

The New Wave of Concept Search Tools

by George Socha
Socha Consulting, LLC
651.336.3940
george@sochaconsulting.com

Published in LJN's Legal Tech Newsletter, Volume XX, Number 12, March 2003.

Although the concept of concept searching has been around for at least 2,000 years in philosophical circles and was first realized in the software world in the 1970's, it is making big news in today's electronic discovery and automated litigation support world. Over the past year, a series of vendors have introduced software solutions they claim can take us far beyond the results we get using tools built around searching full text or coding data using key words, strings of text, and Boolean search algorithms. Whether it is through mimicking the thought processes of high-level aquatic mammals, developing libraries of semantically and geographically related words and terms, or displaying documents as masses of dots within sprays of bubbles, these software programs, we are told, are the automated litigation support equivalent to Big Blue, the computerized chess champion.

Are they? Yes...and no.

Yes: The electronic discovery market has been growing at a stunning rate – probably doubling each year since 1999 and likely to continue to do so for the next couple of years. That growth means that we are starting to have to deal with cases where the potentially relevant data is the equivalent of tens, or even hundreds, of millions of pages of documents. We cannot work through that volume of data using the methods developed in the 1980s and 90s – scanning, manual coding, and then computerized searching using key words, string searches or Boolean algebra. Even in a bet-the-company case, only the rarest of parties can bear the costs of sending all that data through all those processes. And even then, we just cannot wait the time it takes to complete those steps. The methods of the late 90s – capturing electronic data electronically, converting it to a more readily searchable electronic form and then searching it using key words, etc. – are better but still not good enough. Too often, that means setting 50, 100 or more attorneys before computers and asking them to review all the documents online. Even with culling based on hash codes and similar criteria, the volume of data can still be too great for a meaningful and timely human review.

Enter the concept search tools. Point them to a dataset and they process all the data in that set, not just those portions of the data that were deemed worth capturing through coding. They do this quickly, they do it consistently and they do it without thinking about tonight's grocery shopping instead of the data before them. Some of the tools build indexes, others use auto-coding, yet others use approaches known only to the developers and their patent attorneys.

Although each of the concept search tools approaches the task differently, they all are designed to identify relationships between the content of documents or files in the sets that they process. The concept search tools build those relationships as they process the data or documents; some supposedly learning more about the relationships as they work through more information. As a result, when you, the user, start exploring the relationships in your efforts either to identify key data or eliminate materials from further consideration, the concept search engines already have done most of the hard work. They do a substantial part of the thinking for you, returning back the results they believe you want to see.

One of the tools that fall into the “break-through” category is that search results are displayed in a graphical format. For those of us who are visually oriented, the importance of this cannot be overstated.

And no: Sounds great, doesn't it? So why the “and no?” Because if you are going to rely on these tools, you will need to understand and appreciate their potential limitations. This might sound trite, but in my experience all too often attorneys will expect far more of a search engine that it is able to deliver, will believe that searchable data is for more reliable than it really is and will have the most unrealistic expectations when they have the least understanding of how the software is working.

First and foremost, these tools are being oversold. Based on the demonstrations I have seen, they should be great for prioritizing data and documents, letting you hone in on the documents you mostly likely will care about the most. Unfortunately, some of the vendors do not stop there; they promote these tools as a way to cull out materials from further consideration. What is the problem with that? Consider how you would explain to a judge that you failed to produce a key document because a computer programmatically sent it to the trash bin instead of the “consider” pile.

A second and related problem is that most, if not all, of the concept search tools being promoted to the legal community work in proprietary ways that their creators refuse to fully explain.

A third problem has to do with those tools that rely on OCR processes to gather data from scanned documents. Data that gets into the system through an OCR process may not be of very high quality because of the limitations of the OCR processes. A 95% accuracy rate means that on average one of out every twenty characters is not converted properly. The previous sentence had 85 characters in 18 words. With a 95% accuracy rate, 4 of the characters would have been converted incorrectly. If each wrong character were in a different word, 4 out of 18 words would have been wrong – a word error rate of just over 20%.

Finally, some of the concept search tools appear to lack structured datasets that use normalized data. Without those, how does one sort by date, search by author, look for all memoranda created by plaintiff between January and May of the critical year, etc.?

A Concept Search Engine Sampler

All that said, the emerging concept search engines are worth looking at. Presented in alphabetical order, here are a few of the products that have come to market in the last year.

Attenex Patterns Workbench and Attenex Document Manager (www.attenex.com): ASP model. Concept search engine; categorizes electronic data and displays the results in a graphical format. Offered for culling, deduplication, categorizing, review. Drill-down capabilities including display of data in quasi-native format. This set of products is worth the price of admission just for its graphical depiction of data. Go to the web site and look at the Attenex Document Mapper. It shows documents as dots with bubbles. The bubbles represent clusters of related documents. The bubbles then are strung together in vertical, horizontal, and diagonal chains.

Cataphora C-Evidence Service (www.cataphora.com): Strictly speaking, not a concept search tool; focuses on context rather than content. Displays results as a set of linked pictures. Tag line: “making sense of potentially cryptic evidence by putting it into its proper context.” Offered as a service. This solution uses meta-data and text of documents to link "actors" with documents. It tries to link documents with each other, such as chains of e-mail messages. It shows this information as sets of linked pictures. It compares the content of the documents to ontologies (similar to a thesaurus, but with related terms and concepts as well as related words). It uses this to determine the likely topical content of the documents.

LexisNexis e-Discovery powered by DolphinSearch (www.dolphinsearch.com, www.lexis.com): ASP model. Concept search engine; based on neural networking and relational pattern recognition. Native format review rather than TIFF-based approach. Offered for culling, hyper-duping, filtering, review. Described on DolphinSearch's website as “a text-reading robot powered by a computer model of the extraordinary pattern recognition capabilities of a dolphin's brain.” Watch out for these folks, if for no other reason than because Lexis, one of the two anchor tenants of the legal technology mall, is involved.

Prevail (www.fiosinc.com) and Semetric (www.engenium.com): ASP model. Concept search engine; organizes unstructured data by “mapping relationships between each word and every other word in large sets of documents.” Conversion of electronic information to structured database, full text, and TIFF images. Filtering, deduplication, production as TIFF images. Fios recently incorporated Engenium Corporation's Semetric “intelligent concept-based search technology” into Prevail, Fios' web-based electronic discovery management and review tool. This marriage of two sets of tools should yield a whole that is greater than the parts.

SnyDex and Synthetix (www.syngence.com): SynDex auto-codes documents, programmatically populating a set number of fields in a database. They include first and

last page ID numbers, document type as determined by the software, document date, other dates from the document, names, author, recipient, copyee, subject from the re: line of the document, pre-determined key words, and the ORC'd or otherwise captured text of the document. Synthetix is Syngence's search engine that works at a document level. Assume you have a document of great interest and you want to see more like it. Synthetix will look at every word in that document and then look for other documents containing the same or similar sets of words. Synthetix then ranks the results, with the highest rankings going to documents that are the most similar to the starting document. If desired, one can even create an ideal document and have Synthetix search against that.