Advanced Search

C. DAVID SEUSS, CEO, **Northern Light,** discusses "meaning extraction" to identify elements of information and concepts contained within both individual documents and document repositories – a data-sifting process that could impact the data-rich field of pharmaceuticals

Few endeavors rely more heavily on comprehending large numbers of documents than drug discovery. So it stands to reason that any facility that speeds and simplifies the task of mining published or internally generated research — thus accelerating "time to insight" — would be embraced by the pharmaceutical industry.

Enter meaning extraction — an emerging technology that identifies elements of information and concepts contained within documents and document repositories, and surfaces combinations of these informative elements and concepts that imply meaning in the context of the business, professional, or technical purpose of the search process. Today, meaning extraction is beginning to be applied by pharmaceutical companies to searching electronic document repositories and various online resources to dramatically improve and accelerate a searcher's ability to gain insight into a topic and answer specific research questions.

Consider the value of meaning extraction applied to PubMed, a National Institutes of Health (NIH) research database of journal abstracts that is freely available to researchers in life sciences. PubMed indexes the abstracts of over 5,000 journals and 18 million scientific articles. Life sciences researchers can easily access PubMed and execute searches using standard keyword search techniques (commonly known to any individual that works with a Web search engine such as Google). PubMed returns lists of documents to the user that match the search criteria, relevance ranked, in the traditional method of search engines. Over 100,000 searches per month are carried out on PubMed.

Separately, the NIH maintains a controlled vocabulary of life sciences terms under its Medical Subject Headings (MeSH) program. MeSH consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. There are thousands of descriptors in MeSH. Article citations in PubMed are tagged using MeSH, and knowledgeable users that understand the structure and term lists of MeSH can use terms from MeSH to search PubMed at multiple levels of aggregation, since MeSH is a hierarchical system with inheritance.

As useful as this is, PubMed suffers from a severe limitation in that there's no built-in analytic capability. If you do a search using a text string as a query, you will get a traditional search engine result in the form of a list of documents that contain the search terms. Like most search engines, these lists of search results are dauntingly long. Worse still, the summary information contained in the search results often provides little help in deciding if the document would actually be helpful in the particular case to the researcher. The researcher is left to wade through the search results, examining one document after another to find meaningful insights.

EXTRACTING MEANING FROM "AIR POLLUTION"

Assume a researcher is interested in finding out what diseases and drugs are related to air pollution. Using PubMed's search engine on the query "air pollution," there are almost 30,000 hits in total, and there is no way to actually answer the research question without examining each document in detail. In most cases, there is little in the document summary provided by the search engine that would help the research with his or her question.

To add meaning extraction capability to a search in PubMed, the following steps are required:

- Create full-text, metadata, and phrase indexes of the PubMed documents
- Convert MeSH terms to forms suitable for entity extraction/text analytics
- Extract entities from the PubMed document text using the converted MeSH vocabularies
- Create word, phrase and entity proximity indexes of the PubMed documents
- Specify algorithms that can be used by the text analytics technology to discover knowledge
- Embody the indexes, extracted entities, proximity intelligence and analytical algorithms in a user-friendly application that can be used by researchers

With these foundations in place, it is possible to specify algorithms that search automatically across the entire PubMed repository for meaning. For example, an algorithm might be:

- Identify all two and three element combinations of Diseases, Therapies, Drugs, Gene, Proteins and Enzymes that are within 40 words of each other in documents containing a text string specified by the researcher

Using a system designed to accomplish this task, such as Northern Light's MI Analyst, the researcher enters the search term "air pollution," and the search engine returns a list that directly answers the question on diseases related to air pollution. By clicking on the "Diseases" facet on

the search results list, and then opening subcategories — such as "Nervous Systems Diseases," "Respiratory Tract Diseases" or "Pathological Conditions, Signs and Symptoms" — the user instantly identifies diseases mentioned in documents with "air pollution."

While diseases like asthma, cough and rhinitis can hardly be surprising as outcomes related to air pollution, suppose that the researcher did not already know that Williams Syndrome, leukemia or deafness were implicated as a consequence of air pollution. In that case, the results list would be a moment of revelation and discovery. This is an example of an immediate form of meaning extraction. By telling the researcher what is in the documents on the results lists, the search technology contributes to the user's understanding of the topic. The search engine has evolved from just providing document lists into an analytical tool that can assist in understanding. This by itself is a great step forward.

REDUCING "TIME TO INSIGHT"

Meaning extraction enables a researcher to drill even deeper and gain greater insight. Suppose that the researcher wonders what drugs are being discussed as therapy for the diseases he or she has identified in the PubMed repository as being related to air pollution. A researcher could click on a disease of interest, like Williams Syndrome, then click on the Drugs facet, and then see that the research papers that contain "air pollution" and Williams Syndrome mention these drugs:

1. Insulin (27)
2. Bayer ASA (8)
3. Accutane (1)
4. Decadron (1)
5. Folvite (1)
6. Mucomyst (1)
7. Neoral (1)

Within seconds of starting the search process, the researcher knows two new ideas that he or she did not know before: **1**that Williams Syndrome may be related to air pollution; and **2**that insulin and aspirin may be common treatments for Williams Syndrome. The researcher might then be tempted to consider if these drugs would help with the other diseases related to air pollution — and, potentially, the process of "meaning extraction" evolves into "insight creation." This process using the right search tools happens very quickly, and it is this speed that suggests the core benefit of such tools: *time to insight*.

> "IN SOME CASES, **THE RESEARCHER WILL NOT PREVIOUSLY HAVE CONSIDERED THE RELATIONSHIP THAT IS IDENTIFIED;** THAT IS WHEN BREAKTHROUGHS ARE ENABLED."

Without such tools, researchers in the real world typically will examine only a relatively small sample of documents from the nearly 30,000 on the initial search results list, and hope for the best. An automated, comprehensive analysis of the documents on the search result is much more powerful than the hit-or-miss alternative that is the predominant means of literature search today.

FINDING MEANING

Finally, the meaning extraction application can be directed to analyze the documents returned on a search results list and identify relationships that imply meaning, surfacing those to the researcher to consider. In the search on "air pollution," here are some of the relationships that MI Analyst finds in our sample of the PubMed research database:

1. Osteoporosis is related to Skin Disease (95)
2. Chelation Therapy is related to Williams Syndrome (85)
3. Atelectasis is related to Bronchiectasis(85)
4. Bronchiectasis is related to Hemoptysis (85)
5. Ciliary Motility Disorder is related to Dyskinesisas (84)

It's more accurate to call these relationships "research scenarios" because, at this level of automated text analysis, one cannot really tell if the relationships are significant or spurious. All we can say is that the relationships are present in the document repository, and we can measure how many times each one is there.

The meaning-extraction application identifies these scenarios and presents them to the researcher as possibly worthy of follow-up. For example, with a few more mouse clicks, the researcher can determine whether there are common elements contributing to the relationship in the form of overlapping genes or proteins or other elements. The identification of the relationships is done automatically for the researcher, without any specific direction other than the initial restriction — in this case, to documents with the text string "air pollution." After that, the meaning-extraction application analyzes all the text in all the documents of interest and finds the elements and the relationships between them. Amazingly, 1.9 trillion potential relationships between drugs, diseases, genes, proteins and so on are searched in milliseconds.

In many cases, the researcher will already know about a relationship identified by the system; in these cases, meaning extraction is helping the researcher narrow down a document list to those that contain the scenarios he or she finds most interesting.

Yet, in some cases, the researcher will not previously have considered the relationship that is identified; that is when breakthroughs are enabled. Meaning extraction promises to help accelerate the pace of drug discovery by speeding and empowering the critical, but often tedious, time-consuming and, worst of all, hit-or-miss process of literature review. FP

C. DAVID SEUSS is CEO of Northern Light, which provides strategic research portals to pharmaceutical companies and other research-driven businesses. Mr. Seuss joined Northern Light in 1996, and later led an employee group in purchasing Northern Light from its corporate parent in 2003. Prior to Northern Light, Mr. Seuss was founder and CEO of Spinnaker Software Corp., which he led from inception to a public company with $65 million in revenue and 280 employees. Before starting Spinnaker, Mr. Seuss was a consultant and manager for the Boston Consulting Group.