Exploring Data Mining Implementation

Karim K. Hirji

How large volumes of organizational data can be exploited for sustained competitive advantage.

Knowledge is the only factor of production that is not subject to diminishing returns. This inalienable truth is especially important in the current information age, where there is an extraordinary expansion of data generated and stored in computer databases for future access. Identifying potentially useful knowledge from such databases is no trivial task and is resulting in the growing interest in data mining by both practitioners and researchers.

To uncover relationships in data, statistical techniques such as factor analysis have been used in the past. Though traditional statistical techniques continue to be useful and effective for problems involving small data sets and a manageable number of variables, they run into a scalability roadblock when applied to problems where millions of records and thousands of variables exist. Data mining is thus emerging as a class of analytical techniques that go beyond statistics and aim at examining large quantities of data. What is important to keep in mind is the problems associated with data mining are fundamentally statistical in nature; that is, to infer patterns or models from data. In essence, data mining represents an umbrella or catch-all for a wide variety of techniques that aim at examining large quantities of data in search of easy-to-overlook relationships or hints that prove to have business or scientific value. A practical and applied definition of data mining is: the analysis and non-trivial extraction of data from databases for the purpose of discovering new and valuable information, in the form of patterns and rules, from relationships between data elements.

Data mining is receiving widespread attention in the academic and public press literature [5] and case studies and anecdotal evidence [9] suggest companies are increasingly investigating the potential of data mining technology to deliver competitive advantage. It appears interest in data mining is not waning and that at a minimum, its use in the current application areas such as direct target marketing campaigns, fraud detection, and development of models to aid in financial predictions will only intensify. According to the Palo Alto Management Group, the data mining segment is one of the fastest growing in the entire Business Intelligence market.

As a multidisciplinary field, data mining draws from areas such as artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics. Research into data mining has thus far focused on developing new algorithms [1] and on identifying future application areas [6]. Though both research into data mining technology and future application areas are important, the fundamental question in the minds of many early adopters is how to perform data mining. It is likely this question will take on greater importance as data mining becomes viewed as an

integral and necessary component of an organization's portfolio of analytical techniques. Research examining the question of how to do data mining is lacking. Attempting to help fill this void, the study described here provides the first test of the five-stage (or five-step) model proposed by Cabena et al. [4] on how to do data mining. This model is viewed as a type of theory of how to perform data mining and the results of this study suggest that in practice, a more elaborate set of stages is needed.

Basic Data Mining Concepts

Data mining in itself is not an end, but rather a means to an end. The benefits of data mining accrue from the operationalization of data mining results via a business strategy to achieve a specific objective. Although data mining draws from many different disciplines, it in fact has its roots in the statistical community, which predominantly focuses on inferring patterns or models from data, through a hypothesis-driven approach. In contrast, data mining is accomplished through a discovery-driven approach whereby no a priori hypothesis is stated for a particular problem under investigation.

The fields of machine learning, pattern recognition, and statistics have formed the basis for much of the developments in data mining algorithms. Through the pioneering work on Classification and Regression Trees (CART) by Breiman et al. [3], the statistical community has made an important contribution in legitimizing the use of decision trees in data mining for classification and regression. Decision trees are a way of representing a series of rules and consist of nodes and branches. In similar fashion, the field of pattern recognition, which emphasizes the creation of machines to perform tasks more accurately, faster, and cheaper than humans [7], has also made an important contribution to data mining by popularizing the use of neural networks. A feed-forward neural network is a network in which the nodes (or processing units) are numbered so that all connections go from a node to one with a higher number. In practice the nodes are arranged in layers—input, hidden, output—with connections only to higher layers.

Research into data mining algorithms has focused on developing new algorithms, adapting existing algorithms (to exploit memory and processor improvements), and extending the application of existing algorithms to new application areas. A thorough review of recent algorithmic developments is beyond the scope of this article but it is worth mentioning two recent developments to highlight some of the exciting research currently taking place. Agrawal et al. [1] have developed a computationally efficient association algorithm for discovering all significant association rules between items in a large database of transaction data. This contribution is important since previous association algorithms were found to scale poorly to large data sets. In another development, Han, Cai, and Cercone [8] developed an attribute-oriented induction method to extract generalized features of data by focusing on high-level concepts of the data rather than the primitive level of data. This generalization-based technique is leading to the development of a different set of data mining algorithms.

The multitude of data mining algorithms can appear confusing and even threatening. One way to better understand how they are different is by focusing on the three main data mining problem approaches: clustering, association/sequential pattern discovery, and predictive modeling. Clustering (or segmentation) is concerned with partitioning data records into subsets. A cluster is simply defined as a subset of data records and the goal of clustering is to partition a database into clusters of similar records such that records sharing a number of properties are considered to be homogeneous. To uncover affinities among transaction records consisting of several variables, association algorithms are used. These algorithms are used to solve problems where it is important to understand the extent to which the presence of some variables imply the presence of other variables and the prevalence of this particular pattern across all data records. Association algorithms discover rules of the form: if item X is part of a transaction, then for some percent of the time, item Y is part of the transaction. Finally, the predictive modeling data mining problem approach involves the use of a number of algorithms such as binary decision tree, linear discriminant function analysis, radial basis function, back propagation neural network, logistic regression, and standard linear regression. The goal of predictive modeling is either to classify data into one of several predefined categorical classes or to use selected fields from historical data to predict target fields.

To date, data mining has been applied in a number of diverse areas and specific data mining applications have even been developed (see [6]), yet no quantitative or qualitative study has been undertaken to understand how to actually perform data mining. Cabena et al. [4] have proposed a five-stage model of how to do data mining. The stages in this model are business objectives determination, data preparation, data mining, results analysis, and knowledge assimilation. Business objectives determination is concerned with clearly identifying the business problem to be mined; data preparation involves data selection, pre-

processing and transformation; data mining is concerned with algorithm selection and execution; results analysis is concerned with the question of whether anything new or interesting has been found; and finally, knowledge assimilation is concerned with formulating ways to exploit new information. Although a start, this model has some obvious shortcomings: namely, it is not based on any principles or existing body of research, and it is not supported by either quantitative or qualitative research.

Research Method

The study reported in this article is a test of a proposed model about how to do data mining. The model proposed by [4] is viewed as a sort of theory of how data mining should be done. In order to test this model, it was important to find a company willing to participate in this study and at the same time provide full access to the organization during the timeframe of the study. Following Benbasat et al. [2], the case study approach was deemed appropriate since this study was concerned with the larger question of developing a deeper understanding of "how" data mining should be done. The unit of analysis for this study was therefore a data mining project.

As mentioned, an important pragmatic requirement in this study was to find an interested company. Although a multiple-site design would have been preferred due to the generalizability of results, a single-site design was undertaken and this was deemed adequate since this study is a first test of the five-stage model by [4]. To build a rich perspective of how data mining is actually done in practice, multiple methods of data collection were used. While the outputs of these methods are not directly comparable, they did provide a greater area of coverage. The methods used were: archival records; documentation; interviews;[1] and observations.[2]

TAKCO provided an excellent research site since the study was of considerable interest to TAKCO management and staff. The researcher had full access to project team members and was permitted to participate in all stages of the project as a full-time observer. This provided the researcher with a unique opportunity to evidence that otherwise would have been inaccessible as well as, and more importantly, the ability to perceive reality from the viewpoint of someone "inside" the case study [10]. TAKCO is a mature North American fast-food retailer and the Canadian unit is headquartered in Toronto. In Canada, TAKCO is a recognized leader in the Canadian fast-food industry. In many ways, TAKCO is viewed as typical of firms in its industry. Some interesting aspects of the fast-food industry that were evidenced by project team member comments are: 1) it is consumer-driven, 2) firms strive toward operational efficiency and 3) there is an orientation toward extensive marketing analysis to understand and influence consumer choices. At the time this study had been initiated, the IT department at TAKCO was beginning a data mining implementation project that was sponsored by the marketing research department. On this note, it is also important to keep in mind that TAKCO management did not initiate the data mining project for the sake of this study but rather to gain an understanding of the benefits of data mining technology to enhance business decision-making.

Data was carefully collected, categorized (workshop notes, project documentation deliverables, data models, competitor analysis reports, and so forth), and analyzed after each site visit. Since the primary data collection method was direct observation, extensive notes were taken during each site visit and at all meetings. Comments made by project team members were recorded and whenever possible, probed further by asking questions such as "help me understand what you mean?" After each site visit, notes were reviewed for content to ensure an accurate portrayal of events and then logged to reflect each stage of the project. Final data analysis was undertaken after gathering qualitative data from all site visits. The analysis was structured by comparing the data to the model by [4] on a stage-by-stage basis. A total of 10 site visits took place and data collection began in July 1998 and was completed by November 1998.

Case Study Analysis

The results of the test of the model proposed by [4] about how data mining should be done suggest that the model in at least one case is insufficient. Interventions were needed, which suggest that a more elaborated set of stages for how data mining should be done is appropriate. In this study, the data mining project at TAKCO consisted of eight cross-functional team members: a data mining specialist, a project manager, a senior director of strategic planning (the executive sponsor), a research supervisor, a business analyst, an end-user analyst, a data architect, and a database administrator (DBA). One interesting aspect of this project is that it was the first time TAKCO had undertaken a data mining project. Consequently, the executive sponsor and project manager jointly decided, at the outset, that the entire team should be present during key project activities in order to facili-

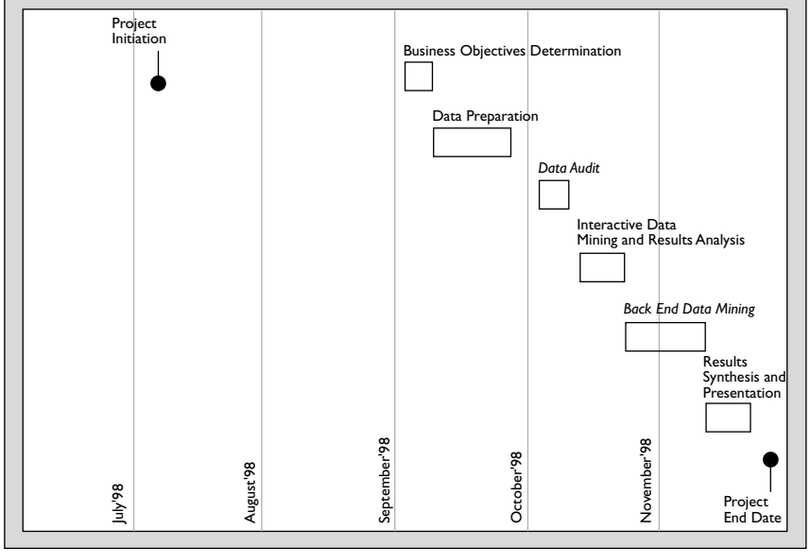
[1]The information gathered from these interviews is not reported for reasons of company confidentiality.

[2]Total participant observation could not be used in this study since time had to be spent on pursuing multiple methods of data collection rather than on participating in a particular task.

tate sharing of ideas. For that matter, the entire project team was assembled during actual execution of the data mining algorithms. Accordingly, the data mining activity was highly interactive as business end users were interpreting results as the data mining specialist generated them.

The relationship between the IT department and business end users at TAKCO was also remarkably cordial. In fact, eight months prior to the start of this study, the IT department had just successfully completed the implementation of an enterprise-wide product purchase transaction data warehouse. This same data warehouse was used for the project under study to provide the required 30 gigabytes of historical product purchase transaction data. To mine the data, a multipurpose industrial-strength data mining tool (such as the IBM Intelligent Miner for Data on AIX) was used as it afforded the opportunity to demonstrate the capabilities of the three distinct classes of data mining algorithms—namely clustering, associations, and predictive modeling. The data mining specialist from an external organization was well versed in the use of this particular tool. Finally, according to the executive sponsor, the data mining project at TAKCO was not deemed a failure as the project was completed on time (that is, before the project end date) and within the budget. It is quite likely that in some companies this same project might be viewed as a failure since completely new and unexpected results were not produced. The figure appearing here depicts a timeline of the project under study.

Actual project timeline.



During the first site visit, a formal meeting was held at TAKCO between the executive sponsor and the project manager to discuss the final parameters of the data mining project. Candidate business problems were identified by the executive sponsor based on the perceived business value of the solution to senior management. The issue of how much historical product purchase transaction data to mine was discussed and agreement reached. At the conclusion of this meeting, the project under study became a formal project. It received a formal project number and budget. The executive sponsor played an important role not only in securing project funding, but also in providing the political will to undertake this project.

The table appearing here represents a summary of the test of the five-stage model proposed by [4]. As shown, the planned project stages mapped very closely to the proposed project stages. However, given the nature of the project under study, the executive sponsor and project manager decided at the project initiation meeting to combine the separate data mining and results analysis stages and to drop the knowledge assimilation stage. Knowledge assimilation was dropped because the intent of this project was not to develop a production data mining application. A fourth and final planned project stage—results synthesis and presentation—was added and involved formally presenting the derived data mining results both to the entire project team as well as to the TAKCO project steering committee. Surprisingly, as shown in the table, this study found that the planned and observed actual project stages were not the same. Two additional stages—data audit and back end data mining—were added as intervening mechanisms during project execution to overcome three project discontinuity points that occurred during the planned project stages 1 and 3.

A workshop for the entire project team was held in early September 1998 to identify the business problems to be mined. In the workshop, the project team members were introduced, roles and responsibilities were assigned, and a high-level project plan was developed. Additionally, there was extensive discussion about the original set of candidate business problems. One of the immediate obstacles to overcome was to ensure that the data to support the data mining business problems was readily available. Input from the data architect and the DBA was invaluable during this stage of the project. In fact, two of the original three business problems were thrown out and two new ones were introduced because of serious data issues pre-

Comparison of proposed vs. actual stages.

Proposed Project Stages	Planned Project Stages	Planned Duration (in days)	Actual Project Stages	Actual Duration (in days)	Actual Days as % of Total	# of Site Visits
Business Objectives Determination	Business Objectives Determination	1	Business Objectives Determination	1	5%	1
Data Preparation	Data Preparation	6	Data Preparation	6	30%	1
Data Mining	Interactive Data Mining and Results Analysis	3	*Data Audit*	1	5%	1
Results Analysis	Results Synthesis and Presentation	3	Interactive Data Mining and Results Analysis	3	15%	3
Knowledge Assimilation			*Back End Data Mining*	6	30%	2
			Results Synthesis and Presentation	3	15%	1

venting the mining of these problems. One point worth noting is that during the workshop, the research supervisor played a dominant role in framing the business problems; at one point these even took on a "hypothesis phraseology."

As mentioned previously, an interesting finding from this study is that during the course of the data mining project, three distinct project discontinuity points emerged, which together with the resulting intervening mechanisms, led to changes to the planned project stages. Anticipation and Anxiety/Apprehension are the two project discontinuity points that occurred in stage 1. Anticipation refers to TAKCO project team member expectations about the potential of data mining to deliver novel and interesting findings. Early into the workshop, it became clear that many of the project team members had unrealistic expectations about what the data mining results would be. Both the research supervisor and end user analyst expected that the results would "… lead to identification of new product bundles so that the lunch time menu could be revolutionized." If left unchecked, it is quite likely that this project discontinuity point would have unfairly positioned the project at the polar end toward project failure, regardless of the derived data mining results. To overcome the Anticipation project discontinuity point, an intervening mechanism—goal alignment—was employed to provide focus and clarity for the project. Emphasis was therefore placed on establishing and reaching consensus on a realistic, measurable, and achievable management business goal and project goal. The agreed-to management business goal was: to gain an understanding of the benefits of data mining technology to enhance business decision-making; the specific project goal was: to demonstrate the potential of data mining technology to provide new and valuable insights into a subset of existing production system data.

Anxiety/Apprehension is the second project discontinuity point that also occurred during the business objectives determination stage. Both the technical and business members of the project team expressed concerns about the nature of the data preparation stage and the potential bias and noise that might be introduced into the data mining data set. Since the actual data mining algorithm execution was going to take place in a highly interactive mode, the data mining specialist planned to mine the data set once all of the required data was in the form of a single database table. After having just built an enterprise-wide product transaction data warehouse where a great deal of effort was dedicated to technical and business aspects of data quality, TAKCO project team members were concerned about data errors that might be introduced because of incorrect interpretation of data fields and improper transformations resulting in the production of a single table structure from an input star database schema. A data audit stage was therefore added after data preparation to demonstrate the validity, reliability, consistency, and integrity of the resulting transformed data set to be mined. In fact, both the data architect and DBA played key roles during the data audit stage to ensure that aspects such as data integrity, consistency, and completeness were preserved after data preparation was performed. With a formal data audit stage, the danger of automatically dismissing potentially anomalous and relevant data mining results and attributing them to data errors alone was minimized.

One interesting characteristic about TAKCO is the detailed knowledge TAKCO project team members possessed about the company's product offerings and their fast-food customer profiles. Determining customer profiles from product transaction data yields information that is uncertain and ambiguous. The Research Supervisor at TAKCO utilized statistical

tools to perform detailed product transaction analysis on an ongoing basis. Both the end-user analyst and business analyst were also power OLAP users. It was not surprising therefore that Frustration, as a project discontinuity point, occurred during the interactive data mining and results analysis stage. "I already know that" comment was made frequently by the Research Supervisor in response to presented data mining results. Overcoming this project discontinuity point was unquestionably critical for the project under study, thus an intervening mechanism was employed to ensure the management business and project goals were met. Back end data mining, which involved data enrichment and additional data mining algorithm execution by the data mining specialist, was introduced as a separate stage. The intent of this additional stage was to increase the dimensionality of the data mining data set with third-party demographic data and then to have the data mining specialist perform additional data mining off-site. It was found that supplementing the original data set with third-party demographic data that TAKCO had not previously considered was effective in providing different and interesting analysis results.

Practical Implications and Discussion

This study provides new insight into how data mining should be done and thus there are a number of practical project planning, management, and execution recommendations that managers and practitioners can follow on future data mining implementation projects. First and foremost, a data mining project appears to follow a more elaborated set of stages than what has been previously reported. In general, the relevant stages a data mining project should follow, where the project focus is not on building a production data mining application, are Business Objectives Determination, Data Preparation, Data Audit, Interactive Data Mining and Results Analysis, Back End Data Mining, and Results Synthesis and Presentation. At the outset of any data mining project, customer expectations must be well understood and where appropriate, during the Business Objectives Determination stage, misperceptions about the potential of data mining technology must be corrected. Unrealistic customer expectations about data mining will not position a data mining project with the greatest opportunity for realizing project success. As far as customizing the set of six stages data mining projects similar to that under study at TAKCO should follow, the Back End Data Mining stage affords the greatest flexibility. The duration of this stage can be increased or decreased (or even omitted altogether) depending on the sophistication of domain experts involved in evaluating and analyzing data mining results. The "I already know that" comment by the Research Supervisor suggests the threshold for data mining is it produce not previously known knowledge. Domain experts are not going to be satisfied with discovering knowledge they already have. Thus, including a Back End Data Mining stage and dedicating more project time to it to enrich the input data set and to execute additional mining runs is a worthwhile project strategy to follow in situations where data mining has already been used and where domain experts are extremely knowledgeable about the data set to be mined.

Second, the findings from this study do not corroborate existing work [4] that found data preparation to be the most resource intensive stage. Previous work on the distribution of effort in a data mining project suggests that around 60% to 70% of the total effort is dedicated to data preparation [4]. For the project under study, a total of 20 actual days of effort across six stages was required. As shown in the table here, 45% of the total project effort was consumed by data mining analysis (Interactive Data Mining and Results Analysis and Back End Data Mining) whereas only 30% was consumed by Data Preparation. One of the possible explanations for this result may have to do with the presence of a data warehouse. Although a data warehouse is not a necessary requirement for performing data mining, at TAKCO, the existence of a data warehouse in advance of pursuing data mining appears to have altered the distribution of project effort away from the data preparation stage. This finding should not be entirely surprising since the presence of a data warehouse implies greater organizational discipline toward data management and therefore less project time should be required to prepare the data prior to actual data mining algorithm execution. For a specific data mining project where a data warehouse is not present, it is possible that more project resources would be required to perform tasks such as selecting, cleaning, transforming, coding, and loading the data.

Third, there are a number of important process aspects relevant for the execution of the Interactive Data Mining and Results Analysis stage. "I don't understand what this means" comment was made by the Research Supervisor in reaction to presented clustering results on the first day of interactive data mining. No data mining briefing was held at the beginning of the project under study to provide TAKCO project team members background information on both data mining algorithms and the IBM Intelligent Miner data mining tool. It is likely such a briefing would have made the Interactive Data Mining and Results Analysis stage more efficient and

effective. Another important process aspect concerns the experience and skills of a data mining specialist. In addition to possessing detailed data mining technology skills, the data mining specialist for this project was also experienced as a facilitator, which proved to be invaluable. Facilitation skills provide control and focus during interactive discussions and it is quite likely that during the interactive data mining stage of this project, discussions would have become unwieldy without the presence of the data mining specialist who was also a skilled and experienced facilitator.

Linking data mining results with business strategy and using application software to perform sensitivity analysis of results obtained to demonstrate how data mining results support business strategy are also important process aspects of executing the Interactive Data Mining and Results Analysis stage. TAKCO is intent on positioning itself in the fast-food industry through a product differentiation strategy, therefore identifying patterns of product combinations as a basis of developing strategies for recombining some product offerings was an important business question that was pursued. Evaluating potential product recombinations from the perspective of TAKCO's business strategy was at a minimum helpful in eliminating results with potentially low business value. Furthermore, comments such as "…this is the kind of stuff I need to know" and "This is good stuff. I can use this in a presentation to support my argument" by the Senior Director of Strategic Planning provide compelling evidence for the importance of contextualizing the Interactive Data Mining and Results Analysis stage with business strategy. Finally, the use of spreadsheet application software to perform sensitivity analysis of potential cross-selling opportunities during analysis of clustering results proved to be highly effective according to project team members. Non-IT project team members from TAKCO were less concerned with 'how' the data mining results were generated than with an analysis of the results using tools they were most familiar with. For example, the Business Analyst, in response to viewing decision tree rules, remarked "how do we turn the rules into business logic?" The benefits of data mining accrue from the operationalization of data mining results via a business strategy. Thus, use of spreadsheet application software to demonstrate how data mining results support business strategy was found to be effective and moreover it provided business end users with a familiar environment to perform business sensitivity analysis.

Directions for Future Research

This study focused on the fundamental question of how data mining should be done. In addition to providing a deeper understanding of data mining implementation, this study points to a number of other research questions. An obvious direction for future research should be first to focus on corroborating this study's findings by undertaking a qualitative cross-sectional study that includes sample companies from other industries as well as other companies from the fast-food industry. With a refined model of how to do data mining, a further test of how to do data mining can be explored through a larger quantitative study. Collectively, these results can be integrated to develop a set of best practices that practitioners can follow when executing data mining implementation projects. Second, this study is novel in that data mining algorithm execution and interpretation of results were performed in a highly interactive fashion. As results were generated, business end users were evaluating them. Further exploration is needed to understand whether this is an efficient and effective way to examine large volumes of possible relationships. In other words, the question of should business end users be present to interpret results as IT specialists are generating a new set of data patterns needs to be explored. Finally, the premise of much of the writing on data mining is that new information will be found after having mined a particular data set. This study has revealed that the threshold of newness may be higher in companies where domain experts are already involved in extensive data analysis using analytical techniques other than data mining. Further investigation is therefore needed to confirm whether in fact this is the case. C

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. Fast discovery of association rules. In Fayyad, U., et al, Eds. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, 1996.
2. Benbasat, I., Goldstein, D.K. and Mead, M. The case research strategy in studies of information systems. *MIS Quarterly 11*, 3 (1987), 369–386.
3. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA, 1984.
4. Cabena P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. *Discovering Data Mining: From Concept to Implementation.* Prentice Hall, Englewood Cliffs, NJ 1998.
5. Chen, M.S., Han, J., and Yu, P.S. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering 8*, 6 (June 1996), 866–883.
6. Fayyad, U., Haussler, D. and Stolorz, P. Mining scientific data. *Commun. ACM 39*, 11 (Nov. 1996), 51–57.
7. Fukunaga, K. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, 1972.
8. Han, J., Cai, Y. and Cercone, N. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. on Knowledge and Data Engineering, 5* (1993), 29–40.
9. Hirji, K.K. Mining for data. *Today's Engineer 2*, 2 (1999), 32–35.
10. Yin, R.K. *Case Study Research: Design and Methods*. SAGE Publications, Inc., 1989.

Karim K. Hirji (khirji@ca.ibm.com) is a consultant with IBM Canada Ltd. in Markham, Ontario, Canada.