Understanding the human cognitive processes involved in creating a search or determining what is the solution to a search is critical to the future of KDD. Better understanding of human cognition and pattern recognition could yield important clues to improved algorithms for computer cognition.

Conclusion

Emerging fields, new approaches, and knowledge discovery all herald change. Indeed, KDD does remold some aspects of research by implementation of a wide variety of tools from an array of disciplines; it advances the interrelatedness of the effort. The techniques have broad potential for application. Some aspects of KDD are indeed rediscovered knowledge. Some would argue that much of modern bibliometrics is kindred to KDD. Others might note that much of it is the application of basic statistics to another set of problems. It is clear that the techniques

are still very preliminary in their current applications, though many of the techniques have existed for some time. Legacy database design problems are bound to KDD because the techniques can detect some of them and because some of them complicate KDD. There is much more work to be done in this area. Clearly, more emphasis and research into designing databases and data warehouses is needed.

REFERENCES

Fayyad, U. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*(5), 220-225.

Fayyad, U.; Piatetsky-Shapiro, G.; & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Ai Magazine, 17*(3), 37-54.

Gardner, S. R. (1998). Building the data warehouse. *Communications of the ACM, 41*(9), 52-60.

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM, 41*(2), 66-71.

Pao, M. L. (1989). *Concepts of information retrieval.* Englewood, CO: Libraries Unlimited.

Raghavan, V. V.; Deogun, J. S.; & Sever, H. (Eds.). (1998). Introduction (In Special Topic Issues: Knowledge Discovery and Data Mining). *Journal of the American Society for Information Science, 49*(5), 397-402.

Rowley, J. E. (1992). *Organizing knowledge: An introduction to information retrieval.* Brookfield, VT: Ashgate.

Sen, A., & Jacob, V. S. (1998). Industrial-strength data warehousing. *Communications of the ACM, 41*(9), 29-31.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.