Analytics of Textual Big Data

Text Exploration of the Big Untapped Data Source

A Whitepaper

Rick F. van der Lans
Independent Business Intelligence Analyst
R20/Consultancy

December 2013

Sponsored by

1 Introduction – Analyzing Textual Big Data

Big Data for Enriching Analytical Capabilities – *Big data* is revolutionizing the world of business intelligence and analytics. Gartner[1] predicts that big data will drive $232 billion in spending through 2016, Wikibon[2] claims that by 2017 big data revenue will have grown to $47.8 billion, and McKinsey Global Institute[3] indicates that big data has the potential to increase the value of the US health care industry by $300 billion and to increase the industry value of Europe's public sector administration by €250 billion.

The big data breakthrough comes from *innovative big data analytics*. For some companies the primary challenge comes from analyzing massive amounts of structured data, primarily numerical, such as credit card companies with millions of cardholders and billions of transactions looking for fraud patterns. Analyzing massive amounts of structured data may require new software strategies and technologies but is generally straightforward and readily achievable.

Not all big data is structured. Big data comes in all shapes and sizes. The greatest big data challenge is that a large portion of it is not structured, often in the form of unstructured text. Think of all the data used or created in a typical business – emails, documents, voice transcripts from customer calls, conferences with note taking, and more. Most of this data is unstructured text. Even in an industry dominated by numerical data, text abounds. For example, in commercial banking, financial statements and loan activity are well-structured data, but to understand the loan you have to read the file, which is full of correspondence, written assessments and notes from every phone call and meeting. To really understand the risk in a lending portfolio you need to read and understand every loan file.

In a medical environment, many structured data sources exist, such as test results over time and coded fields. However, some of the most valuable data is found within a clinician's textual notes: his impressions, what he learned from conversing with the patient, why he reached his diagnosis or ordered a test, what he concluded from various test results, and much more. In most large clinical settings these invaluable notes comprise very large data sets but, while they are increasingly digitized, they are rarely analyzed.

Analyzing Textual Data – Advanced analytical capabilities have always been available for analyzing non-textual data. Almost every organization knows how to turn their own structured data that has been collected over the years by business processes into valuable *business insights*. Countless reporting and analytical tools are available to assist them. Surely, these tools and algorithms may have to be adapted somewhat to be able to run fast on big data (for example, they may have to use in-memory techniques and dedicated hardware), but the algorithms stay the same and are well-known.

---

[1] Gartner, October 2012; see http://techcrunch.com/2012/10/17/big-data-to-drive-232-billion-in-it-spending-through-2016/

[2] Wikibon, *Big Data Vendor Revenue and Market Forecast 2012-21017*, August 26, 2013; See http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017

[3] McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, June 2011; see http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

But what about all the textual data that has been gathered in emails, document management systems, call center log files, instant messaging transcripts and voice transcripts from customer calls? And what about all the *external* textual data, such as blogs, tweets, Facebook messages and informational Websites? A wealth of information is hidden in the vast amounts of textual data being created every day. The challenge for every organization is to extract valuable business insights from this mountain of data that allows it to, for example, optimize its business processes, improve the level of customer care it offers, personalize products, and improve product development.

This paper will outline the benefits and challenges of analyzing textual big data. It will also discuss InterSystems iKnow™ technology, which offers an easier, less time-consuming way to unlock the information contained in textual data.

2 Textual Big Data: The Big Untapped Data Source

Business Reasons for Analyzing Textual Big Data – Almost every industry can benefit from analyzing textual data, especially those industries in which storing text is crucial for business operations, such as advertising, healthcare, legal, pharmaceuticals, publishing and real estate. For example, a hospital may be interested in analyzing the descriptions written by specialists and included in patient files to discover patterns with respect to allergic reactions to medications. An electronics company may want to analyze messages on Twitter to find out if their products are mentioned and whether the tweets are positive or not (a practice commonly referred to as sentiment analysis). Transcripts of call center log files can be analyzed to determine whether popular questions can be identified, or whether over the past couple of weeks specific products have been mentioned more often or in a different context than usual.

What Exactly is Analyzing Text? – We hardly need to "analyze"' text if we only want to know how many words appear in a document, or how often a word appears. This can be determined with a simple, purely mathematical algorithm. But what if we want to answer more complex questions, such as:

- How often do particular symptoms and medications appear together in patient files?
- Does a text carry a positive or negative sentiment and to which concepts applies this sentiment?
- How many texts deal with the bankruptcy of Bank X?
- For each month, how many texts dealt with brain surgery?
- Which concept appears most often in texts together with the concept credit card fraud?
- Which book is, with respect to concepts used on the content of the book, most equal to Jeff Shaara's *Gods and Generals*, and which one is most different?
- How do we identify the characteristics of customer calls that resulted in escalation?

These questions are much harder to handle. For example, how do we determine whether a text has a positive sentiment? How do you "measure" the difference between the contents of two books? This is the type of question for which text analysis is used.

Analyzing text can also be defined as deriving structured data from unstructured text. For example, when a text is analyzed based on whether it's positive or not, the result is a structured

data value: the value yes or no. The answers to the first and fourth questions above also lead to structured data. The advantage of deriving structured data is that this newly created structured data can be combined easily with other structured data sources and processed by well-known algorithms.

The Index and the Thesaurus – Historically, the first developments in trying to understand texts were based on *indexing*. Indexing texts means that terms are selected from a document that provide a sufficient indication of the document's subject-matter, ensuring that it can be retrieved using a specific query. Indexing, however, has its limitations. First, developing an index is time-consuming –which words should be indexed? Second, if the right terms are not indexed, some important and relevant texts may not be found, or incorrect texts may be found.

Quite some time ago, to overcome the problems with indexing, the concept of a *thesaurus* was introduced. With a thesaurus the relations between terms are defined. In a way, a thesaurus can be seen as an intelligent index. The result of deploying a thesaurus is that a more accurate set of texts is found. But building up and managing a thesaurus is also time-consuming. Plus, a thesaurus must be kept up to date as new words are introduced, new domains are introduced, and so on.

The Work Upfront – Most text analysis tools require work in advance, such as setting up a thesaurus. Such tools are only useful if there is enough time to do all this work. What if a new and urgent question arises and in the thesaurus this hasn't been catered for? Or what if new texts become available for analysis and questions have to be asked right away?

Also, with most text analysis tools the goal of the analysis exercise must be clear in advance. In other words, the tool is guided by the analyst. For example, search technology requires that one or more words are entered first. Another example is when patient files are analyzed to discover new insights with respect to the effect a particular medication has on patients with diabetes. As can be imagined, a different thesaurus may be needed when the goal is to look for historical patterns in side effects after surgery, even when the same patient files are analyzed. A thesaurus limits the analytical freedom and thus limits potential outcomes.

3 Exploring Textual Big Data Without the Hassle

Use of Text Analysis Today – Organizations can benefit from analyzing textual data. Unfortunately, most organizations have barely scratched the surface with respect to analyzing textual data. This is a missed opportunity.

One of the dominant reasons why organizations have not tapped into their big data is that most text analysis tools and technologies require time-consuming upfront work. Indexes, thesauri and ontologies have to be developed in advance before any real analytical work can start.

The Need for Text Exploration – Analysis has to be able to follow the speed of business. For text analysis, this means technology is needed that allows text to be analyzed without the need to do all that work in advance. This form of text analysis is called *text exploration*.

A hospital environment is a good example of where text exploration can be used. Imagine a patient is brought to the emergency room. If doctors must act quickly, they probably don't have time to read the full patient file. What they want is a summary that shows all the important aspects related to the patient. Is he diabetic? Does he usually have high blood pressure? What kind of medication is he taking? Has he been here before? This requires text analysis on the spot. The analysis should also be unguided, because the doctors might know nothing about this patient and therefore the analysis technology should not require the clinician to guide the tool, but the other way around.

Another example is analyzing tweets. Everyday new words (acronyms in many cases) and hash tags are invented. It would be impossible to constantly update a thesaurus on them. Furthermore, is there time to develop one?

Many situations exist in which there is no time for all this preparation work. Here, text exploration is needed to give the desirable business insights.

The Three Requirements for Text Exploration – To summarize, text exploration is a form of text analysis that meets the following three requirements:

- **No advance preparations:** There should be no need to develop thesauri or ontologies before the analysis work can be started. It should be possible to start text analysis right away without any preparations. Even if the text covers a new domain.

- **Unguided analysis:** Analysts should be able to invoke the text analysis technology without having to specify a goal in advance. The text analysis technology must be able to analyze the text in an unguided style.

- **Self-service:** Analysts are able to invoke the text analysis without the help of IT experts, although connecting the tool to particular data sources may require some assistance.

4 InterSystems' iKnow Technology for Analyzing Textual Big Data

The Classic Approach of Text Analysis – Tools for analyzing text usually try to identify important concepts in sentences. For example, in the sentence *The enterprise search market is being reshaped by new consumer experiences*, the key concepts are *enterprise search market* and *new consumer experiences*. Most text analysis tools try to locate these concepts by looking at individual words, which results in the words *consumer, enterprise, experience, market and search*. They are considered key concepts in this text.

Some tools search for two-word phrases and even three-word phrases. The result of this approach, however, can be that words are "connected" that should not be connected. Take the following sentence as an example: *Michael Phelps breaks a world record*. If two-word phrases are identified, the result contains the concepts *Michael Phelps* and *Phelps breaks*. Now, the first one is probably a useful one, but the second isn't. And if we would search for all two-word phrases in the first sentence, we get *enterprise search* and *search market*, but not *enterprise search market.* This more classic approach doesn't guarantee that the words that are linked together form the right concept.

In addition, to make sense of the sentences, developers have to build up thesauri and ontologies. This can represent quite an effort and requires domain knowledge. For every domain a new thesaurus and ontology must be created and maintained. In most situations, this process will never end, because the use of words changes over time. New terms are introduced and the meaning of words can change. As an example take tweets – every day new important hash tags are being introduced. New terms are also introduced in the BI domain. Who had heard of the term big data a few years ago?

InterSystems' Approach to Text Analysis – The approach taken by InterSystems to analyze text is different from many other approaches. InterSystems has introduced a technology – called iKnow – that breaks texts into sentences, and then breaks sentences into concepts and relations. Decomposing sentences is done by first identifying the relations in a sentence. Verbs can represent relations between concepts in a sentence, but other language constructs can signify relations as well.

By identifying the relations within a sentence, iKnow has a better chance of discovering the desired concepts. For example, in the sentence *The programmer found bugs*, iKnow considers the verb *found* to be a relation separating the concepts *programmer* and *bugs*. In iKnow this is called a *concept-relation-concept sequence* (CRC). Note that iKnow automatically discards irrelevant stop words from sentences, such as *the* and *an*.

As indicated, other language constructs can indicate a relation. For example, in the sentence snippet *Mammals, such as elephants and lions …* a relation exists between *mammals* and *elephants* and between *mammals* and *lions*. Another example is the sentence *I like the car in the showroom*. Here, the word *in* represents a relation between the concepts *car* and *showroom*. iKnow has been designed to recognize many different language constructs that can identify relations.

If the concepts and relations consist of multiple words, iKnow still recognizes them. For example, in the sentence *The enterprise search market is being reshaped by new consumer experiences,* iKnow discovers that the verb clause *is being reshaped by* represents the relation between the two concepts *enterprise search market* and *new consumer experiences*.

This fast and domain-independent entity identification process effectively decomposes sentences into graph structures where concepts are linked to one another through relationships. These graph structures and the contextual metadata and metrics iKnow collects along the way can then be used for advanced analysis within a text or across a corpus of texts.

iKnow is not restricted to analyze simple sentences consisting of CCs and CRCs. It can handle more complex sentence structures consisting of multiple CRCs. These are called CRC sequences.

Note: InterSystems iKnow technology supports several languages, including Dutch, English, French, German, Portuguese and Spanish. Japanese and Russian are in development.

How InterSystems iKnow Technology Supports the Three Requirements for Text Exploration – iKnow supports all three key requirements for text exploration described in Section 3:

- **No advance preparations:** iKnow doesn't require the development of thesauri and ontologies. It can analyze text coming from a domain or industry it has never analyzed before, and is still able to discover the important concepts.

- **Unguided analysis:** iKnow does not need a goal. It does not, like search technology for example, need a search term before it can analyze the text. iKnow can analyze text in an unguided or *bottom-up* fashion. The result can be studied by the analysts and those results can then trigger them to search in a certain direction.

- **Self-service:** Analysts can use InterSystems DeepSee™ to invoke all the text analytical features of iKnow. DeepSee can be categorized as a self-service analytical technology that allows users to develop their own reports and do their own analysis without the help of IT experts.

Using iKnow With Big Data – InterSystems iKnow technology is embedded with InterSystems Caché®, a high-performance database server. Caché's unique multidimensional data engine makes it ideal for storing, managing and querying all forms of data, including textual data. Its performance and scalability have been proven in many big data environments. Any Caché-based application can invoke iKnow and can therefore analyze both text and structured data.

5 Summary

Everyone agrees, big data can enrich the analytical capabilities of organizations. For many organizations today it means crunching huge amounts of highly structured and mostly numerical data. In other words, the focus has been on analyzing non-textual and highly structured data.

However, a wealth of information is hidden in vast amounts of textual data sources, such as emails, document management systems, call center log files, instant messaging transcripts and voice transcripts from customer calls. Not to mention external textual data, such as blogs, tweets, Facebook messages and informational Websites.

For most organizations these textual data sources are still an untapped source of information. The challenge for many organizations will be to extract valuable business insights from this mountain of textual data in order to, for example, optimize business processes, improve the level of customer care, personalize products and improve product development.

Text exploration is a form of text analysis that allows organizations to analyze textual data at the speed of business. No or minimal upfront work is required. Text can be analyzed when needed by the business.

InterSystems iKnow is a breakthrough technology aimed at text exploration. It allows organizations to analyze their big textual data for business insights as needed.

About the Author Rick F. van der Lans

Rick F. van der Lans is an independent analyst, consultant, author and lecturer specializing in data warehousing, business intelligence, data virtualization and database technology. He works for R20/Consultancy (www.r20.nl), a consultancy company he founded in 1987.

Rick is chairman of the annual European Business Intelligence and Enterprise Data Conference (organized in London). He writes for the eminent B-eye-Network[4] and other Websites. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles[5] all published at BeyeNetwork.com.

He has written several books on SQL. Published in 1987, his popular *Introduction to SQL*[6] was the first English book on the market devoted entirely to SQL. After more than twenty years, this book is still being sold, and has been translated in several languages, including Chinese, German and Italian. His latest book[7] *Data Virtualization for Business Intelligence Systems* was published in 2012.

For more information please visit www.r20.nl, or email to rick@r20.nl. You can also get in touch with him via LinkedIn and via Twitter @Rick_vanderlans.

About InterSystems Corporation

Founded in 1978, InterSystems Corporation is a US$446,000,000 privately held software company with offices in 25 countries and corporate headquarters in Cambridge, Massachusetts. They provide the premier platform for connected healthcare, and their innovative products are widely used in other industries that demand the highest software performance and reliability. Clients include TD Ameritrade, European Space Agency, U.S. Department of Veteran Affairs, Johns Hopkins Hospital, Belgium Police, Mediterranean Shipping Company, and thousands of other successful organizations.

Leading application providers also leverage the high performance and reliability of InterSystems' advanced technology in their own products. These organizations include Epic Systems, Fiserv, GE Healthcare, and hundreds of others.

---

[4] See http://www.b-eye-network.com/channels/5087/articles/
[5] See http://www.b-eye-network.com/channels/5087/view/12495
[6] R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.
[7] R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.