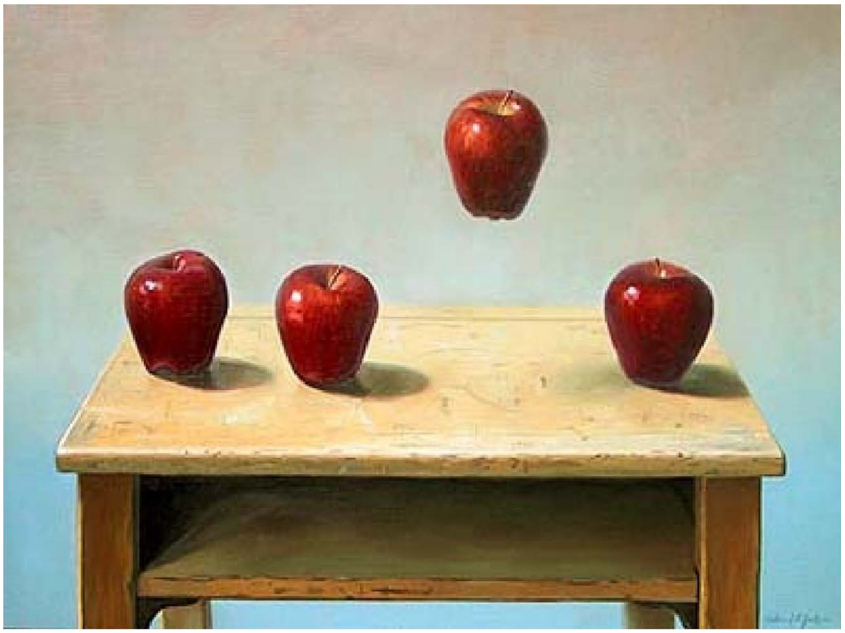
In-Depth Understanding: Teaching Search Engines to Interpret Meaning

By C. DAVID SEUSS
Northern Light Group, LLC, Cambridge, MA 02141 USA



Artist: Robert Jackson, BSEE University of Delaware, 1986

If a person from 1994 jumped forward into 2011, that person would be wowed by many of our information technology advances. Smart mobile phones, the ubiquitous Web, broadband everywhere, wireless networking, cloud computing, digital music, steaming media, software as a service—the list of radical innovations goes on and on. But if that person from 1994 were to use a 2011 search engine, he or she would say, "Finally, something that hasn't changed!" The 1994 user would put a search query into a search box on the 2011 search engine, and would receive back long lists of very briefly summarized documents as a search result to consider, just like he or she would have in 1994. Speaking as a member of the search engine industry, I must observe that the community appears to be perversely enamored by the last good idea we had way back at the dawn of the Web. Sure, there have been tweaks to relevance ranking and substantial gains in scalability to keep up with the growth of the Web and electronic publication repositories. But overall, there has been a stupendous lack of radical innovation in search.

Let us consider an example in the use of search applications by professional users. Imagine a company is considering a move into the Internet telephony business. For competitive analysis purposes, a member of the market intelligence staff at the company decides to analyze the strategy of Cisco Systems in "voice over Internet protocol" (VOIP). Assume there is a content repository available to the user of a few hundred thousand market research reports from scores of authoritative information technology market analysis firms like Gartner, Forrester, and IDC.

When this user searches for "Cisco and VOIP" in the content repository of IT analyst reports using a search engine of the current generation, the search result will list thousands of reports. Having produced this lengthy search result, the search engine then washes its hands of the situation, metaphorically dumping the pile of documents on the user's desk and saying, "Cya!" as the search engine bolts out the office door. The user is left to sort through the pile, and find some documents the user thinks might be interesting to read. The search result itself provides precious little guidance in this process. For example, the search result will be sorted by some secret formula that

Digital Object Identifier: 10.1109/JPROC.2011.2105531

will attempt to put documents estimated to be more relevant nearer to the top of the list. And there will be a little summary provided of each document; perhaps a sentence or two of text that the user can review. Acting on these scant hints, the user selects a few reports to read. Some are helpful, some are not, and the user perseveres for as long as he or she has time, or for as long as he or she can tolerate this hit or miss process.

Because one cannot know what one did not find, there is no objective way for the user to assess whether the documents that he or she actually took the time to read comprehensively represent the body of knowledge contained in the thousands of returned documents on the search result. What the user is actually doing is desperately wishing that the few documents he or she selects to read contain all the important findings, analysis, and perspective available on the topic. As even the most determined researcher will read only a very small percentage, typically a small fraction of one percent, of the reports or journal articles on any given search result, this research strategy can best be characterized as *hope for amazing good luck*.

Hope for amazing good luck as a strategy for dealing with search results is not, of course, the fault of the user. It is the fault of a search engine industry that believes that a list of documents is the right response to a user whose business purpose for doing the search is to gain intellectual command of a body of knowledge, to discover new knowledge, to answer a profound question, to explore the meaning of events and trends, or, in our example, to analyze the business strategy of a leading company that can drive the evolution of a new technology and its market.

I. INTRODUCING MEANING EXTRACTION

So how might search work better? One way is by applying *meaning extraction*, an emerging technology that identifies concepts contained within documents and document repositories, and surfaces combinations of these 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 companies to search electronic document repositories and various online resources to dramatically improve and accelerate a researcher's ability to gain insight into a topic and answer specific research questions.

Meaning extraction works as follows.

1) Extract references to important concepts from every document in the research repository, particularly concepts that imply meaning for the business or professional purpose of the search.
2) Record the location of each concept in each document in the research repository.
3) Identify patterns of proximity-related concepts that imply meaning to a knowledgeable practitioner.
4) Analyze the documents responsive to a search query to identify those patterns and highlight them to the user.

II. MEANING-LOADED ENTITIES

Entity extraction itself has been around the text analytics world for almost two decades. Traditionally, the entities being extracted are proper nouns, specifically: people, places, and organizations. For example, text analytics could tell you that Cisco is in a news story, or in ten thousand news stories in the news article repository.

The first useful extension to entity extraction is to include a relevant taxonomy of *context-specific entities* that go beyond the proper nouns used in traditional text analytics. In an information technology setting, these context-specific entities might be technologies, for example VOIP, cloud computing, or software as a service. In different settings the relevant set of context-specific entities would be different. For example, in a pharmaceutical setting one might include context-specific entities like diabetes, Lipitor, or monoclonal antibodies.

The second and more profound extension to traditional entities is the idea of *meaning-loaded entities*. Meaning-loaded entities have depth- and purpose-driven relevance. Meaning-loaded entities are events, conditions, situations, outcomes, actions, relationships, and trends that imply significance for the professional purpose of the search. For example, in a market intelligence search application, meaning-loaded entities might be *price cut*, *change in market share*, or *strategic partnership*. In a pharmaceutical setting, meaning-loaded entities might be *clinical trial*, *patent lost*, or *generic drug*.

A meaning-extraction-enabled search application can index and record the locations in all documents to all references to traditional entities (e.g., Cisco), context-specific entities (VOIP), and meaning-loaded entities (e.g., strategic partnership). Just for shorthand, let us refer to these three entity types collectively as *concepts*.

An effective meaning extraction application requires tens of thousands of concepts to facilitate the meaning discovery search process. As the concepts are identified they are organized into a taxonomy—a *meaning taxonomy*, if you will. (Practically speaking, the meaning taxonomy usually precedes the concept identification.) The meaning taxonomy is designed using a hierarchy that is specifically relevant to the context. For example, VOIP would be placed into the IT Technologies node of the meaning taxonomy while strategic partnership would be placed in the Corporate Strategies node. For the pharmaceutical example, diabetes is placed in the Diseases node, Lipitor in the Drugs node, and monoclonal antibodies in the Proteins node.

Meaning extraction exposes the concepts found in documents responsive to a search query to the user at both the document level and the search results level. At the document level, the meaning-extraction-enabled

search engine presents the concepts found in a document as an enhancement beyond the all-to-brief document summary of the style supplied by traditional search engines. This assists a user in gaining an at-a-glance understanding of what is really in the document so the user can make a more informed decision about whether this report or journal article should be downloaded and read. This facility helps a user find those reports and articles that are most likely to be of the most value, which is crucial considering that only a few documents from a long list of search results are actually going to be read.

At the search result level, identifying the concepts found in all the documents on the search result represents an overview of the knowledge that is contained in those documents. Such a summary overview provides an opportunity for knowledge discovery that can surprise the user with insights otherwise unavailable, or at least unlikely to be discovered with the *hope for amazing good luck* search strategy.

III. AUTOMATIC IDENTIFICATION OF SCENARIOS

After the concepts are identified and organized into the meaning taxonomy, the next step is to interpret combinations of concepts as potentially significant. A human practitioner in the relevant knowledge domain specifies patterns of concepts that when found in proximate relationship to one another imply meaning to the professional researchers using the search engine. Let us call the relationships among concepts that fit the specified patterns *scenarios*. Because the scenarios can be specified at the taxonomy-node level, a pattern efficiently expressed in simple terms by the human expert can expand at runtime into a search for many, many specific relationships between individual concepts.

The meaning-extraction-enabled search engine analyzes all the documents on a given search result and identifies the scenarios, flagging those found for the user to review. This automated analysis to identify scenarios represents the power of the machine to lever human intellect. For example, in one deployed pharmaceutical application, the meaning-extraction-enabled search engine looks for 1.9 trillion potential scenarios on every search against a repository of 25 million journal articles, returning those scenarios found in a given user query in less than ten seconds.

The efficiency of using meaning taxonomies to create scenarios is illustrated by the fact that the text specifying the above scenario pattern consists of only 38 words that instruct the meaning-extraction-enabled search engine to look for relationships between all of the entries in the taxonomy nodes for drugs, diseases, cells, cell receptors, medical devices, proteins, enzymes, genes, and therapeutic strategies. Then for each document in the 25 million journal article repository that is returned on a search result, the analytical process in the meaning extraction step examines all combinations of concepts in all specified taxonomy nodes to find those concepts that are in proximate relationship to one another according to the pattern of the scenario specification.

The relationships found by the meaning-extraction-enabled search engine should be considered *scenarios*, rather than *conclusions* or *findings*, since it is impossible at present to automatically determine if the identified relationships are obvious, spurious, or significant. That question is left for the human intellect of the user to ponder. The meaning extraction application can only determine that the scenarios are *present*, and it can measure the number of documents each scenario is found in, which is the single most helpful indicator of *weight*.

Once the scenarios have been identified from all the tens of thousands of reports or journal articles on a search result, they are presented to the user to consider. In the example we started with, a search of IT analyst research reports discussing the VOIP market produces these actual search results.

- Cisco is using a corporate strategy of acquisitions.
- Cisco is using a corporate strategy of strategic partnerships.
- Cisco is using a product marketing strategy of market segmentation.
- Cisco is using a product marketing strategy of target market.
- Cisco is using a product marketing strategy of professional services.
- Cisco is using a product marketing strategy of service and support.

It is immediately obvious that these scenarios are no ordinary search results; Cisco's strategy in the VOIP market jumps right off the page without reading a single document. The search engine is suggesting that Cisco is targeting specific market segments in the VOIP market and using a combination of high levels of professional services and support and partnerships/acquisitions, presumably to penetrate the market quickly. Each of the search results listed above is linked to a list of reports that discuss that scenario, sorted by the number of times the scenario is in the report so users can rapidly drill into the documents that best elaborate on Cisco's strategy. So, for example, the user could drill down into the subset of reports that present target markets and market segments to learn what those target market segments are.

Also, consider that had the user in the example above used a current generation search engine that only returned a list of documents, that user was in danger of missing the point entirely. He or she would have read two, three, maybe five reports from the search result of thousands. Since the user did not include "strategic partnership" or "professional services "in the search query, the relevance ranking formula probably would not have placed documents rich in those concepts at the top of the search

result. Current generation search engines have the flaw of giving the user what was asked for, not what the user should have asked for had the user already understood the topic.

IV. IMPLICATIONS FOR RESEARCHERS

Meaning extraction supports knowledge discovery by presenting the user with the concepts and scenarios present in the search result as a whole without the blind spots inherent in the *hope for amazing good luck* search strategy, and by presenting these concepts and scenarios, with easy drill-down into the most interesting ideas, meaning extraction reduces the *time to insight*. Users actually end up reading more reports and journal articles with a meaning-extraction-enabled search engine, despite the helpful automated discovery of meaning, because they find more on target, better, and often surprising, material.

Certainly, the value of meaning-extraction-enabled search extends far beyond the business domain; there are clear applications in scientific research, technology development, and many other areas of endeavor. Scientific research stands the most to gain from meaning extraction. Imagine the search engine could read all the papers published in a technical field and identify the *new* concepts and scenarios for you. Everything required to produce such an implementation of meaning extraction is well understood today. It is inevitable that there will be breakthroughs produced by the machines, or more precisely, produced by the human researchers that skillfully guide them.

An interesting question is whether such meaning-extraction-enabled solutions will be able to be made available to individual researchers as opposed to being only available to those researchers working with corporate-sponsored solutions of the type I have described. Individual users may need different concepts and scenarios to support diverse research questions. There is an investment required in preprocessing the document repositories to extract and locate the concepts contained in the repository. This investment involves building the meaning taxonomy, identifying the concepts, identifying the myriad of ways a concept may be expressed in text, processing the repository with the right search tools, examining test cases of results, and then, of course, iterating. Accomplishing this at the level of one user as opposed to one organization requires no new science to achieve. Rather, the only tasks are of user interface design, software engineering, and network operations organization to grant personalized design and processing control of such a system to an individual user.

While these tasks are not trivial in implementation, and the first such individually controllable meaning-extraction-enabled search solution will be much more expensive to develop and operate than a system in which a common set of concepts will be used by many users, there are already such systems being contemplated by organizations that want to give their users individual control over them. It is only a matter of time before an organization, company, or publisher makes the required investment and combines it with a business model that permits individual users who are not employees of the sponsoring organization to have individualized control over meaning-extraction-enabled search applications.

V. IMPLICATIONS FOR PUBLISHERS

And of course, there are implications for publishers. The most interesting implication is that meaning extraction levers the value of *publishers' repositories*. One of the consequences of Web search engines and massive online content is that the value of having unique coverage of a topic has declined. For example, search on a news topic on a Web news search engine and you will have competent coverage presenting the essential elements returned from thousands of individual news sites. No one site can claim any significant coverage advantage over the collection of all sites assembled on the fly by the search engine. And Web news search engines will instantly and seamlessly switch out the publishers contributing to any particular user query, further reducing the value of being a publisher. The Web news search engine effectively eliminates the value of an individual content repository by creating a virtual repository for each query from all indexed sources. (And it keeps that value for itself as a business model.)

A similar process is rolling ahead in search of scientific and technical literature. As more and more content addressing any scientific question is published by multiple sources and made findable by generalized Web search engines and federated search engines of the type often found in libraries, the value of any one piece of content or any one collection of content declines.

Now apply meaning extraction to equation and the situation changes. Because the first two steps of meaning extraction as discussed above (extracting concepts and recording their locations within documents) have to be done as a preprocessing step and not at runtime, a search engine returning results from multiple uncoordinated content sources will not be able to implement meaning extraction. The publisher with a deep and broad repository can implement meaning extraction against the publisher's own repository. The bigger the repository, the better meaning extraction will perform because the application will find more scenarios and will calculate their weight more accurately. Reversing the game, value is returned to the publisher's repository, and hence to the publisher.

VI. CONCLUSION

Meaning extraction represents a powerful tool to help researchers analyze, comprehend, and apply the flood of information that the modern era of pervasive technology has unleashed, so

that in the future, long lists of documents on search results, search circa 1994, will truly be a thing of the past.

Search engines must evolve to have an in-depth understanding of the searched material and the associated ways of knowing in the user's domain of knowledge. It is necessary that search engines grasp the professional purpose for a given search and that search goes beyond presenting document lists to users. Search engines must interpret and analyze the search results and then present findings that would be considered most significant by the users if they were able to read all of the documents retrieved in the search process. Meaning extraction is the future of search. ■