ACC DOCKET
INFORMED. INDISPENSABLE. IN-HOUSE.

There is Nothing to Fear from Technology Assisted Review

Litigation and Dispute Resolution

Technology, Privacy, and eCommerce

CHEAT SHEET

- ***The history of TAR.*** From as early as 2005, attorneys have stressed the importance of automated discovery in transforming the time, cost, and accuracy required to analyze large sums of documents.
- ***TAR and the United States.*** The 2012 court decision *Da Silva Moore v. Publicis Groupe* allowing for the use of technology assisted review in discovery paved the way for countless court decisions in the United States.
- ***How it works.*** Understanding the use and evolution of vector matrix systems and latent semantic analysis is both easy-to-learn and essential to the use of technology assisted review.
- ***In-house responsibility.*** It is imperative for in-house counsel to accept technology assisted review and its place in modern litigation.

Ever since information has been stored in the memory of a computer, litigants have been retrieving this information for discovery pursuant to litigation. With the advent of email, databases, and the capacity to store even more electronic information, there's no doubt that the number of electronic records will continue to grow. This has and will continue to significantly increase the costs of litigation, especially during the discovery phase. Most of the costs stem from attorneys reviewing each document, determining its relevancy, and then categorizing the document. However, these costs can now be reduced by implementing software programs that perform these activities. In many cases, manually reviewing documents is not acceptable due to the large amount of documents that need to be reviewed. This is a major reason why courts are now starting to accept automated methods of reviewing documents.

These automated methods are known by many names,[1] but the most prevalent name used is technology assisted review (TAR). TAR describes a broad range of processes whereby subject matter experts classify a small portion of an entire set of documents, so that a computer can then automatically extrapolate those judgments and categorize the remaining documents. Predictive coding is a species of TAR and is characterized by using a supervised machine learning algorithm[2] to perform this extrapolation. Machine learning algorithms effectively allow computers to learn without being explicitly programmed for every possible scenario it may encounter.

Machine learning algorithms are not unique to document production. To some extent, most of us interact with machine learning algorithms every day. For example, personal assistants on our smartphones, such as Siri,[3] Cortana, and Google Now all use machine learning algorithms in order to provide more accurate responses to user questions. Target, Amazon, and other large retailers use machine learning algorithms to predict what items customers in certain geographic areas are likely to purchase and then stock warehouses in that area with the predicted items. Retailers also use machine learning algorithms to provide relevant coupons to customers. Many websites now have customer support representatives that are actually computer programs trained using machine learning algorithms. Various news websites now use machine learning programs to write news stories. Other common uses of machine learning algorithms include optical character recognition, email spam filtering, and web search engines.

Machine learning algorithms are also used to automate the production of documents relating to litigation. After storing an image of the document and extracting the text, machine learning algorithms can be trained by attorneys involved in the litigation. To train the algorithms, attorneys categorize (e.g., privileged, confidential, relevant, or not relevant) each document on a small subset of all potentially relevant documents just like traditional manual document review. Algorithms then use these inputs to predict how the attorney would categorize documents it hasn't seen yet based on the new document's similarity to a document in the categorized subset. Litigation costs can be dramatically reduced by using TAR in this manner, since clients are billed for the time it takes attorneys to manually review a small portion of the overall document set.

This article will explain the history of TAR usage, the technical aspects of the most popular TAR algorithms, and the advantages and disadvantages of the various forms of TAR algorithms.

[1] Other names include: "computer-assisted review," "computer-assisted coding," "technology-assisted review," "technology-assisted coding," and "predictive coding."

[2] Supervised machine-learning refers to using a subset of the entire document set and providing the computer with the correct answer for that subset. In contrast, unsupervised machine-learning does

not use a training set and instead attempts to infer relationships among the data.

3 Siri is based on an artificial intelligence project known as Cognitive Assistant that Learns and Organizes (CALO), originally funded by the Defense Advanced Research Projects Agency (DARPA). The goal of CALO was to research how to incorporate many different types of artificial intelligence technologies into one platform.

History of TAR

The notion of using TAR as a tool for reviewing documents has been around for some time. Even before electronic records became prevalent, litigants would scan paper documents into an image file and use optical character recognition algorithms,[4] which is itself a machine learning algorithm, to extract the words from the image. Once the documents were scanned, keyword searches could be performed to retrieve documents that include the keywords.[5] However, these keyword searches did not take into account misspellings, local colloquialisms, or synonyms. Traditional keyword searching missed many relevant documents and there was a need for a better method of retrieving pertinant material. Since the documents returned in a keyword search still had to undergo costly review by humans, there was a need for a method that could further reduce costs.

In response to these needs, Anne Kershaw published the 2005 article “Automated Document Review Proves Its Reliability.” It described how computers can implement statistical analysis instead of keywords to make automated relevancy assessments to categorize documents. This article highlights the cost saving benefits of using TAR for discovery.

In 2006, the US Department of Defense and the National Institute of Standards and Technology (NIST) jointly created the Text Retrieval Conference (TREC) in order to study efficient text retrieval methods. Known as the TREC Legal Track, annual conferences were held from 2006 to 2011 to discuss progress made in the field of electronic business record retrieval for use as evidence in litigation. The TREC Legal Track also made test collections of documents available to the public so that anyone could test their automated retrieval method using the same test documents as everyone else. This provided for a more accurate comparison of test methods.

In 2007, a nonprofit research and educational institute for the study of law and policy in the areas of antitrust law, complex litigation, and intellectual property rights called The Sedona Conference, published an important article relating to the evaluation of automated retrieval methods.[6] The article argues that while keyword searching is an effective way to identify particular documents when the language was relatively predictable (e.g., searching for a date regardless of its context), it has significant disadvantages. For example, keyword searches do not take into account the context of the keyword within the document, resulting in the retrieval of an excessive amount of irrelevant documents. In addition, keyword searches could miss important documents because the keyword is misspelled or a synonym for the keyword is used in the document instead of the keyword itself. The article describes several advanced methods of document retrieval that are more effective at retrieving relevant documents than keyword searching. One of these methods, the machine learning method, is the basis of the most widely used document retrieval method today.

In the fall of 2009, The Sedona Conference® published the Sedona Conference® Cooperation Proclamation,[7] which urged litigants to cooperate with each other during the discovery process to arrive at a mutually agreeable usage of TAR in order to reduce litigation costs. The Cooperation Proclamation was endorsed by 95 active and retired judges from various federal and state courts and has been widely cited throughout the globe.

Many other articles have emphasized the importance of using TAR in document identification. However, the legal profession has been slow to accept TAR technology. The main reason for this lack of acceptance is a misunderstanding or lack of understanding of how TAR actually works. Because of this, attorneys, judges, and even clients, are hesitant to implement a methodology in their case and would rather wait until it has been extensively tested first. Therefore, in order to increase acceptance of TAR, litigants have to become more familiar with it and understand how it operates. There is an ever-growing list of jurisdictions both in the United States and in foreign countries that have accepted the use of TAR.

[4] Optical character recognition is still used today with electronic documents. For example, a .pdf file needs to be processed through optical character recognition software in order to extract the letters from the .pdf file.

[5] The same text searches are utilized by many automated case law retrieval services such as LexisNexis® and Westlaw®, as well as by popular web search engines including Google®.

[6] “The Sedona Conference® Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery,” 8 SEDONA CONF J., 191-223 (2007).

[7] 10 SEDONA CONF. J. 331 (2009 Supp).

Acceptance of TAR in the United States

The United States has more published opinions endorsing TAR than any other country. The first opinion in the United States approving the use of TAR was published on February 24, 2012, by Magistrate Judge Andrew Peck in *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (SDNY 2012). *Da Silva Moore* was an employment discrimination suit in which five female employees accused Publicis Groupe and its US public relations subsidiary MSL Group of gender discrimination. During discovery, three million electronic documents were collected. To determine their relevancy, they all had to be reviewed. MSL Group proposed a predictive coding protocol to reduce the time and costs associated with reviewing the documents. However, the employees filed objections to this protocol on several grounds, including that predictive coding was not a reliable and accurate method and would not perform with 100 percent accuracy. Magistrate Judge Peck held that many studies show that TAR works better than most alternatives. Peck further explained that the goal of TAR is to reduce costs, not to perfectly review all the documents. Peck referenced the TREC Legal Track and other studies for the notion that manual review results in many more errors than a technology-assisted review and went on to formally approve the use of the proposed predictive coding protocol set forth by MSL Group and ordered the parties to cooperate in implementing a mutually agreeable TAR protocol.

While this case has been cited by many as paving the way to the use of TAR in litigation, it has not come without criticism. For example, some critics argue that courts should not mandate the use of predictive coding, but instead allow the parties to decide whether or not to use this technology based on the circumstances in each case.[8] This line of criticism reasons that the “court’s role is to supervise the discovery process and to intervene when it is abused by the parties. Abuse occurs when parties violate the letter or spirit of the FRCP through gamesmanship… A general concern for fairness permeates the process.” However, it could be argued that a party refusing to agree to using TAR could simply be a tactic to induce the other party to spend needless amounts of money on manual document review. Using the high costs of manual document review as leverage seems counter to the idea of fairness, and may be the type of gamesmanship that runs afoul of the spirit of the Federal Rules of Civil Procedure (FRCP). Interestingly, these very concerns were incorporated into the FRCP in an amendment that became effective December 1, 2015. This amendment significantly reduced the scope of discovery by replacing provisions allowing the discovery of evidence if it "appears reasonably calculated to lead to the discovery of admissible evidence" with provisions allowing discovery when it is "proportional to the needs of the case." The notes to the FRCP make clear that litigants should consider using TAR to achieve this proportionality standard:

> The burden or expense of proposed discovery should be determined in a realistic way. This includes the burden or expense of producing electronically stored information. Computer-based methods of searching such information continue to develop, particularly for cases involving large volumes of electronically stored information. Courts and parties should be willing to consider the opportunities for reducing the burden or expense of discovery as reliable means of searching electronically stored information.
>
> FED. R. CIV. P. 26(b)(1).

In light of the new FRCP language, it could be argued that a party unwilling to consider the use of TAR is violating the spirit of the FRCP through gamesmanship allowing the court to require the use of TAR.

Despite criticism, the *Da Silva Moore opinion* has been cited by many cases to support the use of TAR. For example, the United States Tax Court has mandated the use of predictive coding despite objections of one of the parties in *Dynamo Holdings Ltd. Partnership v. C.I.R.*, 143 T.C. 183 (2014). In this case, the commissioner of Internal Revenue wanted access to information contained on two backup tapes in Dynamo's possession. Dynamo refused and argued that the costs of sifting through the information stored on the backup tapes would be excessive. Dynamo proposed using predictive coding to identify relevant documents and to reduce the costs of reviewing the information stored on the backup tapes. The commissioner opposed the use of predictive coding and asserted it was unproven technology. Instead, the commissioner proposed that Dynamo could provide the tapes in their entirety to the commissioner so that the IRS could identify the relevant documents, with the caveat that Dynamo could take back any documents later found to be privileged or not relevant. The judge agreed with Dynamo, citing the principles set forth in *Da Silva Moore*, and allowed Dynamo to use predictive coding to identify relevant and nonprivileged documents for production.

Several other federal decisions throughout the United States have accepted the use of TAR due to the *Da Silva Moore* opinion. Even the Department of Justice accepted the use of TAR in at least two cases in 2013.

While TAR is becoming more accepted in United States courts, even Magistrate Judge Peck himself seems to be backing away from judicially mandating the use of TAR. A recent opinion written by Magistrate Judge Peck states that "[i]n the three years since *Da Silva Moore*, the case law has developed to the point that it is now black letter law that where the producing party wants to utilize TAR for document review, courts will permit it."[9] However, in this case Magistrate Judge Peck did not mandate the use of TAR as in *Da Silva Moore*. Instead, he held that "[t]he court is approving the parties' TAR protocol, but notes that it was the result of the parties' agreement, not court order." Thus, it is unsettled as to whether the use of TAR should be mandated by the courts or if the parties should agree to the use of TAR without a judicial mandate.

8 Tonia H. Murray, Mandating Use of Predictive Coding in Electronic Discover: An Ill-Advised Judicial Intrusion, 50 Am. Bus. L.J. 609 (2013).

[9] *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125, 127 (S.D.N.Y. 2015).

Acceptance of TAR in foreign jurisdictions

One of the first countries in the world to accept the use of TAR is Australia. On March 31, 2011, the Honorable Robert McClelland MP, Australia's attorney-general, issued a report titled "Managing Discovery: Discovery of Documents in Federal Courts." The report indicates that the parties should agree as to how to achieve a reasonable search. The report goes on to state:

> Advanced concept searching software, and more recently predictive coding technology (which is much more accurate than keyword searching), can quickly process large quantities of data and assist in identifying records relating to particular issues. This can be used not only to eliminate clearly irrelevant material, but can significantly reduce the amount of time required.

While the report of the attorney-general of Australia approves the use of TAR, there has been no published judicial acceptance of TAR in Australia.

On October 5, 2011, the Supreme Court of Singapore – High Court issued an order in *Surface Stone Pte Ltd. v. Tay Seng Leon* and another [2011] SGHC 223 (5 October 2011) that emphasized the importance of the proportionality principle in discovery and endorsed the use of TAR:

> [T]he principle of proportionality requires that the procedure to be adopted be tailored to the size of the dispute. For instance, in some cases, the inspection may be handled expeditiously with the use of special software such as de-duplicating software, data sampling software, or predictive coding software.

On December 23, 2011, the Ontario Superior Court of Justice decided *L'Abbé v. Allen-Vanguard*, 2011 ONSC 7575 (CanLII) that openly accepts TAR due to the large number of documents at issue:

> Various e-discovery solutions are available including software solutions such as predictive coding and auditing procedures such as sampling… Suffice to say traditional approaches to production motions cannot be used for production on this scale.
>
> L'Abbé v. Allen-Vanguard, 2011 ONSC 7575 (CanLII)

In its opinion, the court explicitly refers to The Sedona Conference® Cooperation Proclamation for support.

More recently, the Ontario Superior Court of Justice decided *Bennett v. Bennett*, 2016 ONSC 503 (CanLII), wherein the court noted that it was reasonable to use predictive coding for the first level review and then have lawyers and paralegals conduct the next level of review.

In 2013, the New Zealand Justice and Electoral Committee published a recommended Judicature Modernisation Bill. On page 1,113, a discovery checklist endorses the use of "concept searching, clustering technology, document prioritization technology, email threading, and any other new tool or technique." The term "document prioritization technology" is defined on page 1,121 as "technology that analyzes the decisions of a human review of a sample set of documents. The software then prioritizes/ranks the remainder of documents based on the decisions made on the sample documents, which allows the most relevant documents to be identified first." This bill is still waiting to be passed, but litigants in New Zealand could still use the existence of the draft bill to support the use of TAR in their case.

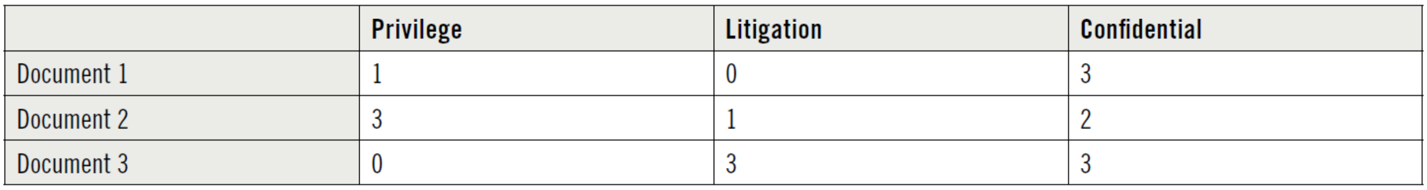
On March 12, 2015, the Irish High Court issued the first opinion in Ireland to endorse TAR in *IBRC & Ors v. Sean Quinn & Ors*. In this case, IBRC calculated that it would take 10 lawyers approximately nine months to manually review the pool of documents in the case. Instead of performing a manual review, IBRC proposed implementing predictive coding to identify relevant documents. When the defendant refused to use predictive coding, the Irish High Court ordered the defendants to cooperate with IBRC's proposed predictive coding plan and held that the use of predictive coding would satisfy a party's discovery obligations. The Irish High Court explicitly cited Judge Peck's decision in *Da Silva Moore* and recited the principles stated in that case. The court also noted that the e-discovery process does not have to be perfect, which is a common sentiment in many jurisdictions.

Similar to many other jurisdictions, the United Kingdom has also been reluctant to embrace TAR. However, earlier this year a recent decision by the England and Wales High Court (Chancery Division) approved the use of predictive coding in Pyrrho Investments Ltd v. MWB Property Ltd & Ors [2016] EWHC 256 (Ch) (16 February 2016). The opinion cites extensively to the decision in *Da Silva Moore*, and also noted that the Irish High Court has also recently endorsed the use of predictive coding. The court then states that because the costs of manual discovery in this case would be well over several million pounds in this case, manual document review was unreasonable and approved the use of predictive coding.

Technical description of TAR

Since it appears that global jurisdictions have approved the use of predictive coding, it is imperative for attorneys, judges, and clients to understand how it operates in order to discuss the advantages and disadvantages of the various types of predictive coding techniques available. The first step in understanding how predictive coding works is to understand how documents are represented in a predictive coding model.

One of the most common models used is called the vector space model. In the vector space model, each document is passed through a filter program in order to extract the base words of interest. For example, the filter program would remove "the," "and," "a," and other unimportant words. Each base word is then represented by an axis in the vector space for that document. The number of times the base word occurs in the document is the length of the vector along that axis. For example, assume that we have three documents that only contain three possible base words after being passed through the filter program: "privilege," "litigation," and "confidential." For each document, the number of times each of these words occur is counted and represented by a document-term matrix as seen below:



	Privilege	Litigation	Confidential
Document 1	1	0	3
Document 2	3	1	2
Document 3	0	3	3

Since each word in the entire document set is represented by an axis, hundreds, thousands, or even millions of axes may be needed to accurately represent the document.[10] Such a large number of axes may slow down the algorithm to unacceptable levels. Luckily, advanced programs are available that reduce the number of axes using a variety of methods. One way to do this is to designate an axis to each phrase or term instead of for an individual word. Another way is to designate an axis to a particular ordering of words or phrases. Yet another way is to remove the noninformative base words or terms that are not relevant to the litigation.

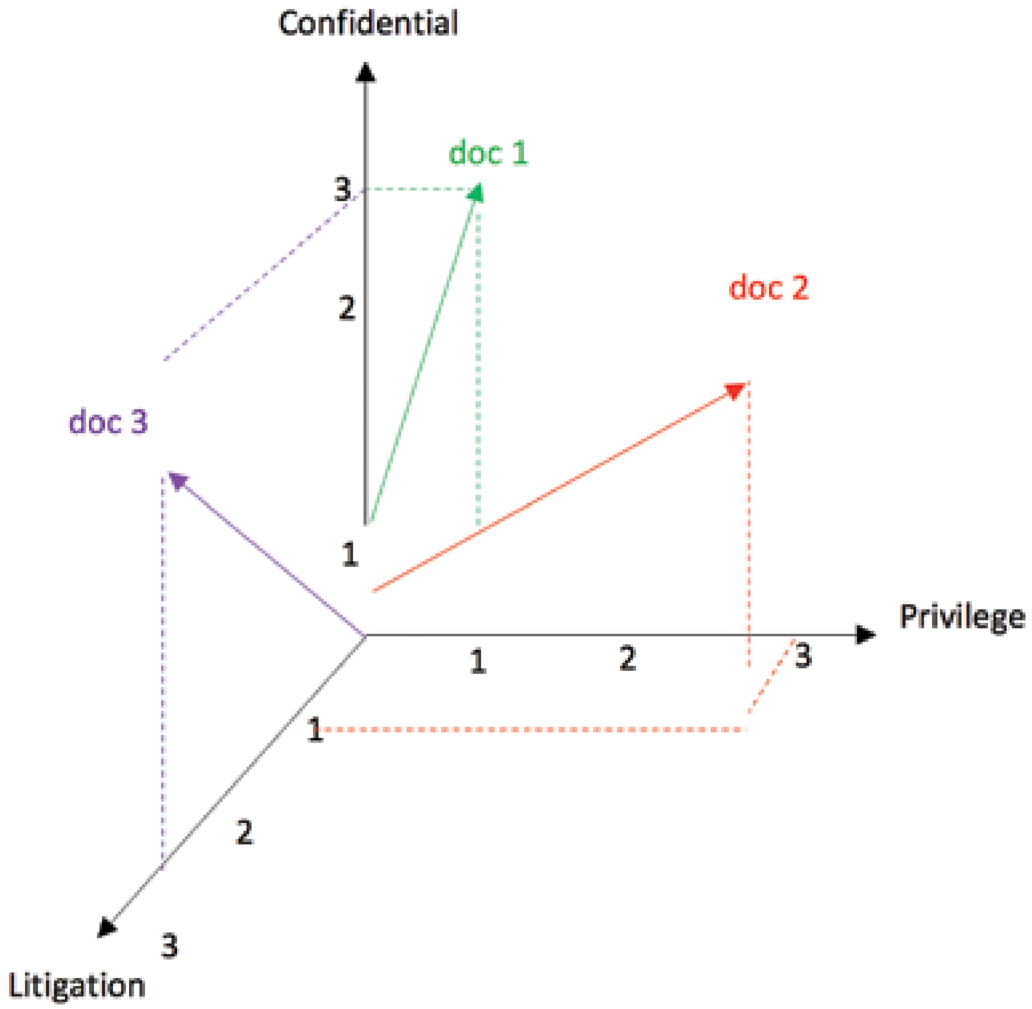
Once all the documents are represented by a vector, subject matter experts can manually review a smaller portion of the entire document set, known as the training set, and designate them into appropriate categories (e.g., relevant, non-relevant, privileged, etc.). This allows the computer to associate each of the vectors, which represent a document in the training set, to a category. Thereafter, the computer can analyze each of the remaining documents and automatically classify them according to the likelihood that the vector representing the uncategorized document is in the same category as a document in the training set. For each newly categorized document, a relevancy score (also called a similarity score) is computed. The relevancy score represents the likelihood that the document being reviewed matches a category to a similar document in the training set. Most programs allow users to designate a relevancy threshold (also called a confidence level) so that documents above the threshold are designated relevant and are produced, while documents below the threshold are designated not relevant.

There are a number of algorithms that a computer can use to determine how similar a new document is to a document in the training set. One of the simplest algorithms is the k-Nearest Neighbor (kNN) algorithm. The user designates a numeric value for k, and the algorithm identifies each document whose vector endpoint (tip of vector) is within k-units from the vector end-point of the current document. For example, if k=3 then the algorithm would return each document having a vector end-point that is within three units of the current document’s vector. The relevancy score is calculated according to the distance between the current document’s vector end-point to another neighboring document’s vector end-point. In other words, the closer the two end-points, the higher the relevancy score.

The advantage to this algorithm is that it is easy to understand. Litigants are more likely to use TAR when they understand how the applicable computer algorithms operate. Another advantage is that it can handle a large of number of documents if there is little word variety among the documents.

One disadvantage is that it is computationally expensive when the number of documents is large. The performance of TAR becomes degraded when there is a large number of dimensions for each vector (e.g., each document contains a large number of different words or terms). To overcome this, litigants should divide the documents into subsets of documents and analyze the subsets in batches, discarding all non-relevant mutually agreed upon words or terms. This can also be overcome by employing advanced algorithms that would automatically discard nonrelevant terms based on statistical probabilities.

Another popular algorithm used to determine the similarity of documents is the Latent Semantic Analysis (LSA).[11] The LSA algorithm assumes that there is a latent or hidden pattern of word occurrence across each document which can be uncovered by statistical techniques that estimate the structure of the document. The LSA algorithm also assumes that words having a similar meaning are more likely to appear in similar portions of the document. The LSA algorithm attempts to solve two problems with the vector space model: (1) the capacity for a word or phrase to have multiple meanings;[12] and (2) the existence of synonyms causing many different words or phrases to refer to the same thing.



After creating a document-term matrix as discussed above in the vector space model, the LSA algorithm assigns weights to certain words or terms and then preforms a well-known matrix-algebra method called Singular Value Composition (SVC) to derive multiple matrices representing similarities between (1) words and other words, (2) passages and words, and (3) passages and other passages. These matrices are used to construct a semantic space that represents each document as a vector. In some programs, the kNN algorithm is then used to find similarities between documents.

The advantages of the LSA algorithm is that it is more accurate than simple vector space models. The LSA algorithm performs well even when there are many grammatical errors in the text, such as when scanning paper documents into a computer using optical character recognition.

One disadvantage is that the LSA algorithm is computationally expensive and requires a large amount of memory, and is therefore not recommended for large document sets. The performance of the LSA algorithm improves as the number of dimensions increases, but then reaches a point of diminishing returns and starts to decrease. This disadvantage can be overcome by performing many test runs with varying dimensions in order to find the optimum number of dimensions.

[10] Unlike the three dimensional physical space represented by length, width, and height axis, a document's vector space normally has many more than three axis.

[11] This is also known as Latent Semantic Indexing.

[12] This is also known as polysemy.

TAR issues in-house counsel should address

Since there are many vendors in the information retrieval industry that employ different algorithms, it is important that in-house counsel understand which algorithm is appropriate for each case. Some issues that in-house counsel may need to address include:

- Should TAR be used at all? For example, some courts have allowed the parties to use traditional methods of document review instead of TAR because the low volume of relevant documents expected to be produced didn’t justify the costs associated with TAR.[13]
- What type of TAR will be used? There are many different methods and algorithms on the market today, each having certain advantages and disadvantages. It may be helpful to hire an expert to guide in-house counsel through the process of determining which vendor uses the appropriate algorithm for your case.
- What if the parties cannot agree on a TAR protocol to use? If the case is in federal court, or a foreign jurisdiction having a published opinion accepting the use of TAR, then it might be possible for one of the parties to propose a TAR protocol and ask the court to approve it after considering the other party’s objections. This may still be available in state courts, but there is little state case law for in-house counsel to reference.
- How large should the training set be? The training set needs to be large enough to (1) include as many different types of documents as possible, and (2) include a statistically representative sample size. Both parties should come to an agreement as to the size of the training set.
- If the TAR being used has an adjustable threshold for a relevancy score or confidence level, what score or confidence level should be applied? In most published cases addressing this issue, the confidence level was somewhere between 90 and 100 percent. For example, the parties in the Da Silva Moore case agreed to use a 95 percent confidence level. Sometimes the parties will lower the confidence level in order to make sure all relevant documents are found. However, in larger document sets the confidence level should be higher so that a manageable amount of documents are produced to the requesting party.
- How will the results of the TAR system be validated? The parties should agree to quantitative metrics for validating the documents identified by the TAR system. For example, after the training set has been analyzed by the program, a test set of documents should be analyzed. The test set should contain different documents than those in the training set, and both parties should agree as to how many documents should be included in the test set. The parties may also agree to review randomly selected documents identified by TAR to verify its accuracy.

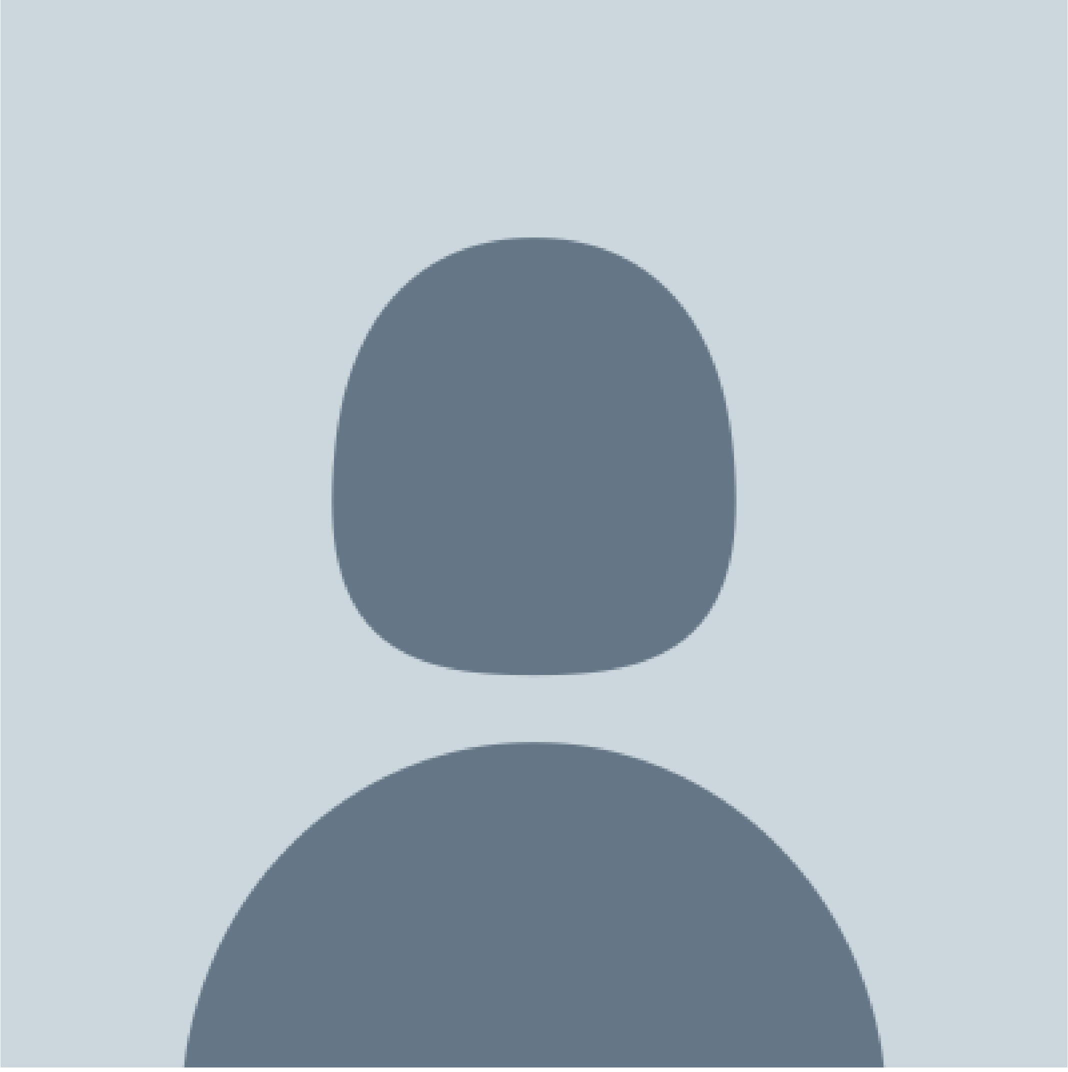
[13] *EORHB, Inc. v. HOA Holdings LLC,* 2013 WL 1960621 (Del. Ch., May 6, 2013).

Conclusion

Even though TAR is becoming accepted in US federal courts and in a few foreign jurisdictions, the adoption rate is slow in US state courts and in most foreign jurisdictions.

The slow adoption rate is likely the result of a lack of understanding as to how TAR operates. While TAR is considered to be a form of artificial intelligence, it is not designed to emulate how the human mind operates. Instead, TAR requires a human to provide an ample problem and the correct solution, so that the computer can then extrapolate the correct solution to examples it hasn’t yet seen. Therefore, TAR is dependent upon humans telling them how to perform its function correctly and is not some independently operating machine that thinks like a human.

Duncan Williams



Corporate Counsel of Intellectual Property

iHeartMedia

He has extensive education and experience protecting intellectual property in the fields of electrical/computer technologies, software, telecommunication systems, and wireless devices.