



Good for you.
Great for your clients.



ORANGE
LEGAL TECHNOLOGIES

Conceptual Search in Electronic Discovery

By Herbert L. Roitblat, Ph.D.

Conceptual Search in Electronic Discovery

According to the Radicati Group, the average business mailbox contains 4.3 GB of potentially discoverable ESI. The burden of dealing with that volume is obvious. Automated tools to help identify responsive and privileged documents are being used more and more as the only viable way to manage these massive volumes of information. Electronic Discovery professionals have been looking for ways to limit the number of documents that need to be reviewed, because review is usually the most expensive part of eDiscovery. Most commonly, these methods revolve around some kind of content search to distinguish those documents that are potentially responsive from those that are not. Content search also has the potential to make document review itself more efficient and effective, further reducing the burden.

Keywords and Boolean Queries

A common approach to document selection is to develop a list of key terms that will be used to select documents for review. Any document that does not have one of the key terms is ignored. Using key word searches or Boolean combinations of key words is risky in that there is a substantial likelihood that responsive documents will be missed and a substantial likelihood that a large proportion of nonresponsive documents will be included.

Jason Baron, at the National Archives, reports that in the case of U.S. v Philip Morris, they started the selection process with a set of individual search terms (e.g., "Lorillard"), and then, on the basis of some sampling, they recognized that they needed a more elaborate search process to get useful results. They used queries like the one below to try to find useful documents.

```
((master settlement agreement OR msa) AND NOT  
(medical savings account OR metropolitan standard area)) OR s. 1415 OR  
(ets AND NOT educational testing service) OR  
(liggett AND NOT sharon a. liggett) OR atco OR lorillard OR  
(pmi AND NOT presidential management intern) OR pm usa OR rjr OR  
(b&w AND NOT photo*) OR phillip morris OR batco OR ftc test method OR star  
scientific OR vector group OR joe camel OR  
(marlboro AND NOT upper marlboro)) AND NOT  
(tobacco* OR cigarette* OR smoking OR tar OR nicotine OR smokeless OR synar  
amendment OR philip morris OR r.j. reynolds OR  
("brown and Williamson") OR  
("brown & Williamson") OR bat industries OR liggett group)
```

Queries like this are very difficult to construct, difficult to get syntactically correct (for example, to get all of the parentheses and "ORs" in the right order), and difficult to validate. Although they certainly have their place in eDiscovery, without a lot of special care, they can fail to yield the expected results.

In addition, Blair and Maron (1985) found that attorneys, like other users, were relatively poor at guessing the right words to search for. Although they thought that their searches had found 75% or more of the responsive documents, they had, in fact, found about 20% of them, in what would now be a small collection of about 40,000 documents.

Concept Searching

More recently, attorneys and the courts have come to recognize that keyword searching is not up to the task of managing ediscovery. Keyword searching misses documents that should be produced and over-burdens the review team with irrelevant documents. For example, Judge Facciola in *Disability Rights Council of Greater Wash. v. Wash. Metro. Area Transit Auth.*, 2007 WL 1 585452 (D.D.C. June 1, 2007) noted:

I bring to the parties' attention recent scholarship that argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results. See George L. Paul & Jason R. Baron, [Information Inflation: Can the Legal System Adapt?](#) 13 Rich. J.L. & Tech. 10 (2007).

There appears to be two fundamental features of human language that make searching inherently difficult. The first is the tendency of humans to be creative. They are always making up new words (for example, "blog") and they frequently use old words in new ways (think of "twitter," or "Macintosh"). As a result, there are often many ways to say the same thing, and the same word can mean many different things.

For example, there are over 200 words that mean roughly "think," including "guess," "surmise," "ponder," and "expect." Familiar words, like "bark" and "bank," show that the same word can have multiple meanings. What is less obvious, though, is that this ambiguity is far more prevalent than most people would guess. The 200 most commonly used words in the English language have about 23 definitions each.

As an exercise, I looked up in a dictionary each of the words of a simple sentence. The number of definitions I found is listed under each word.

The companies have agreed to a brief delay in implementing their agreement.											
37	14	39	17	54	62	20	8	84	8	7	9

If you combine all of the definitions together you get 7 quadrillion possible interpretations of this sentence, yet hardly anyone notices that it is the least bit ambiguous. Each word in the sentence helps to disambiguate the other words in the sentence. You can see this same phenomenon in play when you read a sentence like:

Fatty weighed 300 pounds of grapes.

Or

The young man had his palm examined by a tree surgeon.

By the time you get to the end of the sentence, you may have to re-interpret the earlier words in the sentence.

The difficulty of recognizing this ambiguity is one of the overarching problems with using keywords to select documents in eDiscovery. People just find it hard to imagine all of the ways that an author might have said something or all of the different interpretations that could be attached to a specific word.

Traditional keyword searches try to get at this ambiguity by adding "OR" or "NOT" to the query. Apparently, Baron was interested in finding all documents about a "master settlement agreement," but the document authors probably did not always spell this whole phrase out each time. They may have referred to it as an "MSA," or even as "an agreement." The problem is MSA also refers to a "metropolitan standard area." Baron used "NOT" to exclude documents that mentioned metropolitan standard area, or medical savings account, but if a document contained both master settlement agreement and metropolitan standard agreement, this Boolean expression would exclude that document. It would select documents that used MSA to refer to metropolitan standard area, so long as they did not also use the longer phrase. In short, even after this effort, there is still a potential to reject a lot of potentially responsive documents and to select a lot of non-responsive documents.

No technology can ever completely eliminate this risk. People are just too, well, human, and creative in their writing. Even human readers can fail to recognize responsive documents and falsely select nonresponsive ones. Still, as Judge Facciola noted (quoting, by the way from a paper by the same Jason Baron), concept searching is likely to be more efficient and comprehensive than keyword searching.

There are many technologies that could be called "concept search," each with its own ways to generate concepts. Among these are ontology/taxonomy/thesaurus-based methods, where knowledge engineers, often working with domain experts, construct lists of related terms. A thesaurus, for example, is among the simplest of these. It is just a list of words and their synonyms. Ontologies and taxonomies are more elaborate and can include other kinds of word relationships. Given one word, these systems provide others that may be useful for identifying related documents.

These systems are good at solving the synonymy problem—having multiple words that mean about the same thing. Because each organization has its own jargon and vocabulary, however, these systems require a substantial up-front effort to tune the list to accommodate that specialization. And, they only take advantage of the specific relationships that the knowledge engineers and experts put into them.

A second approach does not use human-designed information structures like a taxonomy or thesaurus, but gets its word relationships from the documents that it processes. This is the approach taken by the OneO® Discovery Platform through its use of the OrcaTec Information Discovery Toolkit. In eDiscovery, it learns the meaning of words from the documents in the collection while it indexes them. For example, if a document has the word "lawyer" in it, then it is also likely to have words like "judge," "case," "attorney," or "court." Conversely, if it has words like "judge," "case," "attorney," or "court," then it is likely to be about lawyers whether it has "lawyer" in it or not. Conversely, a document with the word "court" in it is likely to be about a legal issue if it has words like "lawyer" in it and to be about sports if it has words like "tennis." These word-use patterns are, then, very helpful in identifying what a document is actually about. As in the sentences mentioned earlier, each word helps to disambiguate the other words.

Whether the relations come from the minds of knowledge engineers and experts or from the documents, concepts consist of sets of related words. When a user enters a search query, the query is expanded to include these related words. Whatever terms a user enters, the system finds related terms and adds them to the query. These additional terms focus the search and allow additional documents to be retrieved. Concept search, particularly as performed by the OneO® Discovery Platform, does not just help to expand the number of documents retrieved, it helps to focus the results on the most relevant content.

One problem that concept search is intended to address is the difficulty guessing the right words to search for. Concept search allows users to find additional documents that may be about the same topic as the query term, even if they use different words. The additional query terms allow the system to find related documents even when they don't happen to have the query word in them.

An Example in Japanese

Japanese	Transliteration	Transaction	Weight
化粧	keshou	makeup cosmetics	10.0
美容	biyou	beauty of shape or form	0.0070
肌	hada	skin	0.029
美	bi	beauty	0.069
ケア	kea	care	0.069
dhc		mail order makeup company in Japan	0.08
スキン	sukin	skin	0.11
キレイ	kirei	pretty	0.122
始め	hajime	Beginning, start origin	0.129
口コミ	Kuchikomi	word of mouth	0.192
通販	tsuhan	mail order	0.194

Here is an example of how the OrcaTec Language model generates concepts. This example comes from a set of Japanese Women's blogs. The original search term in this example was 化粧, which is pronounced "keshou," meaning "makeup," or "cosmetics." From having read these documents, the system learned that keshou is related to words like biyou, hada, and so on. No one taught it these relations, rather they were extracted automatically from the patterns of word use in the documents.

These related terms are not necessarily synonyms of keshou, but they do form the context in which this word can be understood. They reflect the interests of the blog writers rather than of some thesaurus writer. Some of them, such as hada or bi, can be found in a dictionary, but others, such as dhc or kirei (as it is written here) do not appear in a dictionary. If the system can learn Japanese from the documents, you know that it will be able to learn an organization's jargon and peculiar word-use patterns.

The rightmost column in the table shows weights. A weight is a measure of the importance of the term to the search, or how closely each word is related to the original search term. The relations between words are not all or nothing, rather they can vary in importance.

Keshou has the highest weight because that is the word we searched for. Biyou has a low weight, meaning that this word is only tenuously related to the original search and tsuhan has the highest weight. This variability is part of what gives the system its power to find relevant documents. The weights control the ranking of the search results so that the documents that best match the search and the context of that search appear at the top of the results list. These are the documents that are most characteristic of the query term as understood in the context of the document collection.

How to Use Concept Searching in Electronic Discovery

The advantage of concept searching is to give the litigation professionals an edge in identifying relevant documents and in understanding the information that is available in a collection.

Concept search focuses the results on the documents that are most characteristic of the search term in the context of the collection. It lets reviewers see immediately how a term is used in the collection. This is an excellent training tool, in that reviewers get to see how a term is used and in what context. If Boolean searches rank documents, it is usually by how often a term appears in a document. Concept searching is also sensitive to the frequency of the query terms, but it also takes full advantage of the related terms and their weights.

Concept search allows reviewers to identify terms that are related to the original query term. Not all of these terms are going to be useful by themselves, but they often suggest important avenues of enquiry that would be difficult to identify any other way. Consider the appearance of DHC among the terms related to keshou. Without special knowledge of beauty products and blogs, it is unlikely that this term would have been predicted.

By organizing documents according to their computed relevance, reviewers have the opportunity to bring together all of the documents in a collection that are relevant to a particular issue. People are limited in the number of categories they can actively keep in mind at one time. After about 5 - 9 categories, they begin to make mistakes. Because the concept search focuses and gathers together documents that are putatively relevant to a query, reviewers can focus their attention and increase the accuracy of their judgments. It is much more efficient for a reviewer to agree or disagree with the recommendation of the computer (e.g., as to whether a document is responsive to an issue) than it is to try to decide which of several issues, if any, a document is responsive to.

Because the concept search ranks the documents according to its estimate of how relevant the documents are, reviewers can see several documents in a row that are likely to be responsive. Ordinary reviews organize the documents in more or less random order relative to the issues of a case. Moreover, usually 80 - 95% of the documents that may be under review are nonresponsive. The more nonresponsive documents a reviewer sees in a row, the more difficult it is to recognize another document as responsive. By organizing the documents topically, the likelihood of encountering responsive documents is increased, thereby increasing accuracy.

The topical organization can also be used to allocate resources. More senior people or specialists can be assigned to specific topics and to documents that are most likely to be responsive to those topics. Less expensive resources can be assigned to documents that are less likely to be critical. Documents that are more likely to be responsive can be prioritized and reviewed first. That way, if there is a shortage of time, the most likely documents will have been reviewed first.

Concept search can be used to cull the documents with more power than can be obtained through just keyword and Boolean searches. The OneO® Discovery Platform search system allows a full complement of Boolean expressions in concert with concept searching. Concept search also computes scores for each document that are an estimate of just how relevant the document is to the query. Documents with higher scores (closer to 1.0) are the most likely to be relevant. Relevance tails off as we move farther down the list to lower and lower scores. We recommend using a cutoff score (e.g., 0.01) when using concept search for culling. Documents with scores below this value are somewhat related to the query, but the relationship is usually too tenuous to be useful.

In the end, the quality of a review depends on the probability of finding relevant documents and the probability of correctly classifying the documents that have been found (e.g., responsive or nonresponsive). Concept search can help with both of these. It increases the likelihood that responsive documents will be presented for review, prioritizes them so that the most likely to be responsive are presented first, and offers to the reviewer a suggestion as to just how a given document is responsive.

Conclusion

Concept Search tools can help to ensure that you achieve rapid and thorough understanding of your case collection. Collections have grown so large and so complicated that without such powerful tools there is no way to reasonably prepare for and execute a review. The alternatives have simply grown too expensive and they require too much time. The OneO® Discovery Platform powered by OrcaTec, is designed to give rapid and effective insight into the nature of the collection and to facilitate the rapid, effective, and inexpensive review of those documents.

Epilogue: Need an Advanced Platform? Many Vendors. One Choice. OneO®.

Orange Legal Technologies' OneO® Discovery Platform provides distinct and quantifiable advancements over current electronic discovery services and provides the following capabilities with in-house proprietary technology:

A Complete Electronic Discovery Platform: OneO® can provide analytics, processing, and review – the core tasks of electronic discovery – from within a single platform. *This means that once data is received and ingested, there is no need for an additional platform or provider to complete these key electronic discovery tasks thus saving clients over 50% of the time and 50% of the money required for electronic discoveryⁱ when compared to traditional offerings.*

An Integrated Electronic Discovery Platform: OneO® architecture provides for integration of electronic discovery tasks at the application level vs. the platform level. *First, this means that data transfer between the key tasks of analytics, processing, and review occurs within the OneO® platform thus increasing the defensibility of evidence by both reducing the risk of potential spoliation that can occur when transferring data between platforms and/or service providers and providing a defensible process. Secondly, this application level integration helps OneO® index documents twice as fast as other leading solutionsⁱⁱ - substantially decreasing the time and cost of electronic discovery.*

An Online Delivery Model: OneO® is delivered to clients via a Software-As-A-Service Model (SaaS). *This means that there is no additional client-side resource or infrastructure investments necessary to implement and maintain the OneO® Discovery Platform – thus providing client's cost savings for today and investment protection for tomorrow.*

About the Author

Herbert L. Roitblat, Ph.D.

Dr. Herbert L. Roitblat is a co-founder and Principal at OrcaTec LLC. Before starting OrcaTec, Herb was Executive Vice President, Chief Scientist, and co-founder of DolphinSearch. He is the primary inventor of the core DolphinSearch technology (patent No. 6,189,002). Herb is also recognized expert in cognitive science, information management, data mining, statistics, and eDiscovery processes. He is the author of numerous papers on dolphin biosonar and neural network models of the dolphin sensory system. More recently he has been writing about data mining and how technology can ease the burden of discovery.

Herb was an award-winning Professor of Psychology at the University of Hawaii from 1985 until 2002. He taught courses in cognitive science and research methods. He has a B.A. degree from Reed College in Portland, Oregon and his Ph.D. in Psychology from the University of California-Berkeley. He served as Assistant Professor of Psychology at Columbia University until he joined the faculty at the University of Hawaii. Herb is a Past President of the Division of Behavioral Neuroscience and Comparative Psychology of the American Psychological Association and Past President of the International Society for Adaptive Behavior. In 2002, he received the Clifford T. Morgan award for distinguished contributions to behavioral neuroscience and comparative psychology. Herb is a member of the Sedona Working Group on Electronic Document Retention and Production.

About Orange Legal Technologies

Orange Legal Technologies is a leading provider of one source litigation, audit, and investigation support services for law firms and corporations seeking insight on electronically stored information. Headquartered in Salt Lake City, Utah, and with four locations nationwide, OrangeLT™ offers a complete suite of electronic discovery services to include collection, analysis, processing, review and production of both digital and paper-based information. Enabled by the OneO® Discovery Platform—an integrated, web-accessible electronic discovery platform that provides online analysis, processing, and review of unstructured data from the security of a hosted centralized repository—and augmented by best of breed electronic discovery partners, Orange Legal Technologies is a member of the Electronic Discovery Reference Model (EDRM) and the International Legal Technology Association (ILTA).

Contact

For more information on Orange Legal Technologies, visit our website at OrangeLT.com, contact us via email at info@orangelt.com, or contact us via one of our four domestic locations.

Salt Lake City – Headquarters

251 South Floral Street
Salt Lake City, UT 84111
801-328-4566

Los Angeles

350 S. Figueroa, Suite 199
Los Angeles, California 90071
213-624-8688

San Francisco

98 Battery St., Suite 250
San Francisco, CA 94111
415-989-7922

Spokane

421 West Riverside Avenue, Suite 319
Spokane, WA 99201
509-744-0200

ⁱ Orange Legal Technologies, Predictive Pricing Estimator, August 2008. (100GB Estimated Client Volume At Initiation).

ⁱⁱ Clearwell Systems Rapid Indexing, <http://www.clearwellsystems.com/products/e-discovery-processing.php>, December 28, 2009. (Clearwell Indexing @ 10-12GB/Hour, OrangeLT® Indexing @ 25GB/Hour).