



[REVIEWED FOR YOU]

mayato book recommendations in the area of business intelligence and customer relationship management

Title of book: **Tapping Into Unstructured Data**
 Author(s): **William H. Inmon, Anthony Nesavich**
 Details: **in English / 240 pages / Prentice-Hall 2008 / Price U.S.\$44.99**
 Reviewed by: **Dr. Marcus Dill**



1 Recommendation

The book is a plausible, compact, and stimulating weekend read for business intelligence (BI) experts, and a recommended introduction for readers who want to find out about opportunities to integrate textual content into existing data analysis environments and procedures. Within this topic, the authors offer many simple and inspiring thoughts. The book establishes an ideal mindset for starting a project on integrating texts into business intelligence. However, it remains on an abstract, general level, steering clear of implementation or architecture details almost completely. It is therefore hardly suitable as a practical handbook. In view of the book's rather fuzzy title, some readers may have greater expectations (and hope to read about, for example, statistical methods or unstructured data other than texts) but they will be disappointed, because the book clearly focuses on texts and how they can be prepared for analyses in the classic BI environment.

2 Summary of content

The authors first provide an overview of the characteristics of unstructured data and compare the treatment of structured and unstructured data. The lion's share of the book is devoted to the question of how unstructured data can be integrated into the world of structured data analysis. The authors explain the special challenges and the fundamental possible solutions associated with this task. They use the term "textual ETL" in analogy to the extraction, transformation, and loading (ETL) processes in a data warehouse. Much of the book deals with questions about identifying key words in texts and how these are mapped and purged. The broad area of textual analysis and textual search is only touched upon. At the end of the book, there are five fictitious scenarios from companies and organizations, whose applications and strategies are illustrated.

3 Assessment criteria

Content: Inmon and Nesavich provide a systematic overview of processing unstructured or semistructured textual data for the purposes of further analyses. Central to the book is the idea of processing texts in such a way that information (key words, etc.) can be extracted with the help of the conventional BI technologies developed for structured data. Even though the title emphasizes unstructured data in general, it becomes clear in the preface that the book is restricted to texts, and other types of unstructured data are not included. Of course, recordings of speech and other sounds – as well as images – also contain information that companies could potentially be interested in. And some readers might have been interested in ways of automatically extracting information from such data. The book's content is characterized by its systematic and simple approach. Many statements are of almost axiomatic succinctness; some may even sound trivial. This simplicity brings many critical aspects home to the reader. For anyone who plans to get to practical grips with this topic in the near future, reading this book will definitely get them in the mood. However, algorithms – or even implementations – are examined only at an abstract level in most of the book. Even the case studies at the end of the book offer little specific information, which, presumably, readers cannot really transfer to their own projects. Understanding textual analytics as text ETL (extraction, transformation, and loading) does not do justice to the business-based and technical breadth of the area – but it definitely reflects the attitude that many business intelligence experts have toward this topic: Texts must be transformed into a structured form so that the data analysis methods commonly used today can be applied to them. Those who have read Bill Inmon's magnum opus "The Corporate Information Factory" and who are familiar with his fundamental and – to a large extent – generally accepted architectural notions of business intelligence will certainly expect this book to contain concepts at an architectural level. However, chapter 6 – the title of which ("Architecture and Textual Analytics") leads us to expect a description of textual analysis within the context of the Corporate Information Factory – is disappointing and contains only a reference to Bill Inmon's Web site. The subject of textual analysis using visualization is also covered only sketchily. Most readers won't be satisfied with just the topic of self-organizing maps as a visualization tool – which is not sufficiently explained and motivated, to boot. Some people might buy the book in the hope of reading about details of or at least finding references to statistical methods for textual analysis. However, the book provides no information at all here.

Readability: One of the main strengths of the book is certainly its intuitiveness and readability. The book avoids superfluous details and thus makes it easier for readers to concentrate on a few central aspects for each topic. In each case, these are presented in a few easily understandable sentences and are often also reiterated using very simple diagrams. These images give the book a much less dense appearance. Because they can be grasped very quickly and also skipped without any loss of information, the reader can make rapid progress through the book. The 200 pages can be easily read in one or two afternoons.

Practical use: In many places, the content of the book is kept at a too simple and abstract level for it to be directly transferable to a practical situation. Algorithms or even suggestions for implementations are largely nonexistent. There are only a few very general recommendations for architectures and technologies. There are no suggestions for suitable data models. Software providers are mentioned sporadically by name, but without going into detail about features, citing specific implementation examples, or making comparisons. The mention in chapter 14 of specific SQL queries for individual steps in a textual analysis comes as quite a surprise. However, these examples describe only specific, isolated solutions. How they are integrated into the architecture as a whole remains unclear. And precisely here, I would have expected more of Bill Inmon. The index and glossary are good and usable. The book is well suited as a short introduction, which makes the lack of a list of further reading (books, periodicals, links to Web sites) all the more bitter. In fact, the book contains no such recommendations whatsoever – apart from links to Bill Inmon's Web site.