News and Semantic Technologies

Edmund W. Schuster

Laboratory for Manufacturing and Productivity

Massachusetts Institute of Technology

The world is moving at a fast pace, and globalization is increasing. Financial markets are interlocked and the role of information becomes more important with each passing day.

Newspapers represent an under-utilized source of information for real-time decision-making in business. Commodity, stock, and bond prices, Federal Reserve developments, and events in Asia and Europe are just a few of the news items that have potential to affect any firm. Putting all of this information together to form a big picture is difficult with current technology.

Computing power exists to process links between news articles for the purpose of enhanced search. However, few computing languages can achieve these links in practice.

RSS feeds have improved the ability to aggregate news for a single topic. Even with this advance, there remain significant issues in analyzing the meaning of text using computing machines. The future is a quick understanding of the exact meaning for a news article in relation to a specific business. Wall Street is waiting for such a break-through.

Since the process of information diffusion is an important factor in the efficiency of financial markets, creating improved semantic links between news articles or any type of knowledge, past and present, is an essential goal. No one wants to be the last to learn of an important news development. Innovative ways of achieving semantic links speeds the diffusion to those in business who need to know.

Since 2003, we have been exploring the relationship between linguistics, computing, and the standards needed to make the next advance in semantic technologies. The result is the M Language, a means of creating semantic links for news articles, among other things, that are machine understandable. You can see an example at mlanguage.mit.edu

With M, it becomes much easier to create meaningful associations between past and present news articles and to reduce the "noise" of keyword and phrase searches. This improvement in linkage and efficiency also has implications for online advertising through targeting messages. Semantic ambiguity becomes less of an issue with the M Language.

Though this approach also applies to information posted on Blogs, news articles represent a source with much higher standards for facts and reporting. In general, financial news publishers have not differentiated this advantage to the general Internet community. For most people, online articles have the same relevance no matter the source.

To achieve semantic integration, Hristo S. Paskov an undergraduate computer science student at MIT has created a new method to place text into a standard form for analysis and linkage. Working in conjunction with Dave Brock and others at the MIT Data Center Program, Paskov's method includes the use of a sentence parser to convert free-form text into parts of speech like subject, verb, object and noun, verb.

Once this is accomplished, a custom program transforms the parsed sentence into XML with M words as tags for parts of speech. At this stage, words from the sentence still appear as natural language. Using a disambiguator, the natural language words held in XML are transformed into words from the M Dictionary.

In the end, each sentence from an article appears as M-XML, creating a standard format for search and analysis. Because Paskov's approach uses words from the M Dictionary, there are additional semantic relationships available for query.

Having a way to express free-form text as M-XML opens a number of new possibilities to understand context. This is an important aspect of broader aspirations involving knowledge management and discovery. By driving to a way of expressing sentences in structured form, combined with the power of parallel computing, the prospect of determining the context and meaning for large amounts of text is a near-term achievable goal. Already, there are plans at the MIT Data Center Program to place all 2 million Wikipedia articles in M-XML form. Using the M Language method mentioned above, the task would take about four months of computing time. However, this would accumulate enormous amounts of knowledge in a standard form that enhances advanced types of search.

While this is an impressive idea, there are several downsides. First, the current sentence parsers and disambiguators are not 100% accurate. Potentially these errors are cumulative, introducing false sentence data into M-XML.

There is also the issue of the changing content of Wikipedia. Every day, there are thousands of additions and so the knowledge is dynamic.

A final issue involves the accuracy of Wikipedia and the corresponding inferences from doing a transformation into M-XML. There is no question that the wiki process is very good at accumulating disparate thoughts, information, and data on a single topic. However, the objective standards of journalism sometimes are not consistently applied. This would introduce discrepancies into the knowledge based expressed in M-XML.

A much better approach would be to apply Paskov's method to a database of articles from the Wall Street Journal, New York Times, or other news publisher. With this approach, vast amounts of knowledge could appear in a standard form. Connecting this information together would be tremendous achievement in knowledge management.

The future is semantic connections for news articles from high quality sources that users (and computers) can query in real time to gain insight for decision-making in business. Unlocking information that is currently contained in publisher databases represents a new potential source of revenue and a way to improve the diffusion process for news. Information always gains value as it moves beyond the four walls of a company. Refining the efficiency of this movement is a dream of economists for the past fifty years. It represents the next stage of the information economy. Many news publishers can be at the forefront of this development through using the M Language.