December 7, 2011

Three Methods for eDiscovery Document Prioritization:

Comparing and Contrasting Keyword Search with Concept Based and Support Vector Based "Technology Assisted Review-Predictive Coding" Platforms

Tom Groom, Vice President, Discovery Engineering | D4

Director, Discovery Services
[illegible]
[illegible]

d4discovery.com

Much has been written lately on how software applications can reduce document volumes in the eDiscovery process. Many vendors in the industry traded in their "Early Case Assessment" banners from a few years ago for the newer "Technology Assisted Review – Predictive Coding" (TAR-PC) shingles found on the exhibitor floor during Legal Tech New York 2012. Some vendors were even saying "Keywords are dead." Nothing could be—or should be—further from the truth.

Keyword Search, Concept Based Search and Support Vector Machines are all three valid approaches for document classification but there are key differences that should be considered before deciding which and—perhaps more importantly, when—to employ these approaches in the eDiscovery workflow. The intent of this white paper is to highlight the differences in the features, functions and benefits of these three approaches and identify potential application areas where they best work in the eDiscovery lifecycle.

Keyword search is not a TAR-PC approach but will be used in combination with TAR-PC for most engagements in order to achieve optimal results. The two TAR-PC platforms considered in this paper are the current version of Relativity Assisted Review for Relativity 7.3 and Equivio Relevance version 3.7. There are significant differences between the conceptual based "categorization" found in Relativity Assisted Review (RAR) as opposed to the active machine learning system coupled with a Support Vector approach with Equivio Relevance. One (Equivio Relevance) is an automated closed loop approach with rigorous statistical validation that can be done at various stages of the EDRM, and the other (Relativity Assisted Review) is a very flexible, open ended, "get more like these" statistically validated categorization system that can be utilized for data loaded in Relativity.

It is critical to keep in mind there is no one standard workflow for using the three approaches described in this paper. Optimal results are obtained through consultation with skilled and experienced professionals who understand where these tools best fit for a given document population. In most cases, a combination of these approaches will be necessary to yield optimal results.

Key Definitions and Measuring Search Efectiveness

Concepts in this paper are well documented in the statistical and information retrieval literature. For convenience, a short list of key definitions, a description of machine learning and a brief discussion of the difference between recall and precision is provided to benefit the reader's understanding.

Definitions

Inferential Statistics is the field of statistics in which evidence from one set of observations is used to make inferences about another set of observations. As used in TAR-PC, a subject matter expert makes "content relevance" decisions[1] on documents from a smaller sample set and then allows the system to infer

[1] The author will hereafter use the term "content relevant" or "content relevance" in this white paper to denote documents that have similar content but may not be relevant to the present dispute because they are outside the date range or are in some other categorical way not relevant. It is critical to note that culling methods such as date filters may have to be used in conjunction with any of the three methods discussed in this paper in order to refine the corpus to a more truly "relevant" dataset.

and propagate those decisions across the larger set of documents.

Random Sampling is a means to obtain a truly representative sample of a large population by enabling all documents to have an equal chance for being selected. Random sampling is necessary for the inference methodology to be "statistically significant" which enhances the overall defensibility.

Statistically Significant indicates the likelihood that a result or relationship is caused by something other than mere random chance. Tests that are validated with a 95% confidence level or above are considered "statistically significant."

Confidence Level is expressed as a percentage and represents how often the true percentage lies within the confidence interval. The 95% confidence level means that if the test were run 100 times, the same results would be delivered 95 times.

Confidence Interval represents the variation from the confidence level for the population. For example, if a confidence interval of +/- 2% were used with a 95% confidence level, the margin of error would be between 93% (95 – 2%) and 97% (95 + 2%). A narrower confidence interval will require a larger sample size. For TAR-PC a confidence interval of +/- 2% is common.

Sample Size is the number of documents required to be sampled to satisfy the confidence level and confidence interval combination. The sample size does not increase linearly on the total size of document population in excess of 100,000 documents (which is considered "very large" by statisticians). A very good primer on sample size and an online sample size calculator can be found here.

Richness is expressed as a percentage of how many documents are actually content relevant for a given population. A document population with low richness would not contain as many content relevant documents as one with high richness for two document populations with equal size. Low richness will likely increase the sample size.

Recall conveys the "completeness" of the content relevant documents that were retrieved by the system. Recall is expressed as a percent and the corollary conveys the percentage of content relevant documents that were not retrieved by the system. (More on Recall below.)

Precision conveys the "purity" of retrieved documents by the system that are content relevant. Precision is expressed as a percentage and has an inverse relationship with recall. (More on Precision below.)

F-measure is the harmonic mean of recall and precision. The F-measure accounts for the balance between precision and recall where an F-measure reaches its best value at 1 and worst score at 0.

Machine Learning Systems

Machine learning is a branch of computer science concerned with the design and development of algorithms that allow computers to infer behaviors based on empirical data. Both Relativity Assisted Review and Equivio Relevance are examples of systems based on machine learning.

Machine learning systems for document prioritization can be organized into the following types:

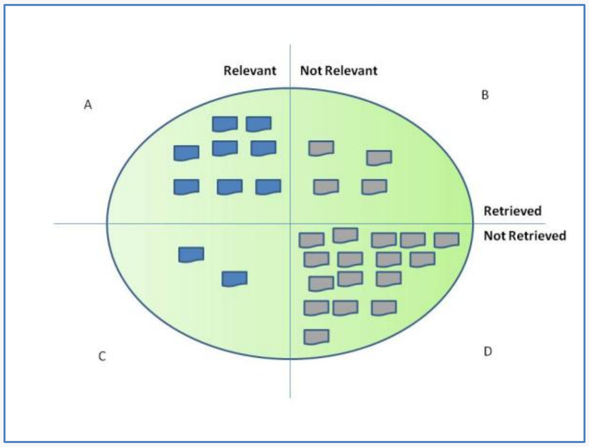
- ***Unsupervised Learning*** systems model with one set of inputs without human interaction. Clustering with Relativity Analytics is an example of unsupervised learning. The document clusters are organized by the software and then labeled by the system.

- ***Supervised Learning*** systems map inputs from a human to desired outputs. Categorization with Relativity Analytics is based on supervised learning where humans chose document exemplars to feed to the system. The system then tags and ranks the remaining documents in the collection based on similarity (or dissimilarity) to exemplars ("find more like this"). Relativity Assisted Review is based on categorization and therefore can be classified as a supervised learning system.

- ***Active Learning*** systems predict new outputs based on training inputs, training outputs, and test inputs. This type of closed loop system chooses the document exemplars to feed to the human who makes the content relevance decision. The system learns from these determinations and iteratively chooses the next exemplars to maximize its learning. Once the system learns and validates all it needs to know (becomes statistically stable) the system applies a relevance score from 0 (not relevant) to 1 (relevant) based on what it has learned to all of documents in the collection. Equivio Relevance is classified as an active learning system.

Recall and Precision

Information Retrieval scientists measure the retrieval system's effectiveness by determining the system's precision and recall. Precision is the fraction of retrieved documents that are content relevant, while recall is the fraction of content relevant documents in what was retrieved. Said another way, high recall means that the system returned most of the content relevant documents. High precision means that the system returned more content relevant documents than non-content relevant documents. Recall is a measure of completeness or quantity whereas precision is the measure of exactness or quality. Both precision and recall are therefore based on an understanding and measure of content relevance which is established by the human "subject matter expert."



Consider this example where we have 30 documents in a collection. The 10 blue documents in the two quadrants on the left side are content relevant and the 20 gray documents on the right side are not content relevant. The 12 documents in the top two quadrants (A and B) were retrieved by the system and the documents in the bottom two quadrants (C & D) were left behind or not retrieved by the system.

Calculating Recall

Systems with better recall means more content relevant documents were identified by the system. The calculation for this example would be:

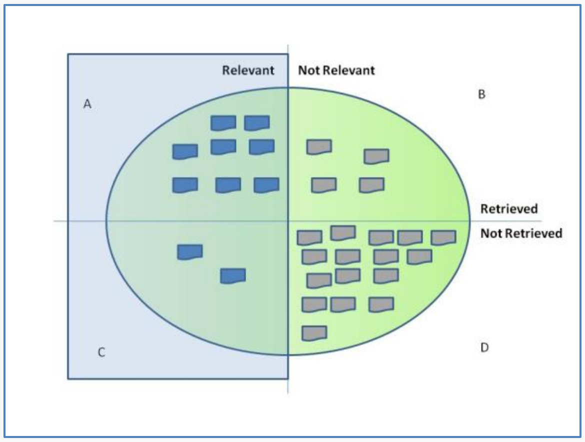
Recall = True R Docs Retrieved/Total R Docs

Recall = A/(A+C)

In this example:

8/10 or 80%

RECALL = 80%



Calculating Precision

Systems with better precision means more content relevant documents were actually retrieved by the system. The calculation for this example would be:

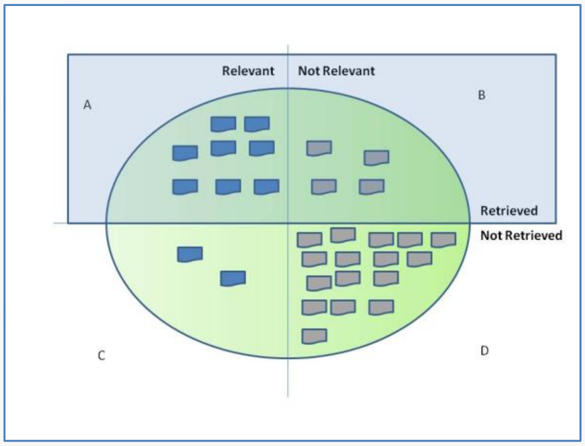
Precision = True R Docs Retrieved/Total Docs Retrieved

Precision = A/(A+B)

In this example:

8/12 = 67%

PRECISION = 67%

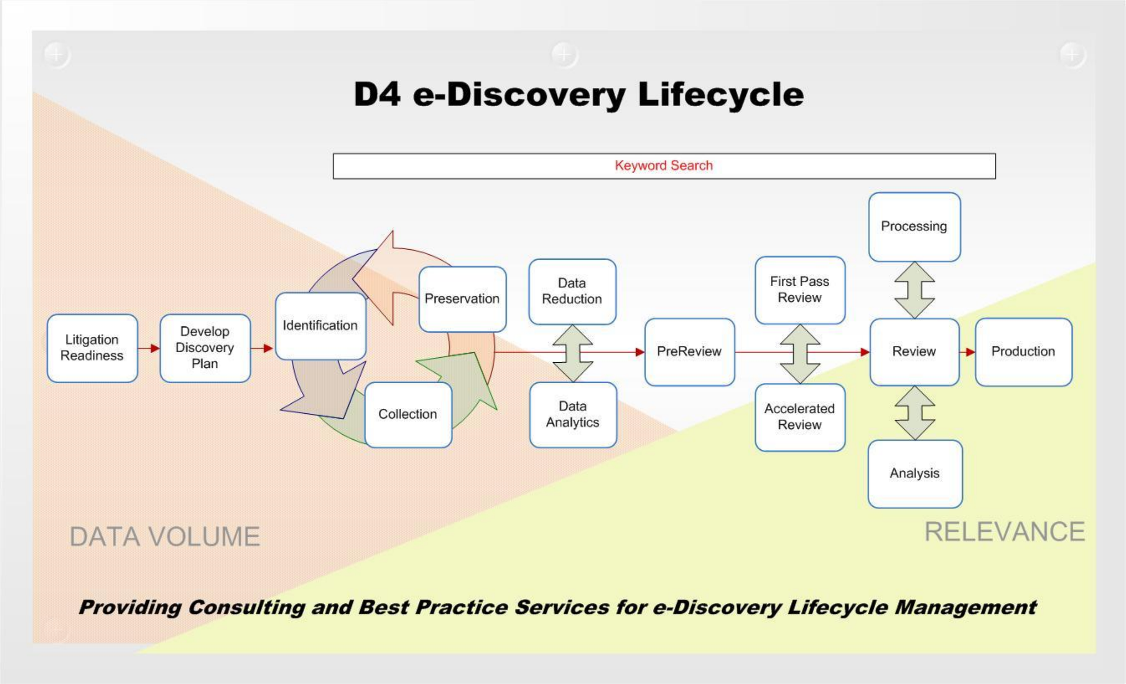


A Close Look at the Three Approaches: Keyword Search, Concept Based and Support Vector Based Technology Assisted Review – Predictive Coding

1. Keyword Search

If used properly Keyword Search can be a very effective approach for cutting down the volume of ESI that has to be reviewed. This can be achieved by employing iterative workflows, random sampling, and result verification to ascertain the best keywords in reducing the volume in the discovery process. Keyword Search can also be very effective in isolating specific boundaries such as date ranges as well as finding proper names in a collection and is often used in conjunction with a TAR-PC approach.

The following graphic shows where keyword search can be applied in the eDiscovery Lifecycle.



Keyword search is built into most eDiscovery tools. It has a wide range of applicability because it can be used during the collection and preservation process all the way through production and the QC of the production. The prioritization of documents using keywords would be either "hit" or "not hit" for each term and the ranking concept will need to be provided by the search engine (such as "priority" found in dtSearch). Keyword search also includes date range search which often can be a very effective means to provide boundaries to identify truly responsive material.

Keyword search does, however, have its limitations. The challenge with using keywords is twofold in that (1) keyword searches normally result in very low recall due to keyword limitations, and (2) it is very difficult to calculate the true recall and precision rates. The Blair Maron study is the most commonly cited study and TREC has subsequently confirmed the best human recall rate is *estimated* to be 20%-40%. This is largely due to the numerous ways to express concepts by the use of different words. Using an iterative approach with keyword expansion and statistical validation can improve keyword limitations but only to a point.

Although not initially apparent, the eDiscovery professional can play an important role in the selection of key words. Initially, counsel and the client must develop a draft list using their knowledge of the case and important players. Non-intuitive client acronyms, usage and jargon can be identified and incorporated into the list. From this list, the eDiscovery professional can identify "noise" words, refine Boolean queries, and efficiently combine words using wild card and proximity search limiters in order to achieve optimal results.

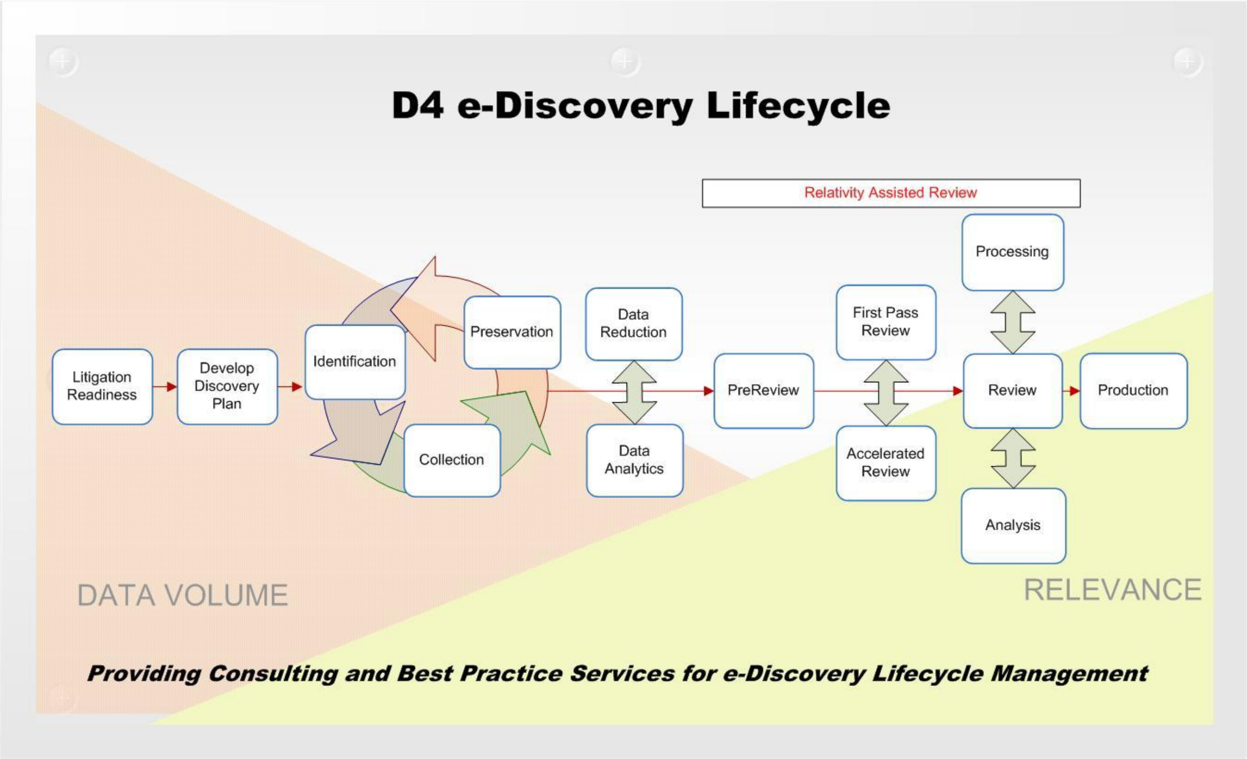
The ideal keyword approach would be for a computerized system to automatically identify the appropriate weighted search terms to include and exclude. Furthermore, if that system could determine the recall and precision based on iterative samples, a statistically significant model could be created to describe the distribution for that corpus of documents. This could be done without having to actually review every document because it approaches the problem from a statistically significant point of view via sampling and verification. That system does exist as a Support Vector Machine, which is the underlying approach found in Equivio Relevance (more on that later in this paper).

2. Concept Based TAR-PC

Conceptual Search Engines translate the meaning of words used in context for a corpus of documents into multi-dimensional mathematical models referred to as "concept space." Once the concept space has been built for a corpus, a "find more like these" classification workflow can be employed to find documents that are similar in conceptual content.

Relativity Assisted Review (RAR) is based on such an approach and is supported by an industry leading Latent Semantic Indexing (LSI) Conceptual engine from Content Analyst surrounded by a statistic model to monitor results. RAR is categorization workflow that uses the conceptual index to "find more like these." The data is processed, loaded into Relativity and indexed with Relativity Analytics (the Content Analyst LSI system). Once that is done, the RAR workflow requires the trainer to supply a confidence level and confidence interval, and the RAR system randomly selects documents for the humans to review and determine content relevance. Once a batch of documents is reviewed, the validation step includes sampling the results of the categorization iteratively until the desired accuracy is met.

For defensibility, the built-in RAR statistical model enables sampling the documents to the point of calculating the "content relevant documents left behind" to generate the recall and precision rate. The graph below shows where Relativity Assisted Review can be used in the eDiscovery workflow.


D4 e-Discovery Lifecycle

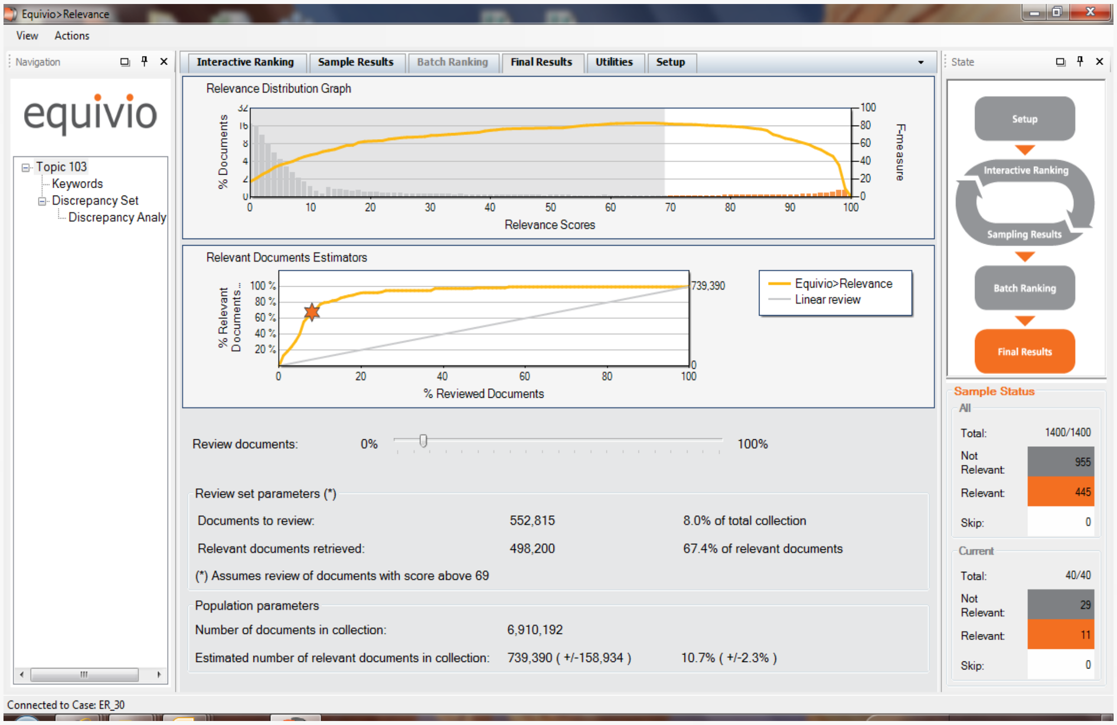
Providing Consulting and Best Practice Services for e-Discovery Lifecycle Management

Definitively, concept-based search engines improve the overall results versus just using keywords. The system can return more potentially content relevant documents through concept analytics without the limitations of Boolean logic, and the false positives can be suppressed through document seeding and refining the exemplar set. An output of categorization is a ranking on how closely the documents resemble the exemplars provided to the system. Importantly, only the documents that are within the similar threshold receive the ranking.

Combining Relativity Assisted Review and keyword search with proper workflow and methodology can yield effective and very defensible results. eDiscovery professionals can build this workflow and help lawyers decide how the technology will make the greatest contribution. The RAR workflow can be an effective means to identify the concept relevant subset of a collection that is in Relativity and then use keyword search and standard categorization to identify truly relevant documents for a specific timeframe, interdependency relationship, and/or conceptual issue. Since RAR is fully integrated in Relativity, review teams can begin the review process without RAR and then leverage the initial determinations should it be decided to implement RAR later. While it can be calculated later in the review cycle, the current version of RAR does not automatically establish "recall" and "precision" rate in the beginning as found in Equivio Relevance.

3. Support Vector Based TAR-PC

The Equivio Relevance approach uses a document classification methodology called “Support Vector Machine” (SVM) to identify a set of weighted terms to include and exclude for predicting and prioritizing results. SVM is a well-established "predictive analytics" statistical modeling methodology, and is widely used in a great variety of industrial applications, including speech recognition, facial expression categorization, handwriting recognition and computational biology. The Support Vector Machine used in Relevance automatically establishes the system's recall and precision of a given population early in the process. Once recall and precision are determined, management can make informed decisions on the best approach for the prioritized document review. Equivio Relevance provides a very helpful user interface to enable decision making functionality.



Review documents: 0% 100%

Review set parameters (*)

Documents to review: 552,815 8.0% of total collection

Relevant documents retrieved: 498,200 67.4% of relevant documents

(*) Assumes review of documents with score above 69

Population parameters

Number of documents in collection: 6,910,192

Estimated number of relevant documents in collection: 739,390 (+/-158,934) 10.7% (+/-2.3%)

Sample Status

All

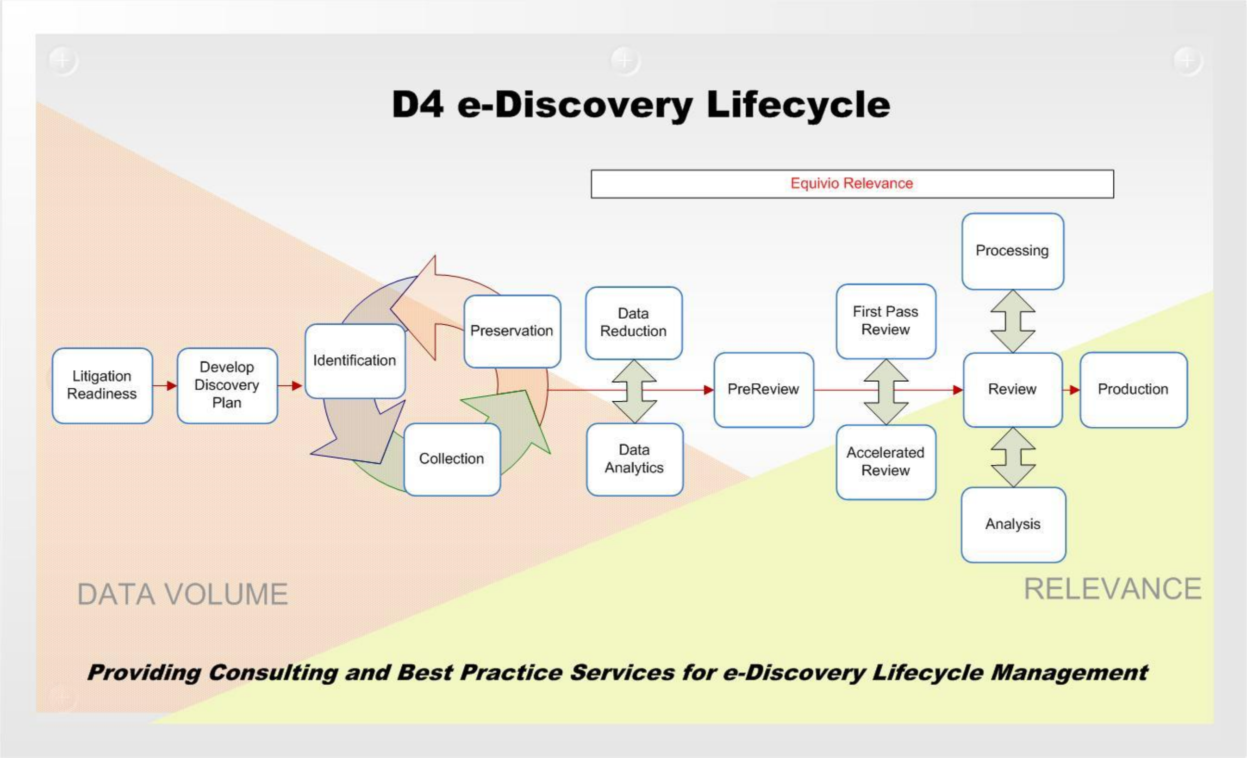
Total:	1400/1400
Not Relevant:	955
Relevant:	445
Skip:	0

Current

Total:	40/40
Not Relevant:	29
Relevant:	11
Skip:	0

Like Relativity, the Equivio Relevance application does not require document seeding. The subject matter expert (SME) is presented with selected documents by the system without queries or search terms. The SVM periodically selects documents that have been previously presented to the SME to insure consistency, as well as documents that are similar to previously selected documents in order to disambiguate between the two. The SVM is generating a set of weighted terms to include and a subset of weighted terms to exclude, which will result in optimal recall and precision outcomes. After the SVM has gathered enough information to ensure statistically significant metrics, and has learned enough to predict what the SME would choose for a given iteration, the SVM can then be tasked to apply a "score" across the entire document population, resulting in a ranking of the corpus in order from the most content relevant to the least content relevant. Unlike Relativity Assisted Review, Equivo Relevance is not a document review platform so the Relevant Scores have to be exported and then imported into the target Review platform.

The graphic below shows where SVM Predictive Coding with Equivio Relevance can be used in the eDiscovery lifecycle.



D4 e-Discovery Lifecycle

Providing Consulting and Best Practice Services for e-Discovery Lifecycle Management

Equivio Relevance requires only the document text (no metadata) to sample and score documents early in the lifecycle. Equivio Relevance can be used as an "Early Data Assessment" technique during data reduction and analytics. The document scores can be used as a means to bypass first pass review altogether for some documents or as a means to divide the review into prioritized groups based on score. The weighted keywords collected by the Relevance SVM can be exported and used in negotiations with opposing counsel or for other purposes when keywords are needed, such as collection filtering. The Relevance SVM is very effective in scoring documents that are "not content relevant" and therefore is an effective approach for removing non-content relevant material from the review phase altogether. More examples on how Equivio Relevance can be used in the eDiscovery process are found in the article Applying Predictive Coding to Reducing Cost in Document Review.

Competent eDiscovery professionals should be consulted to direct lawyers on how to utilize this technology and build the optimal workflow which often requires a hybrid approach. This was recently underscored by Judge Peck in his *Da Silva Moore, et al., v. Publicis Groupe, et al.* opinion:

1. predictive coding is a reasonable, and probably more effective, approach for identifying relevant ESI than currently available alternatives;

2. the appropriate technology must be combined with the right process to be defensible; and,

3. simplistic counter-arguments to the use of predictive, such as "black box", "not perfect" and "using predictive coding means meeting the Daubert standard", will no longer be persuasive.

Comparison Table

The table below highlights the key features of the three approaches described in this paper.

Feature / Attribute	**Keyword Search**	**Relativity Assisted Review**	**Equivio Relevance**
Base Technology	Word search / Boolean Logic	Latent Semantic Indexing	Support Vector Machine
Are all documents scored?	N/A	No. Only the documents above the conceptual threshold are scored. RAR does score Relevant and Not Relevant on separate scales. Documents that are neither Relevant nor Not Relevant will not be scored.	Yes. All documents are provided with a score from 0 to 1.)
Type of Machine Learning	N/A	Supervised Learning	Active Learning
Have to seed with exemplars?	N/A	No. RAR automates the process of identifying the exemplars.	No. The SVM systematically selects samples to optimize its learning.
Can be run in the Relativity document review platform	Yes. Keyword search is available in most eDiscovery tools including Relativity	Yes. Requires Relativity Analytics and the RAR application. RAR is fully integrated with the Relativity Review Platform.	No. Can be run external to REL on extracted text. The document scores can be overlaid into Relativity or other review platforms
Extracted text is needed	Yes	Yes	Yes
Can be used for which phases of the eDiscovery Lifecycle	Collection Data Reduction Review Production QC	PreReview First Pass Review Accelerated Review Review Review Analysis Production QC	Data Reduction Data Analytics PreReview First Pass Review Accelerated Review Review Review Analysis Production QC
Automatically calculates "Recall"	No	No	Yes
Automatically calculates "Precision"	No	No	Yes
Automatically computes "F measure"	No	No	Yes

Provides Sampling system for QC results	N/A	Yes	Yes
Once trained, can apply system to new incoming documents	N/A	Yes	Yes
Generates a set of weighted terms that can be exported	No	No	Yes
Scoring results sortable in the review system	No	Yes	Yes
Can use to score for general "conceptual relevance"	N/A	Yes	Yes
Can use for scoring multiple content relevant issues	N/A	Yes. The RAR workflow is based on one item (i.e. Relevant or Not Relevant) but a reviewer can "tag" exemplars for multiple issues during RAR training and use the "categorization" feature to find documents that are similar.	Yes

Summary

This paper explores the three primary approaches eDiscovery professionals use to systematically prioritize documents that are being considered for review:

1) Keyword Search

2) TAR-PC based on Categorization with Conceptual Search Technology

3) TAR-PC based on an Active Machine Learning approach using a Support Vector Machine

Each of these three approaches has its own unique set of advantages and disadvantages and, when used in conjunction with careful documentation and appropriate iteration, can survive a challenge by a party's opponent in litigation. The optimal workflow will likely require a hybrid approach applying two or perhaps all three of these approaches to reach the best results.

The implementation of Technology Assisted Review – Predictive Coding found in concept based Relativity Assisted Review and Support Vector based Equivio Relevance yields greater results than just using keywords search. These systems can return more potentially content relevant documents without the limitations of Boolean logic. However, a well-constructed keyword search can be more effective for certain conditions such as dates or proper names. Putting these technologies together with proper workflow, methodology and documentation will deliver the most effective results and enhance their defensibility.

Perhaps more important than the technology itself is how the skilled eDiscovery professional -- who understands these systems, their limitations and strengths -- combines them into the optimal workflow best suited to meet the overall requirements for the given project. In the end, the defensibility of these approaches is based on the workflow (the process) and not the technology itself.

About the Author

Tom Groom is a recognized eDiscovery expert with more than three decades of experience in information technology, litigation support, document review, and eDiscovery methodologies. As the Vice President of Discovery Engineering at D4, Tom advises both corporate and law firm clients on issues involving ESI, including defensible collection methodologies, optimized ESI processing and review workflows, and litigation contingency and readiness planning.

Tom has presented numerous CLEs and training seminars on ESI topics including optimized ESI workflows with various approaches to Native File Review and production methodologies. He is also the point person for matters involving complex electronic discovery services, Web-based repositories, technology assisted review – predictive coding and document review projects.

Tom's career began with IBM where he served for eight years as a System Engineer followed by seven years with IKON Digital Litigation Services. Tom then spent three years in the Geospatial software development and computer storage industries before returning to the legal industry in 2004. Tom holds a Master of Science degree in Industrial Engineering with a focus on computer information systems from Arizona State University and a Bachelor of Science, Industrial Engineering from the University of Arkansas.

D4 is here for you.

Call D4 today for immediate help or information: **800.410.7066**

Visit the D4 website for additional resources, white papers and case studies: **d4discovery.com**

Connect with D4 online and a team member will get back to you within 24 hours: **d4discovery.com/contact**