KPMG
cutting through complexity

ForensicFocus
September 2014

Some cautionary thoughts on "predictive coding"

TODAY'S REALITY:

Discovery has always been a part of litigation: it is the identification, review and exchange, by litigation parties, of information relevant to a case. Now that most information is in electronic format, discovery has now mostly become e-discovery. Lawyers need help managing vast amounts of electronic information as they assist their clients with litigation, investigations and regulatory inquiries.

The huge volumes of data and the increasing complexity of matters can make the cost of the e-discovery greater than the value in dispute. To help reduce costs and overcome some of the limitations of keyword searching, a promising technology emerged in the mid-2000s: concept searching (see sidebar). Concept-search technologies can discern similarities in the content and meaning of documents without any particular word needing to be present.

There were boosters and early adopters of concept searching technologies; but there was also a lot of caution: some of it justified. Many practitioners carried on with the same "old" methods: keyword searches and linear review[1] of the results. For many firms, this is still the order of the day.

The volumes of electronic information and the costs to clients continue to rise, once again forcing the question: can new technology do the job as well or better than existing methods and at a lower cost?

In recent years, software companies have been offering the next generation of concept-based search and classification technology: predictive coding or (a more commercially neutral phrase) technology-assisted review ("TAR").[2] The claim – the aspiration – is that, when properly designed and managed, TAR saves time and money and even achieves better results than a linear review of keyword hits. Studies have shown that this is indeed possible. But as with all complex tools, TAR can be misused: the tool itself can be badly handled and the process by which the tool is deployed can be poorly managed. For every success there are likely many wasted hours using TAR for something it is not designed to do.

EVOLUTION OF E-DISCOVERY APPROACHES

The methods employed to identify documents relevant to a matter have evolved to be more efficient and effective.

Keyword searches identify all documents containing user-specified words. Documents containing these words may be "false positives" – documents which contain a search term but which are not in fact relevant to the litigation. False positives increase the time and cost associated with discovery.

Concept searches aim to reduce false positives associated with keyword searches by discerning the conceptual content of a document. Concept indexing reflects how often words are used in and across documents, how close together they are, which words tend to be nearby and so on. Algorithms identify patterns in how words tend to cluster together ("co-occurrence").

In **technology-assisted review** (TAR), a subject-matter expert ("SME"), or a small team of SMEs, trains the concept-based technology to distinguish between wanted and unwanted documents according to the conceptual content of the documents. Over a series of sample-and-review phases, the SMEs essentially train the technology to do a better and better job of finding what is wanted and ignoring what is not.

[1] Linear review refers to document-by-document review, from start to finish. This approach requires significant investment of time and is susceptible to user fatigue and variable performance and reviewer judgment.

[2] The company Recommind has a Patent on "predictive coding" (US7933859).

Conferences and e-discovery websites now include, not just *how-to* sessions and articles relating to TAR, but also cautionary ones. Few seem to be facing up to what may be the greatest weakness: TAR analyzes the content of the document in order to make a simple Yes/No assessment – usually whether it meets a basic definition of relevance. There is no easy way to determine why TAR made the determination it did. If you want to know why the document was considered relevant or not, the document must be opened and read.

This type of intelligent machine assessment – specifically, making determinations at the document level – has been accepted in practice because document reviewers always, at the lowest level, assessed *documents* for relevance ("Produce or not?"). And this is fine and generally works well provided that the conceptual content of a document is fairly homogeneous, but many important documents are not like that. Documents can touch on multiple issues and the legal significance of the text can vary greatly from page to page, from paragraph to paragraph and even sentence to sentence. Assigning a score to an entire document is truly problematic where a particularly important phrase is buried on, say, page 17 of a 56-page document. If you dropped this phrase on its own into a blank document and then scored, TAR would give it a very strong score, but, lost on page 17 of a larger document, its "signal" is lost.[3]

As currently designed, TAR tools simply cannot do what we really need them to do, and what a careful human being can (still) do: detect significance deep in a document and pull that phrase or paragraph forward for special attention.

Some of TAR's inherent limitations which experienced practitioners are aware of include:

(1) TAR uses text and only text; it is hopeless with numbers and images;

(2) TAR does not and can never detect what *kind* of documents it is analyzing (such as letters, emails, memos, medical files, annual reports); and

(3) TAR knows nothing about dates, chronology, sequence, or special time periods.

So beware the hype. When used properly to group large volumes of documents by general content, TAR is extremely helpful. It is good at separating the Finance from the BBQ, the Annual Reports from the hotel bookings, but to expect it to detect – amidst the general population of business-related, case-relevant documents – those documents and, more importantly, those *paragraphs* that are truly interesting is a mistake.

[3] KPMG Canada is working with a Canadian software company, Porfiau, to develop a technology that can detect desired signals at the phrase level. *See* http://www.porfiau.com.

Contact us

Dominic Jaar
514-840-2262
djaar@kpmg.ca

Author of the report
David Sharpe
416-777-3738
davidsharpe@kpmg.ca

kpmg.ca/forensic