act on facts

| BUCHREZENSION

[FÜR SIE GELESEN]

mayato Buchempfehlungen im Umfeld Business Intelligence und Customer Relationship Management

Buchtitel:	**Tapping Into Unstructured Data**
Autor(en):	**William H. Inmon, Anthony Nesavich**
Details:	**Sprache Englisch / 240 Seiten / Prentice-Hall 2008 / Preis US $ 44,99**
Rezensent:	**Dr. Marcus Dill**

1 Empfehlung

Das Buch ist eine eingängige, kompakte und anregende Wochenendlektüre für den Business-Intelligence-(BI-)Kundigen und ein empfehlenswerter Einstieg für Leser, die sich über die Möglichkeiten der Integration von Textinhalten in vorhandene Datenanalyseumgebungen und –verfahren informieren wollen. Innerhalb dieses Themas bieten die Autoren viele einfache, inspirierende Gedanken. Die Lektüre schafft ein ideales Mindset für den Start eines Projektes zur Integration von Texten in Business Intelligence. Das Buch verbleibt allerdings auf einer abstrakten, allgemeinen Ebene und verzichtet fast vollständig auf Implementierungs- oder Architekturdetails. Es eignet sich somit kaum als Praxishandbuch. Wer angesichts des unscharf gehaltenen Buchtitels inhaltlich weiter gehende Erwartungen beim Kauf hat (z.B. statistische Verfahren, nichtstrukturierte Daten jenseits von Texten), wird enttäuscht, da es sich klar auf Texte und deren Vorverarbeitung für Analysen in einem klassischen BI-Umfeld fokussiert.

2 Inhaltszusammenfassung

Die Autoren bieten zunächst eine Übersicht über Eigenschaften von unstrukturierten Daten und vergleichen die Behandlung von strukturierten und unstrukturierten Daten. Der Löwenanteil des Buches ist der Fragestellung gewidmet, wie unstrukturierte Daten in die Welt der Analyse strukturierter Daten integriert werden können. Die Autoren zeigen die besonderen Herausforderungen und die grundsätzlichen Lösungsmöglichkeiten für diese Aufgabe auf. Sie verwenden den Begriff der „Textual ETL“ in Analogie zu den Extraktions-, Transformations- und Lade-(ETL)Prozessen in einem Data Warehouse. Breiten Raum nehmen Fragen rund um die Identifikation von Schlüsselwörtern in Texten und deren Mapping und Bereinigung ein. Nur in Ansätzen wird auf das weite Feld der Textanalyse und Textsuche eingegangen. Am Ende des Buches stehen 5 fiktive Beispielszenarien aus Unternehmen und Organisationen, anhand deren Anwendungen und Herangehensweisen illustriert werden.

© Copyright mayato® GmbH. All rights reserved.

3 Bewertungskriterien

Inhalt: Inmon und Nesavich bieten eine systematische Übersicht über die Verarbeitung von nicht oder halbstrukturierten Textdaten zum Zwecke weitergehender Analysen. Im Zentrum des Buches steht die Idee, Texte so zu verarbeiten, dass sich Informationen (Schlüsselwörter etc.) mit Hilfe von herkömmlichen, für strukturierte Daten entwickelten BI-Technologien gewinnen lassen. Auch wenn der Titel auf unstrukturierte Daten allgemein abhebt, so wird doch schon im Vorwort klar, dass das Buch sich auf Texte beschränkt und andere Arten von unstrukturierten Daten nicht berücksichtigt werden. Bilder, aber auch Sprach- und andere Tonaufzeichnungen, enthalten natürlich ebenfalls Informationen, die potentiell auch für Unternehmen interessant sein können. Mancher Käufer des Buches hätte sich vielleicht auch Hinweise auf die Möglichkeiten von automatischer Informationsextraktion aus solchen Dateien gewünscht. Der Inhalt des Buches ist geprägt von Systematik und Einfachheit. Viele Aussagen sind von fast axiomatischer Kürze, manche mögen vielleicht sogar trivial klingen. Dem Leser werden durch die Einfachheit viele kritische Aspekte klar zu Bewusstsein gebracht. Wer sich in naher Zukunft in der Praxis mit diesem Umfeld beschäftigen wird, kann sich über die Lektüre des Buches sehr gut einstimmen. Von Algorithmen oder gar Implementierungen abstrahieren die Autoren allerdings über weite Strecken. Selbst die Case Studies am Ende des Buches bieten wenig Konkretes und erlauben dem Leser vermutlich nicht wirklich den Transfer auf eigene Projekte. Das Verständnis von Textanalytik als Text-ETL (Extraktion, Transformation und Laden) wird der fachlichen und technischen Breite des Gebietes nicht gerecht, spiegelt aber durchaus die Grundhaltung wider, die vielfach von Business-Intelligence-Experten dem Thema entgegengebracht wird: Texte müssen in eine strukturierte Form transformiert werden, damit man die heute gebräuchlichen Methoden von Datenanalyse auf sie anwenden kann. Wer Bill Inmons Hauptwerk, die Corporate Information Factory, und seine grundlegenden und weitgehend allgemein anerkannten Architekturvorstellungen zu Business Intelligence kennt, der erwartet sicher in diesem Buch gerade auch Konzepte in architektonischer Hinsicht. Kapitel 6, dessen Titel („Architecture and Textual Analytics“) eine Darstellung von Textanalyse im Kontext der Corporate Information Factory erwarten lässt, enttäuscht aber und enthält nur einen Hinweis auf Bill Inmons Webseite. Auch Textanalyse durch Visualisierung bleibt in der Abdeckung des Themas zu punktuell. Nur Selbstorganisierende Karten als Visualisierungstool – und das noch dazu nicht ausreichend erläutert und motiviert – das wird die meisten Leser an dieser Stelle eher unbefriedigt lassen. Mancher wird beim Kauf des Buches ggf. auch Details oder zumindest Hinweise auf statistische Verfahren zur Textanalyse erhofft haben. In dieser Hinsicht bietet das Buch jedoch keinerlei Informationen.

Lesbarkeit: Eine der wesentlichen Stärken des Buches ist sicherlich die Eingängigkeit und einfache Lesbarkeit. Das Buch verzichtet auf überflüssige Details und erleichtert so bei jedem Thema die Konzentration des Lesers auf einige zentrale Aspekte. Diese sind jeweils in wenigen verständlichen Sätzen dargestellt und meist zusätzlich durch sehr einfache Schemadarstellungen wiederholt. Diese Abbildungen lockern das Buch optisch sehr stark auf. Da sie meist sehr schnell verstanden, oft auch ohne Informationsverlust übergangen werden können, kommt der Leser bei der Lektüre auch zügig voran. Die ca. 200 Seiten lassen sich gut an 1-2 Nachmittagen lesen.

Praxistauglichkeit: Die Inhalte des Buches sind an vielen Stellen zu einfach und abstrakt gehalten, als dass ein direkter Praxistransfer möglich wäre. Algorithmen oder gar Implementierungsvorschläge fehlen über weite Strecken. Es gibt nur wenige und sehr allgemein gehaltene Empfehlungen zu Architekturen und Technologien. Anregungen für geeignete Datenmodelle werden nicht gegeben. Softwareanbieter werden vereinzelt namentlich genannt, ohne dass auf detaillierte Features eingegangen, konkrete Implementierungsbeispiele aufgeführt oder Vergleiche angestellt würden. Überraschend kommen dann zwar doch in Kapitel 14 konkrete SQL-Abfragen für einzelne Schritte einer Textanalyse. Allerdings beschreiben diese Beispiele dann auch nur spezifische, isolierte Lösungen. Ihre Integration in eine Gesamtarchitektur bleibt unklar. Gerade hier hätte der Name Bill Inmon mehr erwarten lassen. Index und Glossar sind gut verwendbar. So gut das Buch als kurze Einstiegslektüre geeignet ist, so schmerzlich wird am Ende eine Literaturliste zur weiteren Vertiefung (Bücher, Zeitschriften, Weblinks) vermisst. Tatsächlich enthält das Buch keinerlei solche Empfehlungen (mit Ausnahme des Links auf Bill Inmons Website).

© Copyright **mayato**® GmbH. All rights reserved.