



# What is Discovery Analytics?

An In-depth Perspective on Analytical Search Techniques and Their Application in the eDiscovery Workflow

By Nicholas Croce



### Introduction

As EDD market spending is projected to exceed \$40 Billion by 2011<sup>1</sup>, there is increasing corporate pressure to lower costs while continuing to manage the risks involved in the discovery process. To address this imperative for greater efficiency, the electronic discovery market is undergoing a steady shift toward workflow automation and the application of more sophisticated software tools.

With a diverse and cluttered vendor landscape offering a myriad of “solutions,” one of the most difficult challenges is the evaluation of technologies where diverse terminology and claimed capabilities create “apple and orange” comparisons among solution providers and their product offerings. **Discovery Analytics** is the latest category of software that law firms and corporations have begun to use to help alleviate the continually rising costs of document review. Unfortunately, there is confusion about the definition of analytics and how the newest eDiscovery technologies incorporate different types of analytic capabilities into their systems. This white paper explores the various structures and implementations of Discovery Analytics as it pertains to document review, as well as the many definitions to help purchasers make educated decisions regarding the solution that best fits their needs.

### Acceptance of Search Analytics

Within the last few years, the term “analytics” has become a buzzword in the eDiscovery market, and for good reason. With burgeoning data growth and the need to vastly increase efficiency in the discovery process, both the courts and the vendor community have addressed the challenge with new tools. The terms “analysis” and “analytics” are now being used interchangeably to describe functions ranging from base reporting and review metrics to sophisticated search software and advanced data mining applications.

Over the last year, analytic search technologies have come to the forefront as a set of tools that can add efficiency and reliability to the discovery process, and ensure that the process is defensible and repeatable. This change is evidenced by the recent passage of FRE 502<sup>2</sup>, which states that, “a party that uses advanced analytical software applications and linguistics in screening for privilege and work product may be found to have taken ‘reasonable steps’ to prevent inadvertent disclosure of protected communications or information<sup>3</sup>.” In addition, for the first time, the use of traditional Boolean searching has come into question. Influential and impartial groups like The Sedona Conference, TREC and the *Richmond Journal of Law and Technology* have published opinions demonstrating the weaknesses of

---

<sup>1</sup> Based on Forrester Research’s “Believe It – eDiscovery Technology Spending to Top \$4.8 Billion by 2011,” written by Barry Murphy and published December 11, 2006, and industry expert Michael Clark of EDDix projects a 7-8x multiple for attorney review spending, the combined EDD costs total more than US \$40B by 2011.

<sup>2</sup> Signed into law by President George W. Bush, September 19, 2008, Federal Rule of Evidence 502, or S. 2450, protects against the inadvertent waiver of privileged information during a federal proceeding.

<sup>3</sup> Advisory board comments to Rule 502(b) highlighting that the application of analytical and linguistic tools has become a vital part of the litigation process.

Boolean search and suggest new approaches on how to best utilize analytics for more accurate and effective review. Furthermore, individual cases have concluded similarly—even citing independent research: Magistrate Judge John M. Facciola wrote in a recent opinion, “I bring to the parties’ attention recent scholarship that argues that concept searching as opposed to keyword searching, is more efficient and likely to produce the most comprehensive results<sup>4</sup>.”

The current shift in the electronic discovery market is similar to when legal research went online in the 1990s. During this time, as more legal research tools were introduced to the market, lawyers remained convinced that they could still do a better job using their law libraries. Once the use of online databases and search tools became pervasive and the courts recognized their efficiency, the market shifted and online search became the standard.

Like the change in legal research, mitigating risk is the key issue for attorneys, and because an attorney’s main responsibility is to reduce client risk, how will they be able to become comfortable with concept search technologies that are new and often complex? Can attorneys use analytic search safely today? These questions may continue to be asked for a while longer, until additional laws and judicial opinions are published that accept analytics as “reasonable” and the use of these tools is implemented and defined by a broad spectrum of corporations and law firms.

### **Analytics Defined—Structure of Tools and the Approach to Concept Creation**

To understand conceptual search technology and its applications within the electronic discovery workflow, it is important to review the building blocks of the different types of software available in the market. There are two primary methods utilized to identify concepts and build analytic searches on discovery data: those that are based on a predefined set of terms and those that create an index of concepts from uploaded data.

- **Predefined Terms: Ontological/Linguistic Search.** This method is essentially a comparison against a dictionary or listing of terms or synonyms, often referred to as ontologies. In these systems, searches containing keywords are submitted to the ontology and the system automatically searches for all related terms. For example, should you be looking for “shredded documents” the system will search for each term individually and include the terms “shredded” and “documents,” as well as any synonyms relating to the terms in the initial query. The related terms that could be returned may include “rip,” “tear,” and “destroy,” as well as “paper,” “letters,” and “memos” including all permutations. Therefore, a user looking for “shredded documents” could also find “destroyed letters” and “torn memos.” This functionality can be beneficial when searching for standard words within a single language universe by automatically creating an enhanced Boolean search. However, this technique is less valuable in instances when searching for case-specific information is required.

For example, in the Enron litigation, there was a shell company known as Raptor. Using an ontology-based system, the term “bird” would be substituted and a non-responsive or irrelevant set of documents would be returned, which is not only more ineffective, but will return incorrect results. A hybrid of this method creates a set dictionary of terms as documents are loaded and simply utilizes the number of occurrences of a word to determine relevance: the more times a word is used, the more it must be relevant. This type of method has similar limitations as the ontology-based systems.

---

<sup>4</sup> In his decision, Magistrate Judge Facciola is directly referencing George L. Paul and Jason R. Baron’s research, “Information Inflation: Can the Legal System Adapt?” Rich. J. L. & Tech, 10 (2007). Disability Rights Council of Greater Washington, et al., Plaintiffs, v. Washington Metropolitan Transit Authority, et al., Defendants. Civil Action No. 04-498 (HHK/JMF). United States District of Columbia, 2007 U.S. Dist. Lexis 39605. June 1, 2007 Decided.

- **Index Terms: Mathematical Index Search.** This method builds an index of concepts based on the actual data being reviewed for a matter. Using this technique, as the data is being ingested/uploaded, mathematic algorithms that identify unique patterns of words based on definition, frequency, contextual placement, and a combination of terms are applied. These systems essentially treat each word as a number and each pattern of words as a unique series of numbers, and can therefore perform advanced statistical analyses on these numeric patterns to identify trends. These numeric patterns, or trends, are then translated into “concepts” based on the actual data.

In this more advanced method, the system can track each time the term Raptor is mentioned, but also in what context, based on the algorithms initially applied. Therefore, other terms commonly referred to in relation to the term, such as subsidiary, divestiture or financials, will be identified in conjunction with the initial search. The system will link the term Raptor to other similarly situated terms and extract concepts such as sale transaction or asset transfer, finding crucial, responsive documents.

Although ontology-based search seems to make it easier for users to find the documents they intended to, it can also produce overly inclusive results, as well as simply incorrect documents. In contrast, using mathematically-based concept search may require the user to manually include additional terms (i.e. perform the original search to include shred, rip, tear and destroy), but would return more responsive and conceptually similar results.

It is important to note that there is no single algorithm that works best in every case. The ultimate solution combines both ontological and mathematical logic structures, providing users with multiple methods of application. For example, a document population that consists only of OCR<sup>5</sup> processed data will be rendered completely ineffective in an ontology based system. The converse is true for attorneys who intend to batch documents for strict linear review, where they ultimately want to use keyword and synonyms to batch data.

---

<sup>5</sup> Optical character recognition, or OCR, is the mechanical or electronic translation of images into machine-editable text.

### Preprocessed vs. Dynamic Search

The second building block of an analytics system is how concepts are applied to the user workflow. The goal of concept searching is to extract the meaning of words, helping users search for what they are truly looking for, rather than finding responsive documents because they luckily happen to hit on a specific keyword. Search based on conceptual meaning can be a significant benefit for early risk assessment, enabling attorneys to focus their attention on critical documents instead of irrelevant information. Also, concept search organizes documents for more efficient review by grouping like documents, providing substantial review process acceleration—a factor of 3 to 10 times faster than standard linear review. By dramatically speeding review, the end client saves substantially on the largest EDD cost: attorney labor. While attorneys can get to important documents more quickly thus reducing client risk, and although these systems provide concepts to users in multiple layouts and fashions, there is one major difference between the systems available today—preprocessed versus dynamic concepts.

- **Preprocessed Concepts.** Systems that automatically generate a list of concepts at the inception of a project allow users to only select among that fixed set. The majority of systems using this methodology auto-categorize documents by concept, grouping documents into folders or clusters. This enables reviewers to look at documents that discuss the same concepts together, resulting in an increase in review speed of at least 3x to 10x, as humans can much more easily review documents that discuss similar concepts<sup>6</sup>. The time efficiency and cost savings to this approach are dramatic and very useful within a workflow that essentially performs a linear review. Due to technical limitations in their deployment, these systems normally cluster or “auto-folder” a small set of documents together (5,000 to 10,000 documents). Although this still improves the speed for an individual reviewer, it does not provide attorneys with any ability to look at conceptually similar documents across the entire dataset, limiting the opportunity for in-depth analysis and investigation.

- **Dynamic Concepts.** Systems that calculate the concepts based on a specific document population or subset of the dataset dynamically refine the concepts in real time, allowing users to target review based on specific issues. This methodology provides users with the ability to investigate issues, explore trends, and easily search large datasets, while significantly accelerating the review process. In this user-driven system, attorneys can view concepts within a specific user’s emails, or within a specific date range, or even within a keyword results set.

Attorneys can use these concept groupings to help develop case issues and investigate risk by asking intelligent questions. Most user-driven systems also provide the ability to review the concepts across the entire population of data, so that all conceptually similar documents in the repository can be grouped together.

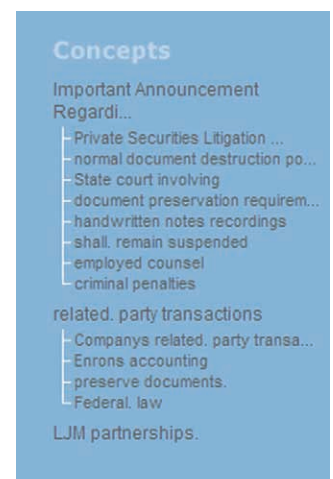


Figure 1 – Inference Dynamic Concepts

<sup>6</sup> Voorhees, E.M. (1985) “The Cluster Hypothesis Revisited.” Pro. of ACM SIGIR, June 1985, ACM Press, New York, US 188-196.

Although both of these approaches provide significant advances beyond standard Boolean keyword search, the ability to dynamically search and review concepts within a set or subset of data provides dramatic benefits to the user as it follows an attorney's specific thought process. Counsel must constantly explore different paths to help identify key documents, having the ability to search concepts that will direct them to the specific area of importance very quickly.

### **Analytics Workflow Presentation**

For both the general public and the legal profession, technology has reached the next stage in interface design, providing enhanced and more efficient methods of interacting with complex information. With millions of documents to be reviewed, using concept searching can initially seem overwhelming for attorneys, which is where presentation and visualization can dramatically help.

Here are two types of analytic workflow presentations (Figures 2 and 3). The first is a "cluster map" or "concept clustering" which helps organize the concept groupings previously discussed. The cluster visualization conveys significant information to the user. Each red square indicates a concept, whereas the proximity of the red squares indicates their relationship to each other. Therefore, the closer they are located on the map together, the more closely related they are in concept. This critical functionality helps users review documents with dramatic efficiency by looking at similar documents in conceptually organized batches. It also helps separate those documents that are irrelevant to the investigation.

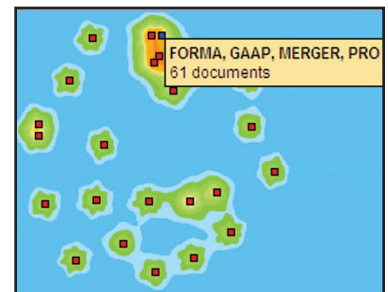


Figure 2 – Inference Cluster Map

Another effective visualization is an "Email Trace." This interface provides the user with an instant view of an email thread communication to determine to whom emails were sent and at what time. The emails are color coded to indicate the original email, replies and forwarded emails to allow the user to quickly navigate the online conversation and to zoom in on relevant communications. At any point, a user can click on the respective line to see the actual email.



Figure 3 – Inference Email Trace

### Conceptual Analytics—Today’s Implementation and Tomorrow’s Standard Use

Concept analytics are quickly becoming a core part of the electronic discovery toolkit and have the strong potential to become the new standard in search technology and legal review workflow. As this rapid evolution takes hold, it is important that the litigation community uses these technologies to maximize benefits and apply them to the appropriate cases where they can be most effective and efficient. Key considerations for the types of technology include:

- **How are the concepts created?** Concept creation may be based on a variety of methodologies, and it is important to understand how the concepts are created to provide a comfort level for attorneys and give them the information needed to design a proper review workflow. “Black box” technologies are dangerous when considering vendors and usually demonstrate an unwillingness (or inability) to explain the technology, therefore limiting defensibility and increasing risk. By simply asking for details on the underlying algorithms used, it will reveal significant information on their reliability and credibility.
- **Do I want a user-driven system that lets me investigate without predetermined foldering, or should I be using auto-categorization?** If you are looking for a streamlined method of linear review, auto-categorization can be helpful in organizing and speeding review. However, if determining risk and prioritizing your review, while also streamlining and accelerating, dynamically created concepts will help to enable a more organized and effective method of review. Moreover, if you plan on any strategic issue coding, deposition preparation or investigation during trial you should focus on a user-driven system.
- **Are baseline review tools integrated?** If you cannot redact, print or produce the data from the same system you are reviewing in, you will ultimately need to switch systems mid-stream. Be sure that you can keep your data in one system until you have completed your objectives for using analytics, as well as your case goals.

The courts and the legal community are in the process of accepting concept analytics as a preferred method of searching and reviewing EDiscovery information. Discovery analytics is certain to increasingly improve review workflow and to help end clients to reduce costs. Taking your first (or next) steps in using these tools is an important advancement for your legal team in this rapidly changing environment.

### About the Author

Nicholas Croce, President of Inference Data, has led the creation and development of Inference, the company’s next-generation analytics software for electronic discovery. Prior to joining Inference, Nick was the president of DOAR Litigation Consulting, a leader in electronic discovery and courtroom technology.

### About Inference Data

Inference Data provides corporations and law firms with next-generation software for analytics-driven assessment, meet-and-confer preparation, and accelerated legal review. The company’s product, Inference, applies conceptual search and analysis to prioritize datasets and find key documents for faster and more cost-effective review. Inference, a web-based solution, is distinguished by its highly scalable Autonomy architecture and user-intuitive, configurable workflow. Inference Data is privately held and headquartered in New York, NY. For more information visit [www.inferencedata.com](http://www.inferencedata.com) or call 877.534.5504.