Discovering Data Mining: From Concept to Implementation

Karim K. Hirji
IBM Canada Ltd.
3600 Steeles Avenue East
Markham, Ontario L3R 9Z7 Canada
905.316.4029
Khirji@ca.ibm.com

ABSTRACT

This paper is a review of the book Discovering Data Mining: From Concept to Implementation – Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi (New Jersey: Prentice Hall, 195 pp., 1998).

Keywords

Data mining, Book review.

1. INTRODUCTION

What do Cabena et al. know about data mining that practitioners, consultants, managers and academics will find useful? A lot! is my answer based on reviewing this easy to read seven chapter book that incorporates over 30 years of combined practical lessons learned by the authors. One of the central themes in this book is the process oriented view of data mining advocated by the authors and later operationalized in a chapter focusing on two detailed case studies. This book however, does not provide an assessment tool for evaluating an organizations' readiness to adopt data mining technology and, like others dealing with the subject of data mining, is a mixture of "good" and "could have been better" parts.

2. REVIEW

Chapters One and Two comprise Part I (i.e., Introduction) of the book and it is here that the authors introduce the building blocks of data mining. Chapter One begins with a vignette of intimate customer-firm relationships that existed over some 50 years ago and argues that although these customer centric relationships do not exist by default today, firms are quickly realizing their potential for economic value and thus are striving towards developing unique customer-firm relationships. The authors then introduce data mining as a fundamental piece in the modern business decision making infrastructure that can assist firms in developing unique customer-firm relationships. In their definition of data mining, the authors neither provide a context for their definition nor do they contrast their definition with others.

Positioning data mining in relation to decision support systems, expert systems, executive information systems, management support systems and knowledge discovery in databases is also not addressed. Furthermore Cabena et al. engage in very limited discussion about the relationship between statistics and data mining.

Chapter Two is a non-exhaustive overview of commercial data mining use since the authors focus only on the consumer goods sector. Data mining applications for marketing management, risk management and fraud management are covered in this chapter. One of the points in the previous chapter that was introduced but not developed is the notion that the data mining process does not end with results but rather the key is to operationalize results, via a strategy, to achieve specific objectives. The authors cite an example of data mining used for forecasting financial futures which nicely illustrates the link between data mining and strategic decision-making. The two concluding topics of this chapter deal with emerging application areas and the danger of misinterpreting data mining results. This last topic is definitely out of place since Chapter Three includes a specific section on results analysis. The discussion on emerging data mining application areas, although thought provoking, was not comprehensive as it focused only on text mining and web analytics.

Chapters Three through Five form the second part of the book that is aptly labeled Discovery. Chapter Three is the centerpiece of the book as it is concerned with instilling discipline in the firm's experience with data mining. Cabena et al. propose a five step linearly sequential data mining process which is likely the product of their practical experiences with data warehousing and data mining projects. Their five step process – Business Objectives Determination, Data Preparation, Data Mining, Analysis of Results, Assimilation of Knowledge - is conceptually logical and their detailed description of the steps in their process includes not only a discussion of the various tasks involved in each step but also identifies appropriate involvement of various technical specialists in each step. One relevant point that is lacking in the authors' description of the business objectives determination step is the importance of correctly framing the problem/opportunity. The reader should be aware that the data mining process proposed by the authors has one significant shortcoming in that it does not reflect the iteration that typically takes place between data mining and results analysis. In addition, the authors include a chart depicting the distribution of effort across their five step process whereby data preparation accounts for over 50% of the total data mining effort. Readers should exercise caution when interpreting this result since it is not corroborated by empirical research. Reading this chapter answers many "procedural type" of questions about data mining however, one question that remains unanswered is the relationship between the five step process proposed by the authors and the Knowledge Discovery in Databases (KDD) process [1].

Face to Face with the Algorithms is the title of Chapter Four. This chapter tackles the various data mining algorithms and begins by introducing a framework for grouping the various algorithms and then continues with a description of the use of specific data mining algorithms for specific problem areas. Generally the authors do an excellent job of introducing and describing the algorithms by way of example while mathematical terms are only gingerly interspersed. This book is written for a wide target audience, thus omitting technical details about a

particular technique – specific implementation of an algorithm - is consistent. In some cases however, it would have been appropriate during the course of the discussions for the authors to infuse contrasts between certain techniques and statistics. After reading the chapter, the reader is not clear about the comparable computational efficiencies and effectiveness of different algorithms for different problem spaces.

Chapters Five through Seven focus on operational aspects of data mining. Chapter Five, which is included in the second part of the book, is concerned with evaluating vendor solutions. Re-organizing the book with chapters five through seven as the final part of the book would have made more logical sense. Moreover, placing the current chapter five as the final chapter of the book would have been perfect since it would have followed a description of case studies and data mining challenges. Chapter Five provides a brief overview of data mining tools, applications and consulting services. Moreover it provides fundamentally important considerations that the reader should keep in mind during the selection process for data mining tools and data mining applications. The prescriptions in this chapter are generally comprehensive, relevant and useful.

Chapter Six is concerned with describing the use of data mining in fraud management and in marketing management. The two case studies – preventing fraud and abuse and improving direct mail responses - are described in the context of the five step data mining process proposed by Cabena et al. Although both case studies are somewhat stylized, the authors provide an end to end perspective of data mining that is reasonably complete. In the marketing management case study, they even perform cross validation of data mining results via different techniques to better understand the direct mail response data! This chapter will undoubtedly be extremely useful to readers interested in concrete examples of data mining in practice.

The book concludes with a chapter that is concerned with two main topics - challenges to data mining and the importance of planning for data mining. Business, technical and social challenges in data mining are addressed and the inclusion of a separate discussion about the ethical use of traceable, individual data records is most appropriate as it hits home the message about responsible data mining use. My final thoughts on this book are that Cabena et al. provide a glimpse into an exciting industry that is relatively immature, without vendor dominance and that has no established industry standards or benchmarks. The book, although far from being the "authority" on data mining, prompts the reader to ask numerous questions and this makes the reading productive, enjoyable and stimulating.

3. REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, Vol. 39, No. 11, pp. 27-34.

About the Author

Karim is currently a Consultant in the Business Intelligence Consulting Services area of IBM Canada Ltd. His project consulting work focuses on providing data warehousing and data mining solutions to both private and public sector clients. Karim joined IBM Canada Ltd. as a professional hire in 1996; he was previously a Technology Consultant at the Bank of Montreal's Institute for Learning.

Karim received the B.Sc. (Co-op) degree in Mathematics and Computing Science from Simon Fraser University, and the M.M.S. (with distinction) degree in Management of Technology from Carleton University. Karim also has a graduate engineering degree from the University of Waterloo.

Karim's current research interests include knowledge mining, representation and management; processes for effective data mining; informatics; data quality; and technology and innovation management. His research work has appeared in journals such as IEEE Transactions in Engineering Management and R&D Management. Karim is a member of ACM, the Academy of Management, IEEE, and INFORMS.