ASIS&T ANNUAL MEETING PLENARY SPEAKERS

Text and Data Mining Meets the Pharmaceutical Industry Markus Bundschus Speaks

by Steve Hardin

2016 Annual Meeting Coverage

EDITOR'S SUMMARY

Text and data mining have proven to greatly impact the world of biomedical research, especially for Roche Diagnostics in Penzberg, Germany. Taking information from such sources as patient literature, genomic cancer samples and PubMed articles, researchers at Roche Diagnostics are able to structure the data in a way that lends itself to creating personalized healthcare. Text mining used to build structured databases tends to yield the most relevant information for biomedical research, so Roche uses unstructured data to build a knowledge base automatically. This knowledge base, the disease marker association database, offers search capabilities for full text, abstracts or curated data. The database is made up of 50-million scientific abstracts and leans on rule-based engines as well as machine learning engines. By combining information from patient care, diagnoses and treatment, the healthcare industry can see a shift to digitization and more efficient care.

KEYWORDS

data mining
biomedical information
pharmacology
knowledge bases
information science

Markus Bundschus has served as head of business and information services at Roche Diagnostics Bavarian Site (Penzberg) since 2010. He can be reached at markus.bundschus<at>roche.com.

Steve Hardin is public services librarian, Cunningham Memorial Library, Indiana State University Steve.Hardin<at>indstate.edu.

When you visit the doctor, you may not be thinking about text and data mining. But Markus Bundschus says they're having a profound impact on biomedical research. Bundschus serves as head of scientific and business information services at Roche Diagnostics' Bavarian Site in Penzberg, Germany. He told the second plenary session at the ASIS&T Annual Meeting that Roche contributes to cutting-edge science and personalized healthcare in numerous ways, such as taking the DNA sequence from a patient and making treatment decisions accordingly. It can take millions of working hours, thousands of experiments and hundreds of scientists to create one new drug.

Bundschus said there are three main sources of external information for the science his division at Roche does: patent literature, PubMed articles and raw data such as genomic cancer samples. It's critical to provide context through data integration, ontologies and data mining. He related a story of researchers whose efforts failed because they missed some crucial literature.

Biomedical research and development depends on text and data mining. How, Bundschus asked, can researchers "burst the dam" holding back information? First, they utilize data integration and aggregation. They look for high quality information. They also try to democratize data and analysis. Data analysts speak a different language than biomedical researchers. They need to learn how to communicate. It's also important to learn from the past, to understand the impact of older research and data.

How can we bring text mining successfully to the end user? First, Bundschus quoted Google's Ten Things We Know to be True: "Focus on the user and all else will follow." Apply this advice to the domain of life sciences as well as to the day-to-day work of the industry researcher. What are the favorite tools of both domains? They want to make the look and feel of the text mining tools resemble the look and feel of the favorite tools.

HARDIN, continued

He said there are two main modi operandi for text mining: there's text mining as a way to complement the traditional literature search, and there's also text mining to build structured knowledge bases. You get more information from text mining in a structured database [1].

One example involves the pharmacological parameters of antibody drug conjugates (ADCs). ADCs are a new class of highly potent biological drugs built by attaching a small molecule anticancer drug or therapeutic agent to an antibody. Can we learn from existing research parameters? You can search on the full text or the abstract or curated data. You can get a lot of additional knowledge from the full text. But not always; it depends on the situation.

Text mining can be used to help build structured knowledge bases automatically from unstructured data. Roche researchers built the disease marker association database with 50-million scientific abstracts. They went from an unstructured world to text mining with relation classification for relevant articles/information. The motivation is to avoid information overload, reduce the time component to manageable levels, enhance flexibility and provide links to the literature.

Roche scientists use a flexible platform to create structured knowledge bases. There are rule-based engines with machine learning engines feeding into the normalization module. There is also an optional curation module. Then the indexing and representation layer is added. A graphical user interface (GUI) makes it easier to use.

Bundschus showed the user interface for the disease marker association database. He searched for bladder cancer; the program suggested various topics it considers related. The user can export the results to an Excel file which he or she can then use for other data collections.

Bundschus and his Roche colleagues want to link the database world and the literature world. Utopia Documents is a semantic, scientific PDF reader from the University of Manchester [2]. The unstructured literature world represents the most complete human knowledge base. It's growing exponentially. It is used to create a structured database world, representing human knowledge in a machine-readable format, after being created by humans who analyze the unstructured world. There are plugins available to show additional data. Once scientific PDF articles are loaded, users can make queries of those articles. They are usually able to go directly to the PDF instead of having to go to a web page and download it separately. They can create a chart showing the relationships among the various articles. Users can also go into the gene sequences and further determine what's going on. There's also a figure browser to search illustrations.

Text and data mining can be used for outcome prediction in clinical trials. The results are quite encouraging, Bundschus said. They can assess the probabilities of a launch by uncovering publication data patterns from thousands of scientists. For example, in 2011, Zelboraf, a medicine used to treat melanoma, was launched. Prior to that launch, the biggest component of research for this product involved genetics. But drug therapy and drug effects are also important components. The idea is to gain new insights into the drug introduction process.

Article counts can be useful too. Consider the number of papers with co-occurring target and indication per year. The article count is significantly higher for the drugs that make it to market. Successful topics have on average a stronger author commitment as well as a lower average density of gene names in their abstracts before approval.

Look at the evolving digital health ecosystem and the role of scientific literature. Is there a big misunderstanding about text mining? On one hand,

text mining enables researchers to read less. On the other hand, text mining enables researchers to read more relevant things. The healthcare system is always learning, and digitization holds a lot of promise. Researchers always say, "I should read more; but when?" Also, they ask, "Can I trust the data?"

Medical care includes continuous monitoring of the patient, diagnoses and treatment. If we can achieve the integration of all these streams of information, patients will be helped, and the results will work their way into the scientific literature. To sum up, the healthcare system will undergo a digitization shift, he said.

Holistic information science approaches will become a differentiator. He wants to not have to go to various repositories and aggregate information. It's better to get it at one site.

"The scientific literature ecosystem," Bundschus concluded, "has the potential to connect all the dots and act as glue that holds it all together." ■

Resources Mentioned in the Article

[1] Bundschus, M., Bauer-Mehren, A., Tresp, V., Furlong, L., & Kriegel, H-P. (2010). Digging for knowledge with information extraction: A case study on human gene-disease associations. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1845-1848.

[2] Utopia: http://utopiadocs.com/