Knowledge Management World article – Text analytics and beyond

Proactive measures for compliance

It is growing increasingly common to use text analytics to proactively manage data against breaches and other risks. “One of the issues that companies face is that their information is scattered in so many places, in hundreds of different formats,” says Jake Frazier, senior managing director and global practice lead for information governance and compliance services at FTI Technology, which provides e-discovery and information governance software and services. It is one of the business practices within FTI Consulting, a global business advisory firm focusing on matters involving corporate finance, economics, forensics, litigation and strategic communication for crisis management.

FTI Technology’s software uses text analytics to more quickly and easily find important information within the sea of enterprise big data. That enables organizations to better protect sensitive data or more easily collect it for regulatory or legal matters. One example is a recent engagement with a healthcare industry company. “We created several thousand rules under a master taxonomy for classifying documents,” Frazier explains. When the software finds a subset of documents that might fit the category, human reviewers validate or correct the classification, so the system can begin inferring other rules. “Over time, the software can scour petabytes of data and know with fairly high certainty that all the purchase orders for medical equipment have been located, for example,” he adds, “and if the company is

audited, the documents can be easily found."

The two products employed for that purpose are Ringtail, which is used for enterprise discovery and legal e-discovery, and Radiance, which performs analytics and visualization on extremely large data sets. With those products, companies can identify key information such as social security numbers or credit card data and create rules to better archive or remediate data. "The ultimate goal is to put resources where they are most needed to ensure enterprisewide governance and compliance," says Frazier. "Once a company identifies the subset of systems containing highly sensitive information, such as patient data and other personally identifiable information, that data can receive priority protection."

Innovative software for disorganized data

Quite a few companies have been working on software products that operate on data that has no schema and is heterogeneous. Cognitive technology from Coseer is being used to build a database for a company that deals with healthcare products. The company managed 10 million SKUs representing healthcare products. "There was no standardized product database," says Praful Krishna, CEO of Coseer, "and about 35 million PDF documents were in the company's files or posted on the Web. No table of attributes had been created, so searching for the right product or comparing different products was extremely difficult."

Coseer's software ingested that diverse collection of unstructured content, which included product brochures, catalogs, surgical protocols and white papers. It detected patterns in the content that enabled it to create metatags and organize the information. "After going through this process, we are able to determine when one SKU is very similar to another one that may, for example, be much cheaper," Krishna

says. “Analyzing this much content with absolutely no structure is impossible to do without cognitive processing of text.”

The resulting database allows executives to find all the potential candidates, focus on certain attributes and help the care providers make better choices regarding products that meet their needs. “This project uses a model developed specifically for the domain,” Krishna explains, “and it took several months to train the system. But at the end, the results are more accurate than those for a sample categorized by humans, and the fill rate for identifying attributes was 12 percent higher.”

Precision medicine

Healthcare provides fertile territory for the use of analyzing unstructured content. Its value is multiplied when information from a large number of diverse sources is integrated with structured data in many different contexts. One current effort involves a partnership between Montefiore Health System and Franz. The goal is to move toward precision (also called “personalized”) medicine, which takes into account an individual’s characteristics, including genetic makeup, medical history and demographics.

A semantic data lake for healthcare stores data originating from numerous sources, both structured and unstructured. AllegroGraph, a graph database developed by Franz, stores medical vocabularies, taxonomies and ontologies to provide a semantic overlay to the data. “We are mining unstructured text and electronic medical records and feeding the results back into a graph that shows relationships among data elements,” says Jans Aasman, CEO of Franz. “The system includes many unstructured notes as well as structured data, and with the semantic data lake, it is possible to do very sophisticated analyses on a real-time basis.”

Having ready access to a large volume of medical data also presents the opportunity to achieve another goal: to predict medical incidents and either take action pre-emptively or be prepared to respond rapidly. In the pilot program at Montefiore, the hospital can use predictive analytics to flag patients who are hospitalized in its various locations who are at risk of death or for intubation within the next 48 hours and find almost all the high-risk patients with only a one percent error rate.