Predictive coding: a helpful new technology or merely the latest buzzword?

This blog was first published by PLC Construction on 30th July 2013 as part of Berwin Leighton Paisner's regular construction blog series.

Sara Paradisi, associate, Berwin Leighton Paisner LLP

There has been a lot of hype in the past year about predictive coding. Lawyers have prided themselves on being fantastically "in the moment" when proposing this advanced technology to clients, in the context of large disclosure exercises, with the promise that it will save time and money. Indeed, using this latest technology has become an effective way for law firms to set themselves apart from their competitors. However, clearly clients should not use it simply because it is the latest trend. It is critical that we, as legal advisers, and our clients understand how it works and when its use is appropriate, so that we can make an informed decision.

How does predictive coding work?

Predictive coding software takes all the documents related to an issue, ranks and tags them so that a human reviewer can look over the documents to confirm relevance. The beauty of this technology, but also its potential downfall, is that the computer uses the human decisions to learn what types of document are related to each issue, making its relevancy suggestions more accurate over time.

One can use the technology for different purposes, though the ultimate aim is to make a large scale disclosure exercise more manageable in terms of time and expense. For example, it enables one to organise large data sets of emails that have not already been structured. It can also give you a ranking of likely relevance down from a 100% to zero.

Potential downfalls

Sounds great, right? However, picture a disclosure exercise on a complex construction dispute about a process plant:

- You have collected more than two million documents from the client.
- The issues in dispute range from matters such as which documents form part of the contract, and the negotiations that led to those documents, to delay and defects (both in design and the workmanship). The case and therefore the documents contain very complicated technical information and associated terminology
- Plus, don't forget that, as it is a construction matter, you will have hundreds of thousands of construction-specific documents, such as spreadsheets, programmes, AutoCAD drawings, a large number of PDFs, photographs and videos.

So, how does one teach the software whether, for example, a spreadsheet of data (which includes no text), a video, a photograph or an AutoCAD drawing is related to one issue as opposed to another? Good question. The answer is, simply, that you can't. These are file types that the software's algorithm cannot read. In this regard, predictive coding software is no more sophisticated than old school keyword searching.

Moreover, with construction disputes, the technical issues in the case often develop over time, as the experts investigate and report. The emphasis placed on certain technical issues may shift throughout the duration of the case. Often, we don't know enough at the beginning of a case to appreciate the finer technical points of the dispute or the significance of certain factors, so we cannot teach the system what documents are relevant to one particular issue as opposed to another. We have to teach ourselves first – and then "train" the system, but do we always have sufficient time available to us to do this? Indeed, does the court/tribunal's timetable recognise the time required for undertaking this process effectively?

Overcoming these difficulties

From personal experience, I have found that the way to overcome these "construction-specific" difficulties is to use predictive coding software alongside other traditional e-disclosure tools such as keyword searches, de-duplication, date ranges and file filters. These help narrow down the pool of documents, making it easier to train the system. In addition, predictive coding can be very useful in addressing the issue of false positives, in particular keywords, and therefore significantly reduce the number of hours of review time and cost.

Moreover, although difficult to use in a detailed issues-based review, predictive coding software is very helpful at the beginning of a case, where there is often a tactical advantage to prioritising the most relevant or interesting documents for review.

Also, given the time implications of training the system, it is important to get the court/tribunal on board when making disclosure orders to ensure that timescales are realistic where predictive coding is to be employed.

When should you use predictive coding?

In deciding whether this technology is appropriate for a specific case, a lawyer should consider the following:

- Is the technical nature of the dispute likely to create problems? If the matter is very technical, it may be difficult to train the system properly.

- What are you trying to achieve? Is it an early case assessment? Is it a first pass relevance review or a detailed issue review?

- What are the timescales?

- Are the issues well defined or still under consideration?

- Is there anything unusual about the documents?

- Will there be a sufficiently senior person who is involved in the review process and able to invest time in training the software?

These are all complications that can reduce and limit the benefit of predictive coding software. As such, they should be highlighted to the client so it is not misled into thinking that predictive coding is a magic button that will quickly and neatly categorise its mountain of documents at a far lower cost than other technology-assisted review options.

Take it case by case

Having used this technology, it seems to me that it is too often simplified. It is not magic. It does not automatically find all documents related to an issue. It is important to understand that this software is only as good as the time invested in training it. The decision to use it should be made on a case by case basis.