

Preface

There have been two environments that have grown up side by side—the structured environment and the unstructured environment. The structured environment is typified by transactions, databases, records, keys, and attributes. The unstructured environment is typified by email, spreadsheets, medical records, documents, and reports.

It is amazing that at the same time that these worlds have grown up side by side, they have grown separately. It is as if these worlds exist in alternate universes.

The world of analytics and business intelligence has grown up around structured information. With business intelligence, we have displays of information, summaries, pivots, and an entire world of analytical processing. With business intelligence, we can make sense of the numbers, facts, and figures that hide out in the systems that run our corporations.

For analyses of text—unstructured information—there is nowhere near the amount of sophistication that exists in the structured environment. In the unstructured world, a few search engines can find documents and that is about it.

Does that mean that there is no important or useful information in the unstructured environment? The answer is—of course not. There is a wealth of important and useful information in the unstructured environment, but it is not as easily recoverable as information in the structured environment. The information in the unstructured environment is much more difficult to get a handle on.

There are many reasons why textual data is more difficult to handle than structured, transaction-oriented data. The primary reason is the lack of repeatability of textual data and the lack of predictability about the contents of the data. Textual data is hard to handle because it is hard to find, and it is hard to find because it does not entail repetition to any great degree.

This book is about doing textual analytics and the technologies that can be used to do textual analytics.

Two major architectural and technological approaches to doing textual analytics are used. One approach is to look at and gather the textual data in the unstructured environment. When there, the textual data is analyzed and manipulated in the unstructured environment. The unstructured environment seems like a natural place to do textual analytics because, after all, the text resides in the unstructured environment.

The other architectural approach is to look at and gather the textual data in the unstructured environment and then bring the textual data to the structured environment to do the textual analytics there.

It might seem strange or even unnatural to take the approach of accessing and gathering textual data in the unstructured environment and then bringing the textual data to the structured environment for analytical processing; however, there are good reasons for doing exactly that. Some of those reasons follow:

- The analytical environment has already been created in the structured environment. If we bring unstructured data to that environment, we can leverage existing investments. We already have trained end users, trained support staff, and licenses in place. So, why not bring the unstructured text to the structured environment where analytical tools are already in place?
- Proprietary software. When we bring in technology to do analytical processing in the unstructured environment, that technology is proprietary. Do we actually want more proprietary software in our world? Isn't it a much more rational approach to use open software that has thousands of users and uses around the world, rather than bring in proprietary software that might or might not meet the long-term goals of the organization?
- By bringing unstructured text to the structured environment, we can create links between the unstructured data and our structured data, making possible analysis that otherwise would not have been possible. In doing so, we can build an integrated data warehouse that takes into account both structured and unstructured data.
- If we don't bring unstructured data to the structured environment, we are going to have to re-create the analytical infrastructure in the unstructured environment. Is that advisable? We already have an analytical infrastructure. Why not use it?

For these reasons, this book is about what is required to go to the unstructured environment, find and integrate the textual data there, and then bring the unstructured textual data to the structured environment and organize it in a meaningful manner. After the textual data is in the structured analytical environment, a new world of analyses opens up.

One of the recurring themes of this book is the need for integration of text before it is useful. In most environments and in most circumstances, text is nonhomogeneous. People might talk in English, but for all practical purposes, they speak in dialects. Before analytical processing can be done effectively, there must be a common tongue established. Stated differently, if all you do is gather text and throw it into a database, you end up with the Tower of Babel. The Tower of Babel led nowhere, certainly not up to God.

One of the requirements of textual analytical processing is accessing and analyzing text in a colloquial vocabulary and a common vocabulary. The textual analyst needs *both* abilities.

The classical approach to text and text processing is to use semantics and natural language processing. This book describes a different approach. Without fail, the approach taken in this book is that text—made up of words—is just another form of data. The approach that looks at words as just another unit of data frees the analyst from the trap of context. It is true that words taken out of context can have twisted meanings in some occasions. It is also true that freeing words from context opens up the door to entirely new and novel kinds of processing that simply are not possible when having to stop and consider the context of text at every turn.

There is a tradeoff. Paying attention to context when dealing with text entails a certain set of opportunities and precision. However, freeing text from context opens up entirely new and exciting vistas.

This book assumes that words are treated as just another unit of data and does not take context into consideration in 99.99 percent of the cases.

This book is for a wide audience. It is for students of computer science, general managers, database designers, data modelers, database administrators, researchers, and end users—in short, it is for anyone facing the challenge of taking a body of text and trying to make sense of it. In addition, this book answers the questions, “How do we bridge the gap between structured and unstructured systems?” and “How do we create an integrated data warehouse that incorporates both structured and unstructured data?”

The discipline of textual analytics is in its infancy; it is entirely predictable that more discussion and more advances will be made in the future about this subject. This book represents merely the first step in what is likely to be a massive field of endeavor in years to come.

We hope that you find the book full of useful information. We hope the book at least sets you down the right path to enjoying the fruits of textual analytics.

Bill Inmon, January 11, 2007

Tony Nesavich, January 11, 2007