Northern Light

Northern Light Perspectives

Revisiting Controlled Vocabularies

Controlled vocabularies are a valuable idea, but they were used the wrong way during the early days of document searching. Going back to the Dialog system, there was a controlled vocabulary that one would use to structure a search, and only those documents that had those keywords assigned to them would be returned as a search result. This process gave rise to the still-enduring term “keyword search.” The problem with that was the controlled vocabulary was a restriction on the search process, and it caused many frustrations.

Then free text searching came along with systems like Alta Vista, which offered the first full-text search of entire web pages. This was document search nirvana – at least momentarily.

Full-text searching does indeed serve many needs. The user can enter some words without being trained to use a controlled vocabulary, and the documents do not have to be tagged, which was always a heavy burden on the entire process. Suddenly the screen is filled with wondrous results that completely match the search criteria. One is therefore tempted to dispense with controlled vocabularies.

Not so fast. It turns out controlled vocabularies have an important place in an entirely new part of the search process: text analytics. This involves taking lists of words that are meaningful to the business purpose of the search and mining the documents for the occurrence of those terms and their co-occurrence with each other. Employing text analytics makes the search process a more productive event for the free text searcher in the relevant domain.

For example, Northern Light took the MeSH (Medical Subject Headings) controlled vocabulary and applied it to the NIH’s Medline database of 25 million peer-reviewed journal article abstracts and looked for the co-occurrences of terms relating to diseases, drugs, cells, proteins, or designs. In a text analytics demonstration to a major pharmaceutical company, we found that in the Medline database certain diseases were discussed in connection with others. The head of research for a major disease state who attended the demonstration asked, “Are you telling me that disease A is related to disease B?” (This was unexpected because the two diseases frequently affect different body systems and age groups.) Our response was, “We are not saying that we have scientific proof of a relationship, but in the literature of the best research on the topic, disease A and disease B are related.” The research scientist exclaimed, “If that is true, then this is an ‘Aha!’ moment!”

This example illustrates a very powerful use of controlled vocabularies – facilitating a post-search analytical process (data mining and data analysis) to address the business problem of the search. Effectively the system asks the user, “Did you know that these two terms are related to each other in the documents on your results list?”

Here’s another example: In a recent search about a company in the transportation industry, the embedded text analytics application gave results that said the most common phrases in news reports on the company were “earnings disappointment”, “fuel prices”, “energy costs”, and “gasoline prices”. Without reading any of the 343 news reports that the search retrieved, it was possible to glean significant intelligence about that company’s current set of business issues. If you were a sales executive planning a sales call on the CIO of the transportation company, as our client was, you would know to pitch your new software systems as ways to save fuel for the fleet.

The point is this: Having the controlled vocabulary and doing a little bit of simple statistical analysis can lead a researcher in a wholly new direction.

And it really is straightforward. There is academic research showing that the co-occurrence of two terms is the single best way of knowing that they are related to each other. There’s no need to worry about an ontology and the nature of the relationship between the terms. In the life sciences, MeSH is an excellent starting point; some industries, like information technology, require creation of an original comprehensive controlled vocabulary. Users need to be very thoughtful about authorities and how much effort has been put into developing them. Once the vocabulary exists, looking at the co-occurrence of terms and thinking about the business meaning when they co-occur is important.

In most searches, too much is returned on results lists. Search engines must evolve to have an in-depth understanding of the search material. Applied creatively, controlled vocabularies will be useful for a long time. They went through a hiatus where we got away from systems like Dialog and fell in love with full-text search, but now text analytics changes the picture. Today it is possible to use a controlled vocabulary in a whole new way to infer business meaning from search results.

# # #

(10/15)