



Unique Features of Content Analyst™ for the Legal Market

Electronically Stored Information (ESI) has only recently been included in the Federal Rules of Civil Procedure (FRCP), while corporate businesses have been quietly converting to a digital world for years. The challenge that ESI presents to the legal world is manifold: a potentially daunting store of discoverable information, a plethora of file types, metadata which itself can be considered discoverable, and finally content that is *dynamic*. In a world where facts are facts and what's done is done, content that continually changes is a vexing problem.

Content Analyst's technology, using Latent Semantic Indexing (LSI), enables our legal partners to create products that address these challenges in new and more economical ways. Competing solutions cannot match the straightforward – *and cost effective* – manner in which Content Analyst deals with many major content challenges facing the legal market.

The Accuracy or Defensibility of LSI

Content Analyst's software uses a patented and proven technology called Latent Semantic Indexing (LSI) that along with numerous other patented software routines yields a mathematical representation of documents that is inherently as accurate as human review.

When conducting any search, documents being returned (or not) will fall into one of three categories: relevant, non-relevant, and documents missed (not returned but relevant). The measurement of accuracy that any solution provides is the combination of Precision and Recall:

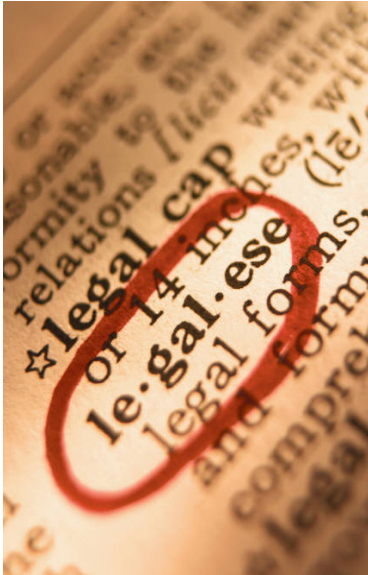
Precision: The precision rate is the percentage of identified documents that were relevant. For example 100 documents were identified, but only 80 of them were relevant. So precision is 80%.

Recall: The recall rate is the proportion of relevant documents identified, compared to the total number of relevant documents that exist. For example the total number of relevant documents is 100, but only 75 were identified. Your recall is 75%.

Content Analyst has demonstrated Precision and Recall rates of 89% and 88.3% respectively - a magnitude of performance better than other techniques.



There are Ten Significant Benefits provided by Content Analyst for the Legal Market:



1. Prevention of Inadvertent Disclosure of Privileged Information – *using machine learning to automatically flag content before it is released*
2. Automated Summarization – *how to figure out the relevance of a document without reading it first*
3. Automatic Categorization – *adding in the structure needed for ESI that most corporations lack*
4. Auto-Taxonomy Generation – *“jump-starting” the process of identifying a category structure where none exists*
5. Multi-Lingual and Cross-Lingual Text Analysis – *identifying relevant content regardless of the language in which it’s written.*
6. Addressing Synonymy and Polysemy – *synonymy refers to the common practice of having many names for the same object (think car=auto=automobile) while polysemy refers to identical words whose meaning is wholly context-dependent (for example, is “spam” a food product or junk mail? – or is a bank where you keep money or the side of a river?)*
7. Overcoming Obfuscation – *how to overcome deliberate attempts to hide information from discovery.*
8. Sentiment Determination – *how to identify favorable versus hostile witnesses before they are deposed*
9. Domain Agnostic Analysis – *doesn’t depend upon extensive dictionaries and term lists that need constant refinement.*
10. Contextual Explanation and Relationship Discovery – *at your fingertips, instant explanation for unknown terms as well as their relationship to information you may already know.*

Prevention of Inadvertent Disclosure of Privileged Information

Why it happens: Unless an entire document is read by a knowledgeable associate, privileged information may be inadvertently disclosed. Of greater concern is the disclosure of *partly* privileged information, i.e. a document that by itself is privileged but meaningless – triggering the necessary disclosure of additional privileged documents to explain the first one (and often releasing even more information).

How can it be prevented: A deep contextual search performed by Content Analyst not only finds the privileged information in its original form, *it finds and identifies derivations of that privileged information*, since both are related by content and context. A “sweep” through Content Analyst’s engine catches documents that might inadvertently be disclosed.

Why this is valuable: Despite recent FRCP regulations regarding “sneak-a-peek” and “clawback” provisions to address the growing problem of inadvertent disclosure, preventing the problem in the first place saves the time (and cost) of having to create new motions, review disclosed information, and meet with opposing council to resolve the issue.

Automatic Summation

How this works: Since Content Analyst “reads” an entire document as part of its analysis routines, our software includes an additional module called *automatic summation*. This feature automatically pulls-out sentences within the document that best represent key concepts within that document.

Why it is useful: Often only a small section of a large document is relevant – other summation techniques relying on key words or phrases totally miss the *context* of those terms and have limited use. By summarizing based on *only* those items that were both most relevant to the case *and* most significant in the document, users gain a quick summary of the entire document.

Why it is valuable: It’s a simple math exercise: every document that has to be read and extracted by a knowledge worker is hours of time and money – automating this process can save \$000’s *for every case in which it is used*.

Automatic Categorization

What it is: Content Analyst allows users to define categories by means of examples. Based on these examples (called *exemplars*), Content Analyst can automatically assign “responsive” vs. “non-responsive,” or “client-privileged” materials to the appropriate categories.

Why it is useful: Most companies don’t employ a content management scheme that is useful for ESI. Worse, legal companies who develop their own thesauri or taxonomies often find they don’t “fit” the content of particular clients.

Why it is valuable: Categorizing large volumes of discovered information by hand adds many billable hours of time to any case; automatic categorization adds virtually none since it occurs during the discovery process.

Automatic Taxonomy Generation

What it is: Content Analyst also has the ability to actually create a taxonomy from the material it is analyzing.

Why it is important: A typical taxonomy automatically generated from random unstructured text can eliminate 60% or more of the work required to create the same taxonomy manually.

Why it is valuable: Creating a taxonomy from scratch can take several weeks; modifying an existing one may not match the data being analyzed; turning weeks into days has a bottom-line impact on the cost of litigation.

Multi-Lingual and Cross-Lingual Text Analysis

What it is: Content Analyst’s cross-lingual support lets users submit queries in English while searching documents in 17 major languages. Multi-lingual support exists for any language that can be represented in the Unicode® encoding system.

Why it is important: Not all documents may be in English – if there are other languages mixed in, information discovery becomes a huge challenge – unless you’re using Content Analyst.

Why it is valuable: Without multi-lingual capabilities, the only other way to correlate data in different languages is human translation (upwards of \$5/page) since machine translation is inherently unreliable for these purposes.



Addressing Synonymy and Polysemy

What it is: Content Analyst automatically identifies synonyms and polysemic usage of terminology.

Why it is important: Synonyms can cause 20% or more of relevant information to be missed. Polysemy requires numerous key phrases and dictionaries and thesauri in other systems – and will always miss some terms.

Why it is valuable: On average, these two word challenges have a direct impact on undiscovered information (at least 22%) or will require human intervention with a cost of \$000's in additional hours.

Overcoming Obfuscation

What it is: Content Analyst reads content and context, so attempts to “hide” information by using confusing terminology or omitting key terms is of no consequence.

Why it is important: A practice once limited to patents, obfuscation has crept into general practice when people want to disguise or hide content, sometimes innocently, but with the same end result of preventing its discovery.

Why it is valuable: Systems that use phrase dictionaries do a poor job at uncovering obfuscation; that requires manual intervention at a cost of many \$000's – and is vulnerable to FRCP Section 26 cost challenges as well.

Sentiment Determination

What it is: The way in which information is conveyed in unstructured text is generally narrative. Unlike structured information, narration includes many clues as to the sentiment of the writer. Content Analyst is a *learning* system: it can be trained to recognize those clues and ascertain the sentiment – positive or negative – of the writer.

Why it is Useful: Evidence ultimately comes down to the individuals responsible for that evidence. In legal proceedings, the notions of positive and negative usually translate into favorable versus hostile witnesses and evidence.

Why it is valuable: Evidence produced by a favorable individual is much more valuable to a case than the same evidence connected to a hostile individual. If this individual becomes a witness, the outcome of the case may rest on this knowledge.

Domain Agnostic Analysis

What it is: Content Analyst reads the content and context of the unstructured text it is analyzing in its entirety. That content as well as the *context in which it is used* becomes the basis for further understanding and analysis.

Why it is important: Other solutions rely on precisely-defined term lists and phrase dictionaries, often created for specific industries or even specific legal cases. The issue quickly becomes lack of domain-specific words or phrases in these lists and dictionaries.

Why it is valuable: Systems that use phrase dictionaries and term lists require constant refinements during the query process. In addition, *multiple* dictionaries are required for complicated cases and multi-lingual requirements. Not only do these lists take time to develop (or additional \$\$ to license), they miss up to 20% of relevant content.

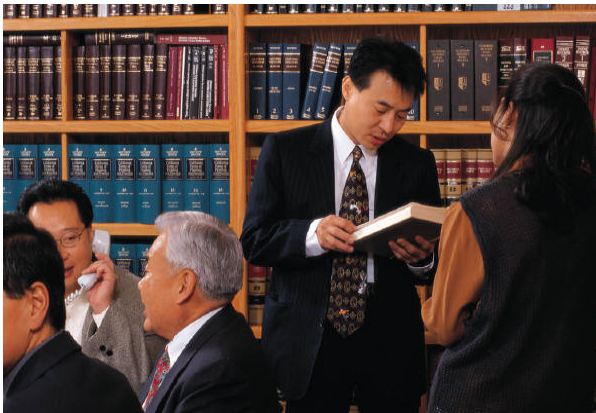


Contextual Explanation and Relationship Discovery

What it is: Content Analyst has a unique pop-up feature that aids in dealing with unfamiliar terminology or uncovering non-obvious relationships between entities. By clicking on a term within a given document, a pop-up displays the terms or entities that are closely related. This provides the user with a good understanding of the general term meaning, *as well as* possible relationships between two entities.

Why it is important: Every industry has its own set of acronyms and technical jargon. The learning curve for any reviewer can be steep if not already a subject matter expert in that space. Hidden relationships between entities could provide valuable leads.

Why it is Valuable: Think of it as a force multiplier. You are reducing the time proficiency surrounding industry specific lingo. The potential to discover hidden relationships is invaluable as the case is developed.



Content Analyst's technology has been designed to integrate with standard software products - programmers appreciate our easy-to-use APIs and Java-based adapters. Our software uses English language commands, handles complicated file types like XML, HTML, and PDF with ease, and works well in conjunction with other search and content management solutions.

Isn't it time you found out how Content Analyst can help make *your* legal software solutions more compelling and cost effective for your customers?

Learn More: Call 888.349.9442/703.391.8700 or Email info@contentanalyst.com



Stop searching. Start doing.™