**Are Lawyers Being Replaced by Artificial Intelligence?**

**Moving Beyond Keyword Search: An Introduction to Advanced Search & Retrieval Technologies**

**By Sonya L. Sigler**[1]

Keyword search is quickly losing favor as a method for culling data to a manageable size or as a mechanism to produce relevant data. That message is coming through unmistakably loud and clear from many courts that are opining about the shortcomings of keyword search. Lawyers trying the cases, however, have not necessarily heard or understood that same message. For example, see the opening pronouncement of the Gross Construction Order: [2]

> "This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or "keywords" to be used to produce emails or other electronically stored information ("ESI"). While this message has appeared in several cases from outside this Circuit, it appears that the message has not reached many members of our Bar."

Who would have thought lawyers would need to become search and retrieval experts? The Gross Construction case happened in New York, which is home to some of the best and brightest lawyers in the world; specifically, the Southern District of New York, which is the same district that spawned the five Zubulake opinions beginning in 2003.[3] This district is not a district of lawyers who are unfamiliar with finding the right electronic data for their cases or who practice in front of judges uneducated on the topic of electronic discovery. Yet we have a scathing order telling lawyers to get a clue about the shortcomings of keyword search and its ineffectiveness in retrieving the right information for review and production.

There are so many search and retrieval technologies being applied in the litigation context today that it is difficult to keep up with them all. It is even more difficult to differentiate the pros and cons of using these different technologies. An even more vexing question on the minds of many lawyers is whether these search and retrieval technologies are replacing lawyers.

This paper explores how lawyers can move beyond keyword search and provides an

---

[1] Sonya Sigler is the Vice President, Operations & General Counsel at Cataphora, Inc. She is a frequent speaker on advanced search and retrieval techniques and other eDiscovery topics. If you wish to contact Ms. Sigler, she can be reached at sonya.sigler@cataphora.com.

[2] *William A. Gross Const. Associates, Inc. v. American Mfrs. Mut. Ins. Co.*, _F.R.D._, 2009 WL 724954 (S.D.N.Y. March 19, 2009).

[3] Judge Scheindlin wrote five Zubulake opinions beginning in 2003 and she sits in the SNDY, with Judge Andrew Peck of the Gross Construction Order. Judge Scheindlin recently published two new textbooks co-written with Professor Daniel Capra and The Sedona Conference, entitled *Electronic Discovery and Digital Evidence: Cases and Materials* (West 2009), and *Supplementary Materials on Electronic Discovery: For Use in Civil Procedure Courses*.

introduction to advanced search and retrieval[4] approaches using linguistic methods, statistical methods, and pattern analysis. This paper also explores whether these search and retrieval technologies are replacing lawyers or merely supplementing their work and work product.

**Overview—Search & Retrieval Methodologies**

*Linguistic methods* focus on the use of language to retrieve items that contain specified words or patterns of words. Two examples of linguistic search and retrieval methods are keyword search and ontology-based search. *Statistical methods*, not surprisingly, use statistics and probabilities in order to group similar documents together. Statistical search and retrieval methods include clustering, latent semantic indexing, and Bayesian classification. *Pattern Analysis* models a data set in part by associating with it, a baseline of behavior and then measures particular behavior against that baseline to detect anomalies.

Keyword search is easy to use and easy to understand, especially given lawyers' widespread use of Lexis Nexis and Westlaw and other commercial search tools such as Google. Using keyword search, one types in a word or two, maybe uses wildcards (*), Boolean operators (AND, OR, NOT), or proximity indicators (within 20 words of this word), and then almost instantaneously receives a list of search results that may even be ranked or prioritized, depending on the search engine used; then the laborious process of wading through these search results begins in order to see if anything relevant or useful has been retrieved. With a quick scan of the results, it is readily apparent that using keyword search alone will return results that are over-inclusive or under-inclusive.[5] For example, a search for "hous*" will return documents containing *house, household, housemate, housing, Houston*, etc. Searching for names is even more challenging–a search for *Ted* will return documents containing *Ted*, but may miss documents containing *Theodore*. Depending on the search parameters in force, it may also return documents containing words like *united* or *drafted*. Crafting an effective query (as keyword searches are often called) is difficult and time consuming. Courts are recognizing that keyword search is not the be all *and* end all of finding the right information and are cautioning against using keyword search alone or without appropriate search and retrieval expertise.[6]

Ontologies are another linguistic method that focuses on the use of language in a dataset. Corporations usually have their own way of referring to things–products, people, groups, etc., and that language must be deciphered to effectively retrieve relevant information. Ontologies are like assembling a super query to retrieve information related to a concept or topic. A simple example is the concept of *aircraft*. In

[4] Search and retrieval as used in this paper also refers to and encompasses the categorization and classification of data as well as the searching and retrieval of data. These two concepts are used interchangeably when "search and retrieval" is referenced.

[5] For a more comprehensive discussion on keyword search and other search and retrieval methodologies, see the Sedona Conference's *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (August 2007).*

[6] See Henry v. Quicken Loans, Inc., (E.D. Mich., Feb 15, 2008), *United States v. O'Keefe*, No. 06-249 (D.D.C. Feb. 18, 2008), *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 2008 WL 2221841 (D. Md., May 29, 2008) and In re Fannie Mae Securities Litigation, _ F.3d _, 2009 WL 215282009, U.S. App. LEXIS 9 (D.C. App. Jan. 6, 2009).

this case, an ontology-based query may retrieve documents that contain any of these words rather than just the word *aircraft*: *airplane, plane, Boeing, 747, Cessna, Glider*, etc. This method is sometimes referred to as concept searching.

Statistical methods rely on counting words in a document and on establishing probabilities that having one word in the same document as another word means that it is more likely to be similar to another document containing these two words than to a document that does not contain this same word combination. Using statistical methods involves indexing a data set and counting the number of times a word appears in the data set and within individual items in the data set. For example, if the word *diamond* appears in the same document as *ball* or *base* it is more likely to be similar to another document that contains those same words than a document that contains the word *diamond* and *pendant*, but not *ball* or *base*. Bayesian Classification is a system that assigns probabilities to a document in terms of how likely it is to be related to another document with the same words in it. Latent Semantic Indexing involves extracting multiple concepts from data sets through a statistical semantic analysis of each file. The theory is that unstructured files comprise *latent* concepts that are not readily recognized and remain hidden until a more precise lexicon is developed out of the whole collection. These methodologies are often referred to as clustering or statistical clustering.

Pattern analysis is a more advanced search and retrieval or classification method. It involves modeling the entire set of data to establish a baseline of behavior for individuals or groups present in the data. Once this baseline of behavior is determined, anomaly detection (showing good or bad intent) can be done, showing communication patterns, deletion patterns, changes in behavior, centers of power, social networks, decision-making patterns, and any number of other things. Any of these types of pattern analysis can be used to quickly help locate information or illuminate patterns of behavior that can be invaluable in litigation or investigations.

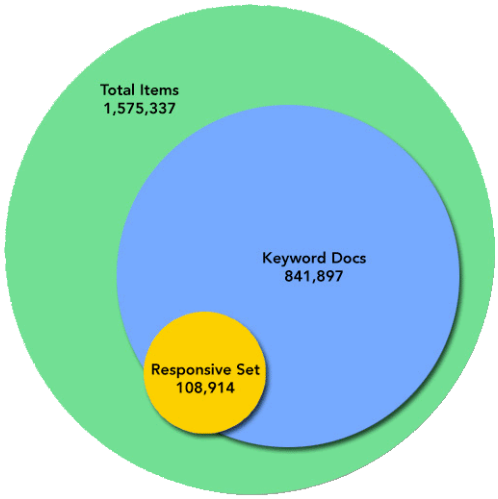
**Keyword Search Shortcomings**

Although keyword search is easy to use and lawyers are familiar and comfortable using it in tools such as Westlaw and Lexis Nexis, keyword search results can be over-inclusive and under-inclusive. This means that keyword searches retrieve too many documents that are irrelevant (false positives) and, at the same time, fail to return documents that are relevant (false negatives). Take the keyword searches done in the case *In re Fannie Mae Securities Litigation*[7] where the OFHEO lawyers agreed to use over 400 search terms and in performing those searches (or queries), retrieved 80% of the entire data set. This was a mere 20% reduction of the data set, which was not effective in trying to reduce a review set to a manageable amount of data nor was it helpful in pinpointing the most relevant data.

[7] *In re Fannie Mae Securities Litigation,* _ F.3d _, 2009 WL 215282009, U.S. App. LEXIS 9 (D.C. App. Jan. 6, 2009)

In Gross Construction,[8] Judge Peck makes the obvious point that using thousands of search terms for construction is unlikely to reduce the data set. In fact he opines:

> "This case is just the latest example of lawyers designing keyword searches in the dark, by the seat of the pants, without adequate (indeed, here, apparently without any) discussion with those who wrote the emails. Prior decisions from Magistrate Judges in the Baltimore-Washington Beltway have warned counsel of this problem, but the message has not gotten through to the Bar in this District."




As a further illustration see the Venn diagram of data from an actual review set of data. In this review data set of 1,575,337 items, consisting of emails and their attachments as well as loose Microsoft Office files (the green circle), keyword searches were run over the data set and retrieved 841,897 items (the blue circle) or 53% of the entire data set.[9] Sometimes this is referred to as the key-term positive set.

Of those 841,897 key-term positive items, 88% were irrelevant, meaning the keyword searches run were over-inclusive to the tune of 741,536 items. The keyword searches retrieved 92% of the responsive documents but missed 8% of the responsive documents, meaning the searches were under-inclusive and missed 8,853 responsive items. This particular number of missed items does not seem like a lot unless the items that the case team needed to make their case were in that missing set (the portion of the yellow circle outside the blue circle).

Of course, different data sets and different keyword searches may produce distinctive percentages and results, but the end result of running keyword searches over any data set will be similar to the example used here, resulting in under-inclusiveness (missing relevant documents) and over-inclusiveness (retrieving too many irrelevant documents).[10]

---

[8] *William A. Gross Const. Associates, Inc. v. American Mfrs. Mut. Ins. Co.*, _F.R.D._, 2009 WL 724954 (S.D.N.Y. March 19, 2009).

[9] In the In Re Fannie Mae Securities Litigation case, this blue circle would have been 80% of the size of the green circle. In Gross Construction, the 1,000 search terms would have made the blue circle the same size as the green circle (100%). See footnotes 7 and 8 for these case cites.

[10] The technical terms for these search results are Precision and Recall. Precision is the measurement of correctness: the ratio of Responsive Items Retrieved to All Items Retrieved shows how "clean" the retrieved document set is. Recall is a measurement of the completeness of a search: the proportion of Responsive Items Retrieved to All Responsive Items. In the Venn Diagram example, the keyword searches retrieved 92% of the responsive set (100,361 of the 108,914 Responsive Items), represented by the portion of the yellow circle (Responsive) overlapping the blue circle (Retrieved); The precision of the keyword searches, represented by the

Keyword search is often completely ineffective in identifying relevant documents in the following situations:

- When the words are not used (who actually uses the word fraud when they commit fraud?)
- Foreign language is used (although keywords can be run in other languages, they often are not considered)
- Acronyms (ESI, short for Electronically Stored Information)
- Short Messages or Instant Messages (don't contain many words)
- Misspellings (Priviledged, priveleged, etc.)
- Cryptic language is used (i.e. slang, obscure personal references, or shorthand)
- IM language–*BFF*, *NSL* (Best friends forever; Name, sex, location)
- Nicknames or location names (the Barnes place, Midtown)

The goal of using any search and retrieval technology is to find the relevant data set (the yellow circle) in the most effective and efficient manner. Overcoming the shortcomings of keyword search requires the use of other search and retrieval methodologies. Supplementing keyword search with other information retrieval methodologies or using these other methodologies *instead* of keyword search will lead to a smaller universe of documents to review (i.e. shrinking the blue circle) (correctness) *and* will result in more of the relevant documents being retrieved on the first attempt (having more of the yellow circle overlapping with the blue circle) (completeness).

**Moving Beyond Keyword Search**

What, other than keyword search, can be used to find the relevant or responsive data in an efficient and effective manner? Employing statistical and other linguistic methods can be much more effective than using keyword search in isolation. These other methodologies can be used alone, together, or in conjunction with keyword search. In addition, these methodologies can be used to categorize data sets to focus a review or production. In general, categorization is the grouping of objects, people or ideas on the basis of some kind of "similarity."

As applied to electronic discovery, categorization describes the grouping of documents according to some desired criteria. Categorization may be done by topic or by legal criteria. Categorizing documents about specific products, or documents that relate to sales in a given country are examples of categorizing by topic. Documents can also be categorized by foreign language by assigning a primary language to a particular item (especially if it contains more than one language) and grouping each language set together. Legal criteria might include categorizing responsive, non-responsive, and privileged documents.

---

relative size of the yellow circle overlapping the blue circle (Responsive Items Retrieved) to the blue circle as a whole (All Items Retrieved) was only 12% (100,361 Responsive items in 841,897 documents retrieved). So, the keyword searches were effective (92% of the "right" or responsive documents were returned) but not efficient (only 12% of the entire retrieved set was responsive). In using any search and retrieval methodology there is always a trade off between precision and recall. Achieving 100% on either variable while maintaining reasonable results for the other is usually time and cost prohibitive. An ideal is to strike a balance between the two measures.

Taxonomy is the practice and science of classification. Taxonomies, which are composed of *taxonomic units* known as *taxa* (singular *taxon*), are frequently hierarchical in structure, commonly displaying parent-child relationships. Although taxonomies are a way of classification for data, it is unlikely that this methodology will be used for litigation document review, it is more likely that categorization of data will be more useful for litigation or investigative purposes.

**Statistical Methods**

*Categorization* may be based on statistical analysis of the similarity of documents. For this, a document is mathematically represented by a set of features such as the occurrence of words, or their proximity to other words in the documents. Different weights (levels of importance) may be assigned to the various features. Documents are then deemed to be similar (and therefore belong to the same category), based on the degree to which their features resemble each other.

*Clustering* is the grouping of information by some category or statistical similarity. This is done by comparing various lexical (vocabulary), syntactic (grammatical use), semantic (meaning), and even orthographic (punctuation) features to detect topics rather than just individual keywords. Statistical clustering can be accomplished by counting words and their frequency, and then grouping those documents with similar statistics together in a cluster. When items are determined to be about the same or similar topics, they are clustered together, and usually displayed in some kind of graphical relationship that facilitates reviewing similar documents together.

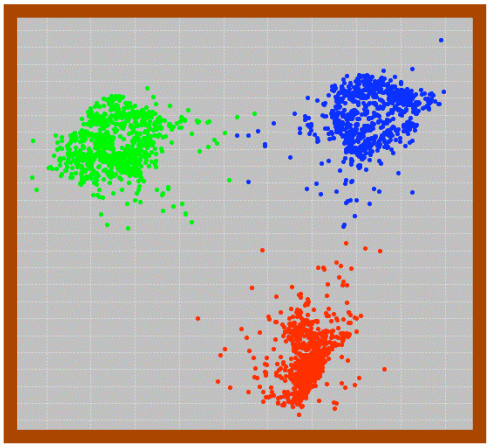
In general, *Bayesian Classification* is based on the statistical probability of a class and the features associated with that class. This type of classification utilizes a training set composed of classes that have correctly assigned features. Once the probabilities of the training set features and classes have been stored, new data is compared against the training set. During this comparison of the "learned" classification of the training set with the new data, the new data's features are calculated and the new data are assigned classes whose probability of matching the training set's classes and features is highest.

*Latent Semantic Indexing* involves extracting multiple concepts from the data collections through a statistical semantic analysis of each file. The theory is that unstructured files comprise *latent* concepts that are not readily recognized and remain hidden until a more precise lexicon is developed out of the whole collection. These concepts then form a dictionary (lexicon) for the collection that can be weighted for both frequency of occurrence and relevance. At that point each file in the collection is compared to the concepts list, and it is assigned a *fingerprint* (or value) that uniquely defines the file according to those criteria. Searches can then be conducted by requesting files that are statistically similar, i.e. that have similar fingerprints, under the presumption that they are statistically similar and conceptually related.

A more in depth look at *Clustering* among these statistical methodologies reveals its appeal to lawyers as a categorization methodology for quicker document review.

Simply put, clustering just means putting documents into groups that have something in common. Clustering can be done manually, which is what humans do during manual document review with issue tagging. Keyword searches can also be used to cluster or group documents together that all contain that same key word or key phrase. Again this is manual tagging or coding. Ontologies (or linguistic filters) can be used to cluster data into groups with those items caught in the linguistic filter. Automated clustering (using technology) can be done many ways: by document type (all the Word documents go into one basket); by creation date; by Actor (person in the data set, not necessarily a custodian); by statistical similarity (statistical clustering); and many other approaches.

Luckily, software can be used to implement statistical methods of finding groups of "similar" documents. The case team can help define what "similar" means so that similarity is defined appropriately for the application. Once this is done, documents can be categorized with very little effort from the user. This use of clustering technology can help immensely with document review. A single reviewer can look at similar documents together, producing consistent review decisions.



Shortcomings of clustering can be its unpredictability as compared to keyword search or taxonomies. Clusters can be as narrow or as wide as the software defines them, sometimes these may be irrelevant clusters for document review purposes. Tight (or narrow) clustering can be used to detect "near duplicates" or textual errors caused by OCR (Optical Character Recognition) programs.

The items that **look** very similar (to the clustering algorithm) may not actually **be** similar in ways that make a difference to a legal team. For example, responsiveness or relevancy may depend upon fine legal distinctions. Responsiveness or relevancy may vary in the same matter by subpoena and/or jurisdiction. These subtleties can make the use of clustering ineffective. Approaches that use attorneys' definition of similarity rather than other properties of the clustered items can make for a more effective use of clustering technology.

A further example of how tight or loose clustering can affect a review is using the word *option*. Does the word *option* make one cluster or four clusters? Is it one cluster around the word *option* or is it four clusters, each containing the word *option* in a different context?

- Financial/energy trading options
- Email/computer menu-driven options
- Stock options (ISO's)
- The generic idea of an available choice of action

Depending on the case, making these distinctions in the tightness of a cluster can make a huge difference to the effectiveness and speed of a document review. If none of these categories are relevant or responsive then a wide cluster containing all four meanings is fine. If stock option backdating is the focus of the investigation or litigation, then building four narrowly defined clusters (that disambiguate among the different senses of the word *option*) around the word *option* will be vital to an effective review.

## Linguistic Techniques

Linguistic search or categorization techniques are based on the analysis of language features of documents, in contrast with *statistical* techniques. Key word search and ontologies are an example of a linguistic technique. Some linguistic search methods are often referred to as concept search. Concept search attempts to find documents that address some concept that a user is interested in. To do so, it goes beyond *keyword* search for documents that contain a specified word or phrase, and tries to find other documents that address the underlying concept. For example, a concept search for *fiber* might return documents that refer to the concept of *fiber* using alternative terms such as *cloth, textile, material, cotton*, etc.

An *ontology* is an arrangement of words, phrases, and search terms under a *concept*. Here is a simple example:

AIRCRAFT CONCEPT

- Boeing
- 747
- Cessna
- Glider

It is a useful idealization to suppose that a document containing one of more of the terms under the AIRCRAFT CONCEPT in fact discusses or deals with the concept of aircraft. This process can be automated, so that a computer does the work. If the computer finds a document that contains one or more of the four terms, it concludes that the document is (at least partially) about aircraft. The document might also discuss other concepts, but a reference to the concept of aircraft is clearly present in the document.

Ontologies are basically super queries that can be crafted and saved for reuse across different matters or investigations. This reuse is especially important for corporations experiencing repeat litigation over the same product or topic. Ontologies can be used to compensate for the shortcomings of keyword search where the word is not there or the word is not the correct word that is used.

In the Gross Construction[11] case, Judge Peck recommends that lawyers do not come up with keywords in a vacuum, but rather, that they craft them 1) using the words in the data, which can be gleaned from building an index of words in the data and 2) asking

[11] *William A. Gross Const. Associates, Inc. v. American Mfrs. Mut. Ins. Co.*, _F.R.D._, 2009 WL 724954 (S.D.N.Y. March 19, 2009).

the people whose data is being searched what words they used to refer to certain things. Both of these recommendations (or admonitions) are common sense:

> "… where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of "false positives." It is time that the Bar–even those lawyers who did not come of age in the computer era–understand this."

The shortcomings of keyword search can be addressed by building an ontology to deal with missing items, misspellings, abbreviations, and other issues. Examples of how this can be done are listed below:

Missing abbreviations, acronyms, clippings:
- *incentive stock option* but not *ISO*
- *Board of Directors* but not *BOD*
- *1998 plan* but not *98 plan*

Missing inflectional variants:
- *grant*

but not
- *grants*
- *granted*
- *granting*

Missing spellings or common misspellings:
- *gray* but not *grey*
- *privileged*

but not
- *privileged*
- *priviledged*
- *privilidged*
- *priveliged*
- *privelidged*
- *priveledged, etc…*

Missing syntactic variants
- board of directors meeting

but not
- meeting of the board of directors
- BOD meeting
- Board meeting
- BOD mtg
- Board mtg
- Directors' meeting
- Director's mtg
- mtg of the BOD
- mtg of the directors
- BOD meetings
- Board meetings
- BOD mtgs
- Board mtgs
- Directors' mtg
- Directors' meetings
- mtgs of the BOD
- mtgs of the directors

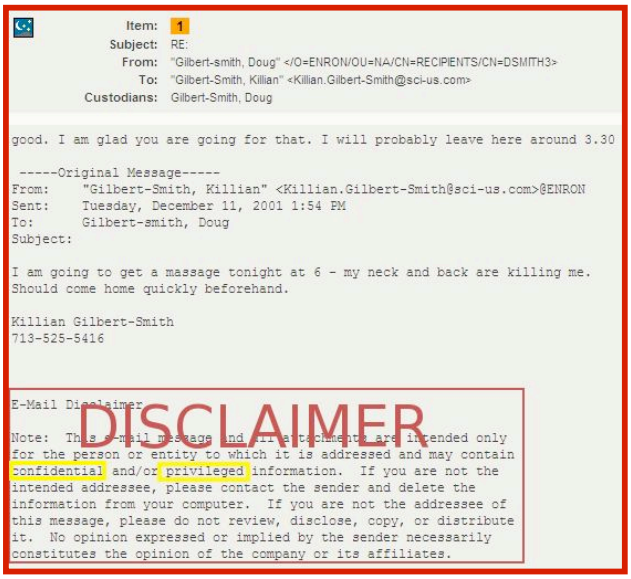
Missing synonyms/paraphrases
- hire date but not start date
- approved by Smith

but not
- Smith' approval
- the approval of Smith
- Smith's OK
- Smith's go-ahead
- Smith's goahead
- the go-ahead from Smith
- the goahead from Smith
- the nod from Smoth
- Smith's signature
- Smith's sign-off
- the sign-off of Smith
- the signoff from Smith

As a keyword item, the address
- *101 E. Bergen Ave., Temple, CA 90200*

does not match any of these:
- *101 East Bergen Avenue*
- *the Bergen site*
- *the Temple location*
- *our 90200 outlet*

Ontologies can be useful in distinguishing between *indicators o*f privileged information in the *email contents and indicators of privileged information or words* found in boiler-plate disclaimers.[12] Disclaimers, which are often found on every email, no matter how trivial, can confound clustering software, but can be addressed using ontologies to detect the presence of disclaimers and automatically exclude the disclaimer from the search results. The use of ontologies can also be helpful in distinguishing words as used in text versus words used in a title or address of a signature block. This use of ontologies is a very cost effective use of search & retrieval technologies up front in any effort to analyze data.




Ontologies are also very effective in dealing with multi-lingual issues, which are omnipresent in all large datasets. Most data sets from companies that have international offices are multi-lingual in a significant way. Many different languages appear in the data set and they are often mixed in the same email or email strings. Ontologies can be built to address product names, colloquialisms, company cultural issues, differences in address or city names, and other issues brought to the forefront because of language differences. As an example, when searching for a particular office where action or behavior took place, it may matter what language was used. It may matter that *Lucern, Lucerne, Luzerne,* and *Lucerna* may are be used to refer to the same office in the same Swiss city. Keyword search will not address this issue unless the search used contains all of those terms, and clustering may not be granular enough to catch these in the same cluster. Ontologies can be used to cover all of these variants.

**Other Methods**

A *neural network* is a computer program whose operation is loosely inspired by the way a human or animal brain works (though the neural network is much, much simpler). A neural network can be "trained" by giving it sample inputs and the correct outputs associated with these. The network can analyze the difference between the answers it is generating and the "correct" answers. It can then automatically adjust its internal workings (weights), until its answers on the training set adequately match the given outputs. The idea is that you can now feed it new inputs (the answers to which are unknown) and it should now be able to provide the correct outputs for these. For

[12] Hopson v. City of Baltimore, 232 F.R.D. 228, 244 (D.Md. 2005) Electronic discovery may encompass "millions of documents" and to insist upon "record-by-record pre-production privilege review, on pain of subject matter waiver, would impose upon parties costs of production that bear no proportionality to what is at stake in the litigation." This type of privilege review is rendered obsolete with the effective use of search and retrieval technologies.

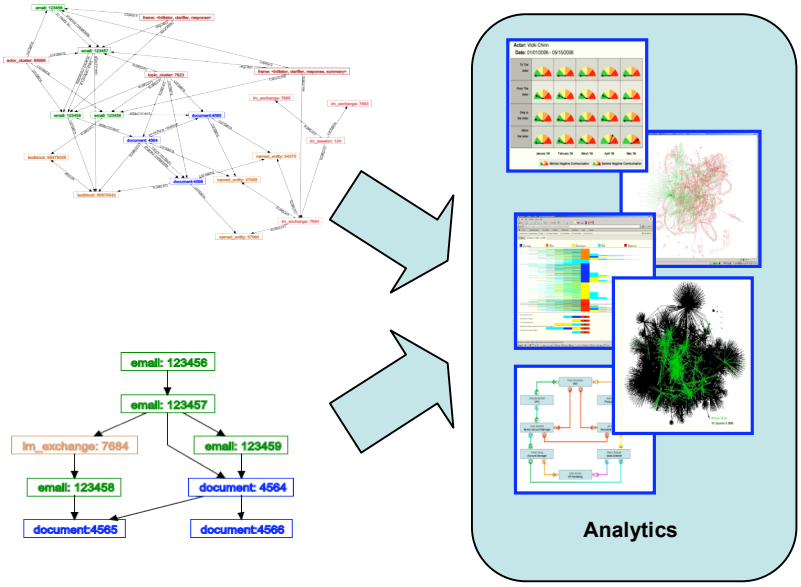
purposes of electronic discovery, the inputs might be information about documents and the outputs a *categorization* of those documents.

*Vector Space Modeling* (VSM) is a concept that first came into favor in the early 1970s and it has provided some additional guidance in automated document review even to this day. It is based on building vectors that describe the relationships between each search query and each file in the collection. Each vector, by its magnitude and direction then maps to other files that are closest to it in relation to the same *feature* as emphasized by the search query. Each file thus becomes a compilation of *features* that place it in a multi-dimensional construct. That construct can be realized in a graphical display depicting all the relationships as vector lines between and among separate files.

All of these linguistic and statistical search and retrieval methodologies return data that meets specific criteria. Each methodology has its own deficiencies and strengths. Using these methodologies together can help compensate for these shortcomings. Using them in conjunction with keyword search can help illuminate gaps in the searches as well. Using multiple methodologies will help focus on the data that is most likely to be responsive.

**Pattern analysis**

Although statistical and other linguistic methods will give better or more targeted results than keyword search will, it may not be enough to locate the most relevant or useful data. Social networking or behavioral analysis can also be used for more insight into the data set, revealing actions and decisions of the various actors. The entire set of data is built into a model to establish a baseline of behavior for individuals or groups present in the data. Causally related items can be grouped together for faster document review.




Using a data model, once a baseline of behavior is established, variations in behaviors can be shown. Anomalies in the behavior can be detected and exposed. Some anomalies in behavior can be explained. For example, if a person's email communication pattern reveals them sending several hundred messages a day and then it drops down to 5-10 emails a day, is that person sick, on vacation, or traveling? Or is there a more disturbing explanation? Perhaps all communication has moved over to the phone or instant messaging? Modeling the data set allows for an intimate look into the social networking relationships and patterns of behavior of an individual, group, or company.

Anomaly detection, whether there was good or bad intent, can be done, showing communication patterns, deletion patterns, changes in behavior, centers of power, social networks, decision-making patterns, and any number of other things. This type of analysis can be used to gain invaluable insight in litigation or investigations. This type of pattern analysis and behavioral analysis can go to the heart of a case and quickly help attorneys locate the information needed to craft their story. These are the kinds of questions to pose and look at the pattern or behavior analysis for answers:

- When were customary work practices circumvented?
- When did established norms of behavior change?
- Who knew, or likely knew, what facts?
- Who interacted with whom and how intimately?
- Who was involved in what types of decisions or meetings?
- Who are the real 'insiders'?
- What data is hidden or missing?
- When were electronically documented conversations "taken off line," possibly in an attempt to avoid detection?
- How did the importance of different actors change over time?

These are the kinds of questions lawyers ask themselves when they are trying to make a case or figure out a defense strategy. This behavior or pattern analysis is a more sophisticated look into the data than merely retrieving individual items in a data set. This behavioral analysis allows an all-encompassing view into the data rather than merely looking at a slice or segment of the data. This comprehension and insight into the data is invaluable.

## Conclusion

Moving beyond keyword search (and reading through masses of under-inclusive and over-inclusive search results) and making use of any of these advanced search and retrieval techniques to find and focus on the more likely to be responsive or privileged data allows lawyers to spend more time "lawyering" rather than searching and trying to find the right information. In a seminal study on the efficacy of human document review, humans thought they were retrieving 75% of the relevant documents, when in reality they were retrieving less than 20% of the relevant documents.[13] Using technology is necessarily more consistent than human review because such work is done by a machine, which does not get tired or make mistakes. These search and retrieval methodologies do not replace lawyers; they supplement each lawyer's judgment.[14]

---

[13] *An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System*
*Blair & Maron (1985).* Human beings retrieved less than 20% of the relevant documents when they believed they were retrieving over 75%

[14] FRE 502(b) Explanatory Note on Evidence Rule 502 Prepared by the Judicial Conference Advisory Committee on Evidence Rules (Revised 11/28/2007). "Other considerations bearing on the reasonableness of a producing party's efforts include the number of documents to be reviewed and the time constraints for production. Depending on the circumstances, *a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken "reasonable steps" to prevent inadvertent disclosure.*" (emphasis added)

Judgment remains the primary reason for hiring a particular lawyer or lawyers. Using these advanced search and retrieval techniques allows lawyers to focus on particular documents rather than trying to review every document. Lawyers play a significant role when any of these search and retrieval technologies are used – this role, however, is focused on retrieving the information they need rather than in reviewing documents. Using these techniques or methodologies isn't replacing lawyers with artificial intelligence (or letting machines do the lawyers' thinking for them), it is an effective use of technology and lawyers' skills and judgment to focus on finding the right information a quickly possible.