



WHITE PAPER

**There has to be a  
better way to search...  
and there is!**

---

Fios, Inc  
921 SW Washington Street  
8th Floor  
Portland, Oregon 97206  
(503) 265-3467 Toll Free  
[www.fiosinc.com](http://www.fiosinc.com)

---

Content Analyst Company, LLC  
11720 Sunrise Valley Drive  
4th Floor  
Reston, Virginia 20191  
(888) 349-9442 Toll Free  
[www.contentanalyst.com](http://www.contentanalyst.com)

---

## Searching for defensibility...and results

It is no secret that the majority of the cost of discovery resides in the cost of the review. Often, more than 80% of total electronic discovery costs can land here. It is exactly that metric that leads to the existence of e-discovery providers, who have the experience and capacity to ingest large (as in huge) amounts of raw data, disassemble that data to its lowest common level, and then systematically and defensibly separate the chaff from the “potentially responsive.” Full-service providers utilize many techniques and industry standards to accomplish this modern miracle, such as:

- » The repository of known software, file profiles and file signatures maintained by the National Software Reference Library (NSRL) from the National Institute of Standards and Technology (NIST)
- » Standard data de-duplication methods (believe it or not, more than 27 standard de-duplication measures exist)
- » File signaturing to identify the thousands of file types that may exist throughout an enterprise (your file names can run, but they cannot hide!)
- » Complex set theory
- » Boolean search filters – the most relied upon and yet potentially the most error-prone culling technology.

The effectiveness of a search filter depends a lot on the underlying data population as well as the ability to identify keyword filters that will accurately isolate the potentially responsive data. Normally, it is less risky to err on the side of over-inclusiveness if there is a choice to be made. However, that choice leaves more data to be manually reviewed and categorized, with obvious impact on costs.

Furthermore, the effectiveness of any keyword filter is dependent upon the understanding and imagination of the list developer. For instance, if you want to find all items related to, say, “seatbelt failure,” is your strategy expansive enough to have included related references such as “passive restraint system breakdown”? In another example, all patents related to “motorcycles” are requested. Is the author of the keyword list knowledgeable enough to include “open air bi-wheeled frame”? (Yes, there really is a patent for that, and, yes, it really is a motorcycle.) Other barriers to successful searching include failing to think of all of the creative ways people can misspell words and names, as well as the nicknames, acronyms or non-case-sensitive terms that may be used to describe a product or person that might not be known to the search creator.

Of course you may be able to construct a very suitable list of search terms, but if someone does not properly “optimize” the list, the desired results may prove elusive. In most modern search technologies, arcane words and symbols are used to specify things like “wildcard” substitution characters or the proximity of words to each other (e.g., term 1 within 12 words of term 2). A lack of expertise in the “lingo” of search syntax can severely hamper the efficacy of a search. In one real-life example, a company was in the midst of a Department of Justice second request as a result of a pending acquisition and had only 45 days to respond. After collecting two terabytes of data, counsel needed to cull down the data to what was potentially responsive to the request. They began the culling process by searching for the stock symbol of the acquired entity. Imagine their surprise when all but one item in the entire document set came back as potentially relevant because the stock symbol was included in the footers and email signatures of nearly all the documents.

The challenge with search term filtering is that it is an imperfect science. While some filters, such as date ranges, are usually reliable when properly applied, keyword filters can be notoriously inaccurate. Keyword filters are simultaneously over-inclusive and under-inclusive in that they have the potential to capture non-relevant documents as well as leave behind relevant documents. In a study of the efficacy of search and retrieval techniques used in the legal community, presented at the 2006 and 2007 Text REtrieval Conference (TREC) proceedings, researchers found that Boolean searches yielded only 22 to 53 percent of the actual relevant content when used to cull the seven million documents stored in the tobacco litigation Master Settlement Agreement database.



More evidence of the inadequacy of conventional search term filtering was recently provided by The Sedona Conference. Of particular note are the two following practice points:

- » “In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.”
- » “The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.”

Even the courts are taking a much deeper view of how potentially relevant data is discovered. Judge Grimm, writing for the US District Court for the District of Maryland in the *Victor Stanley v. Creative Pipe* case stated, “[W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying on such searches for privilege review.”

A similar sentiment was also expressed by Judge Facciola, writing for the US District Court for the District of Columbia in *U.S. v. O’Keefe*: “Whether search terms or ‘keywords’ will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. See George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 *RICH. J.L. & TECH.* 10 (2007) \* \* \* Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread.”



NLP is very similar. It relies on a wide range of outside information, or “auxiliary structures,” that help differentiate between similar words, variations of words and phrases, slang, technical terms and so on. If you want an NLP system to deconstruct a sentence like, “The bank is on the left,” it doesn’t know if you mean First National Bank in Peoria, Illinois, in La Rive Gauche along the Seine in Paris, or on an airplane that needs to dip its wing to make a left-hand turn. What is “bank”? How is it being used? An NLP technique needs to be told, in essence, how to think about “bank” in this context.

Needing to define the context is a real problem for two reasons. The first is that you don’t always know how a particular term, like “bank,” is being used. The second and far more serious problem is that language changes rapidly and regularly. Think about today’s nomenclature. To “google” something is to look it up on the Internet using a popular search engine; however, as recently as five years ago, if you offered to “google” something you’d raise an eyebrow or worse. This problem is exacerbated in e-discovery, where everything happens in the past. When we are trying to analyze a large collection of documents, those documents may likely have been produced five, 10 or more years ago using the terms and language of the time. Let’s take cell phones and hand-held organizers as an example. What did we have for personal digital assistants (PDAs) or smartphones 10 years ago? Are those the same as Palm Pilots? What’s an Apple Newton? Could it make a phone call? We might be experts in forensics, but we’d rather leave history to the academicians.

The two competing mathematical approaches overcome these NLP limitations, but they also present some of their own challenges.

---

## Bayesian Inference

Bayesian Inference has been around for at least 250 years. “Bayes’ theorem” is credited to Reverend Thomas Bayes in a paper published in 1763 after Bayes’ death by his friend Richard Price. It turns out that Bayes himself might not have invented his “theorem.” There’s compelling evidence that one of Isaac Newton’s close

friends, the blind mathematician Nicholas Saunderson, developed it in the early 1700s.

The main idea behind the theorem is conditional probability. It’s actually fairly simple: If a given event occurs a number of times under certain conditions in the past, and if you can identify other events and similar conditions, there’s a high probability that these events are related. Bayesian Inference has been used to successfully predict any number of things, from drug test results to poker game outcomes.

It can also be used to identify “concepts.” The idea here is to look at a whole collection of documents and identify how people, places, things and events are described in a similar manner. It actually works quite well. Bayesian Inference will be able to determine that “the bank on the left” and “the First National Bank in Peoria, Illinois” and “the financial institution located left of Sanderson in Peoria” are all phrases about the same place. Bayesian inference is at the heart of some very large and successful enterprise search solutions.

For e-discovery, however, there’s a problem. The more conditions you introduce into Bayes’ theorem, the more “variants” of that formula you produce. Put more simply, it’s like factoring:  $2 \times 2 = 4$ , but  $2 \times 2 \times 2 = 8$  and  $2 \times 2 \times 2 \times 2 = 16$ . For each additional factor, the increase in the number of potential results is exponential – a result exactly the opposite of our stated objective of reducing the volume of potentially responsive ESI.

Thus, with Bayesian Inference, the more complex the events and conditions, the more “math” the system is going to have to perform to get a result. In the world of e-discovery, it’s rarely as simple as “Mary told Frank about the bank draft.” Very complicated situations can cause software solutions based on Bayesian Inference to go into overdrive. Worse, they may tend to “narrow in” on a given set of results, either because they aren’t able to see all the variations or because the software has been “choked” to yield the desired performance.



---

## Latent Semantic Indexing

Latent Semantic Indexing (LSI) was first developed at Bell Labs, which subsequently abandoned it. The idea behind LSI is to create a kind of “hyperspace” with lots of dimensions and other mathematical “things” that we mere mortals struggle to understand, and then plot every word and document into this “space.” On some dimension or plane somewhere in this space, documents that are similar are going to get plotted close together. The “closeness” or proximity of similar documents to each other can even be measured. Give an LSI system enough documents and relationships start to emerge.

Like Bayesian Inference, LSI is based purely on mathematics. LSI-based applications don’t get “hung up” on specific words or synonyms. Multiple meanings can be attributed to a single term, such as in our “bank” example. They also have the ability to compare concepts across language boundaries. With a mathematical approach, it’s easy to “train” the system to recognize that “Comment allez-vous?” in French means “How are you?” in English simply by having it analyze two documents that are identical in content but in different languages.

Unlike Bayesian inference, LSI doesn’t expand in scope as variations increase. Every word and every document gets a “plot” or value, regardless of how obscure the word or complex the document may be. This is how the “index,” or vector space, that LSI uses is created. When you go to search an LSI index, it functions as if the system has total recall. In reality, it is finding only the “plots” or “vectors” that are most similar to what you’re searching. The Achilles’ heel of LSI has traditionally been hardware. LSI uses computer memory – lots of it. Let’s face it, “total recall” means you have to “remember” lots and lots of mathematical relationships. This was a problem in the Bell Labs days, when the Intel 8086 microprocessor represented the most advanced processing technology. Today, however, memory is available and cheap, and today’s PCs and servers have 64-bit architectures that allow them to access tons of this inexpensive memory.

Concept-based searching using LSI has been applied to the e-discovery process by leading providers as early as 2003. Newer search applications exploit LSI in ways many companies and their legal teams have never considered. By taking full advantage of conceptual search methodologies, recent technologies are able to offer powerful new features, such as the automatic determination of themes within the data, visual mapping of the relevance of one concept to another and support for cross-lingual capabilities.

We accept Boolean searches as the standard for e-discovery. Most people, including lawyers, are very comfortable with keyword searching. Put in a string of keywords and get back documents that contain those keywords. But



Boolean searches are only de facto standards, and the courts are growing increasingly frustrated with legal wrangling and maneuvering involving ESI. Grimm, in *Victor Stanley*, clearly stated that keyword searching is not enough for privilege review.

The amendments to the Federal Rules of Civil Procedures (FRCP), enacted in December 2006, were intended to change that. One outcome of the FRCP amendments is the requirement for opposing counsel to “meet and confer” early in the process and present the courts with their strategy for searching, identifying and preserving ESI relevant to their case. But that requirement has hardly resolved the many difficulties related to limiting the scope and costs of e-discovery. In fact, to the increasing dismay of many in the legal community, the “meet and confer” requirement often results in a situation where the less-prepared party is compelled to agree to keywords, time frames or other search parameters that virtually guarantee either an over-inclusive, overly expansive document set requiring review or a much greater risk that that responsive documents will be missed.

The recent rulings by Judges John M. Facciola and Paul W. Grimm both cite the shortcomings of keyword and Boolean searches and illustrate how an e-discovery process based solely on keyword techniques runs the risk of derailment. These justices, along with dozens of others e-discovery experts, are pressing defendants and plaintiffs alike to augment their “comfortable” keyword search techniques with advanced searching. One might ask if a single watershed case will settle the defensibility of advanced searching, or will dozens of narrower rulings pry the door open and gradually elevate advanced search techniques to the same admissibility as keyword? The details are hard to predict, but the general direction of the courts is clear: Advanced technologies, such as concept searching, supported by sound, defensible process and expert management, will be in increasing demand, especially with judges who are well-versed in ESI.

Are you prepared?

---

## About Content Analyst

Headquartered in Reston, Virginia, Content Analyst® is a leader of advanced search tools and technology. Content Analyst’s software, which includes patented Latent Semantic Indexing technology, provides advanced, conceptual-based search and document analysis for a wide range of customers, from highly-classified intelligence installations to world-class publishers to cutting-edge eDiscovery providers. Content Analyst’s technology exponentially reduces the time needed to discern relevant information from large volumes of documents and data. To learn more about how Content Analyst helps organizations improve productivity and speed the discovery, analysis, and delivery of information, visit their website at [www.contentanalyst.com](http://www.contentanalyst.com), or email [info@contentanalyst.com](mailto:info@contentanalyst.com).

---

## About Fios

For over a decade, Fios has helped corporations and their outside counsel reduce risk, control costs and gain management control over the entire spectrum of e-discovery. We are dedicated exclusively to delivering comprehensive services and expert guidance that transform the burdensome nature of electronic discovery into a streamlined, legally defensible business process. Our proven services and methodologies are based on an integrated, in-depth knowledge of technology, legal and human resource requirements to meet the ever-changing demands of complex e-discovery. Headquartered in Portland, Ore., Fios is backed by top investors, including 3i, Digital Partners, Banyan Capital Partners, FBR CoMotion and Fluke Venture Partners, and has offices throughout the continental United States. For more information about the company and its services, visit <http://www.fiosinc.com>.