An array of clusters and interlacing threads

(Deep dive into other technologies)

The technology available to assist investigations and document review is an ever-evolving landscape. No list of options is ever complete for very long, as there is a constant stream of new additions – some more effective than others! That being said, there are a number of tried and tested stalwarts with which any competition law (or other) practitioners would benefit from being familiar.

Concept searching and clustering - To-May-Do, To-Mah-To... what difference does it make?

While Predictive Coding or Active Learning fall into the category of supervised learning (i.e. the system learns continuously based on human input), clustering and conceptual searching are considered to be unsupervised learning. There are no iterations or options to refine the model – the system simply analyses the content of the documents it has been provided with and makes connections between words and phrases based on the usage within the data set.

This can be particularly useful in cartel investigations or real-time monitoring where employees use code words in order to collude and evade detection. Clustering and conceptual searching identifies code words which would not be usually used in day-to-day commercial communications with third parties or in particular sector/industry, or can simply ensure that industry-specific terms are correctly interpreted. The system has no pre-defined understanding of words so the meaning it assigns to them is purely contextual. In one data set, "chips" might refer to food and be associated with other edible terms such as fish or vinegar in the UK, or dips in the US; in data collected from an electronics firm, however, the word is more likely to relate to microchips and be linked to words like processor and memory. Where code words are being used, the system will automatically associate the words with whatever they are referring to, rather than any common meaning; for example, people often prefer to speak offline about illicit activities and so terms like sandwich or lunch become synonymous with the conversations they have over lunch (e.g. the "Gardening Club" in the Air Freight Forwarding Cartel). As the system had no other frame of reference for the meaning of the word "sandwich", in one of our previous cases, it assumed it was related to rate fixing because that was always the context in which it was used. This helped the legal or compliance team expand their searches and find more useful conversations.

Clustering is a great way to visualise this analysis. It groups documents by content, and then defines each of these groups by the four key topics represented in them. The

exact visualisation varies by tool – bubbles, or segments of a wheel are popular – but all have the common design of showing bigger sections for the groupings that contain the most documents, and allowing the user to drill down into sub-groupings to get more specific. Clustering is very useful in an investigation or at the early stages of a review. When faced with potentially millions of documents, it can be difficult to know where to start. Being able to visualise the key topic groups that appear in the case data can help investigators to spot unusual trends such as unexpectedly common topics, surprisingly uncommon topics, or unanticipated connections between seemingly unrelated topics. It can also be useful to identify large sets of documents that can either be prioritised or deprioritised for review, or perhaps excluded altogether.

Taking a different approach to leveraging the same process of extracting meaning from the words in the data set, conceptual searching allows competition lawyers or investigators to hone in on documents related to specific topics. While this looks similar to standard keyword searching, it is actually quite different and requires a very different mindset. Instead of searching for the exact words the culprits would have used (requiring the investigator to know those words), the search can now be about describing the topic they would have been discussing. Longer passages are better, as these allow the system to get a better understanding of what is being searched for. Rather than searching for the common LIBOR investigation keyword of "collude", a savvy investigator might apply a conceptual search illustrating the principle e.g. "we need to work together but keep it to ourselves. If we join forces and watch each other's backs, we could do very well here".

These searches have a number of applications, varying from answering specific questions to simply kick-starting an investigation or review by identifying key documents early on.

Reduction, reduction, reduction. The art of threading and near-duplicate identification

The first step before embarking on review of a large dataset is always to find defensible ways to reduce that volume for review. With email usage steadily increasing year on year, and the scope of investigations broadening, the mountain of potentially relevant emails grows taller and taller. Not only are there more emails to review, but anyone who has participated in such reviews will no doubt recall the frustration at having to read the same content over and over as they encounter an original email, then the reply to that email, then the reply to that one…. And so on.

This is where Email Threading can save a lot of time and effort. Email Threading is a process that ascertains which emails are part of the same conversation. Once they have been grouped together, it sorts and analyses them to work out which ones have unique content. These emails are referred to as "inclusive". Emails that are fully contained within later emails in the chain do not provide any unique content and are referred to as "non-inclusive". For example, scrolling down in a reply to an email usually includes the original email in full; in that case, the original email could be excluded without any reduction in the integrity of the review.

The most common example of an inclusive email is as described above, so the latest email in a chain would always be inclusive. If the conversation branches off into multiple strands (e.g. by forwarding an email to another recipient while continuing the original conversation, or through multiple replies to the same email) then the most recent email in each strand will be inclusive. There can also be inclusive emails in the middle of a chain, for example where an attachment is only found on that email, or where someone adds inline comments to an earlier email, preventing it matching the original email.

In all of these cases, the inclusive emails would be retained and only those which are entirely contained within one or more of the inclusive emails would be suppressed. This means the non-inclusive emails can be confidently removed from the review with no loss of information whilst keeping the cost of reviews down to a minimum.

While Email Threading works to reduce the volume of emails to be reviewed, Textual Near Deduplication can help to reduce the number of standalone non-email documents to be reviewed. (It's not generally recommended that this is applied to email attachments, as they will be factored in to Email Threading, and for most reviews it's advisable to keep emails and their attachments together.)

Textual Near-Deduplication compares the exact characters in the text of the document and calculates the percentage similarity to other documents in the set. It then groups documents with identical or similar content together so that they can either be reviewed at the same time, or some copies excluded from the review. Depending on the intended use, the minimum similarity threshold within each group can be set: if duplicates are to be excluded, this should be set to 100% as even 1% difference could be key. (It may seem unnecessary to run Textual Near-Deduplication at 100% for data that had duplicates removed in processing, but it is possible for duplicate content to exist in different documents e.g. a Word document, and a PDF of that document.) For use as a review accelerator, to group similar documents together, this threshold can be lowered, and it may be useful to experiment with different levels before determining the optimal level for the data set in question.

Once again, far-reaching potential but no universal panacea

As with any use of technology in competition law (or other) investigations or compliance, these features can be very helpful and save a lot of time and effort, but must be used appropriately. Clustering and concept searching are based on the system's interpretation of words and phrases from their usage, but it won't necessarily provide a definitive list of all topics in the data. Just as keyword searching runs the risk of being incomplete due to misspellings, synonyms and other inconsistencies, so too concept searching can be subject to confusion caused by misspellings, homonyms or simply not enough information to build up an understanding of the meaning of a term. It is very useful to seek out evidence of an activity, or answers to a particular question, but a lack of evidence should not be considered proof the activity did not occur or the answers don't exist.

Even Email Threading and Textual Near-Deduplication, while less fuzzy and based on solid information in the text, should still be used with intelligent caution. Thought should be given to whether the specifics of the data or issues mean that removing earlier emails in the chain is inadvisable; consideration should also be given to the most appropriate threshold for Textual Near-Deduplication.

There are many benefits to be had in terms of efficiency, cost, and even reduction of risk, when using technology to streamline review and identify crucial issues or documents. The key is to always consult with your E-Discovery vendor or competition lawyers to ensure you are using them as part of a robust and defensible workflow.

More about the Authors

Rebecca Cronin is Legility's Director of Technical Solutions based in London.

A Relativity Master and experienced e-disclosure project manager, Rebecca joined Legility as a Senior Project Manager in 2014 after 7 years with KPMG. With a degree in Computer Engineering and a masters in Digital Forensics, Rebecca has worked with numerous review tools over the years, but now primarily focuses on Relativity including developing and managing processes, and advising clients on workflow and strategy. With extensive knowledge of linear review, analytics, and technology-assisted review workflows, Rebecca is an expert in understanding a client's needs and tailoring a solution to them. As Director of Technical Solutions, Rebecca is responsible for identifying and evaluating both new products and updates to existing technologies.

Marie Leppard is a partner at Euclid Law, The Competition Law Firm. Before joining Euclid Law, Marie was a senior associate at Clifford Chance's antitrust practice (London and Paris).

Marie assists clients on French, UK and EU antitrust investigations, complex multi-jurisdictional mergers and abuse of dominance cases. Marie's practice focuses mainly on cartel investigations. Marie has worked on numerous high-profile cross-border cartel investigations before the European Commission, the CMA, the FCA and the US Department of Justice.

Marie has also vast experience in providing clients with compliance framework and training as well as advising on the use of the latest technologies and artificial intelligence in internal audits and cartel investigations.