

The problem of analysing unstructured data

The limits of computers in electronic discovery
Andrew Harbison & Pearse Ryan

The problem of unstructured data

As anyone even peripherally involved in litigation can tell you, the advent of modern computer systems in recent times has greatly increased the quantity and complexity of the discovery process. Even as relatively recently as 20 years ago a complex case would have involved no more than a few hundred documents. Nowadays, modern computer systems ensure that even routine cases can require the analysis of gigabytes of documentary data.

This does not sound much until you realise that printing 20 gigabytes of data would produce a stack of A4 paper over a kilometre high. **The use of terms such as ‘gigabytes’ and ‘terabytes’ sometime leads us to under-estimate just how much data is stored on modern computer systems. All of Shakespeare’s works can be stored in less than 10 megabytes** (one-hundredth of a gigabyte), yet even junior staff members regularly carry around gigabytes of data in their pockets stored on PDAs, Blackberries, and pen drives, not to mention the additional gigabytes stored in e-mail accounts, computers and backups.

What complicates the situation even further is that while computers are exceptionally useful in accumulating and storing all this material, they are of limited help in searching and analysing it when the need arises. This is because computers, at their core are still mathematical devices, essentially ‘number crunchers’, and as such they are much better at analysing structured data, the kind of material found in databases, lists and spreadsheets, than they are the unstructured data, the documents, e-mails, graphics, sound, video and other files which makes up the vast majority of material on modern computer systems.

Unstructured data

It is estimated that around 80 to 85 percent of the useful data stored on modern computer systems is unstructured in format. Of course, most people already possess a computer optimised for analysing such data, namely their brains. Unfortunately, while we are very good at dealing with unstructured data in small amounts, we cannot easily handle the vast quantities of data produced in even moderately sized legal cases today. We need electronic assistance to manage the vast amounts of information delivered to us.

A further complicating factor is the existence of different problems in analysing different kinds of unstructured data. Specialists in the field often divide unstructured data types into two broad groups - textual and non-textual data. Textual data, obviously enough, includes

all the documents, e-mails, text files and chat files that accumulate on computers. Non-textual data, on the other hand, comprises everything else - graphics, sound, movies etc. Of course not even this simple split is straightforward. For example, is a Powerpoint presentation textual or non-textual data? If sound or graphics are embedded in a Word document or email, is it textual or non-textual data? Putting these problems of categorisation aside, textual and non-textual data raise different problems for those attempting to analyse them.

Textual data analysis

In the textual case there is a key problem of context. The classic example often given is the difference between the statements that “John rides in a mustang” and “John rides on a mustang”.

A human analyst will see a great deal of difference between these two sentences. Our experience adds enormously to our understanding of both. We know, for a start, that the first statement refers to a car, the second to a horse. But we will also understand that in the first statement John is a man, and he is probably in the United States, because Ford Mustangs are not sold in large numbers outside the US. In the second, we may consider that the event might have occurred in the US as the descriptive term is generally associated with that country, and a long time ago, as there are not many wild horses left in the US. It might even occur to us reading one of the two sentences that, because the O and I keys are beside one another on a standard keyboard, there could have been a typographical error and the other sentence may be the correct one.

The human brain picks up all of this data almost instantaneously – our understanding is implicit. Computers cannot deal with implicit information and have to be told how to understand it. Consequently they deal with this ‘tacit’ information very badly, if at all.

Complicating the matter still further is the fact that the writing style, sentence structure and vocabulary used in formal documents are very different to those used in e-mails, which are in turn different to those used in text messaging. **Similarly, we do not often appreciate that we do not speak one language but a number of closely related yet subtly different ones, depending on the context.**¹ We speak to our friends, children, and bosses, all in very different ways, our brains usually handling the ‘translations’ effortlessly. But for computers it can all be an impenetrable problem.

Classically, the principal tool used to analyse textual data has been the keyword or keyphrase search. In the past data analysts and lawyers faced with a mass of information have prepared long lists of possibly relevant keywords and have used computers to search for these in the mass of data. This method is far from perfect. If the set of search terms is too narrow it can miss vital information, if it is too broad the resulting set of ‘hits’ can contain large numbers of totally irrelevant ‘false positives’.

¹ It is worth considering that in many other cultures, the language of personal dealings are entirely different. Business communication may be in English, while personal communication will be in the Lingua Franca.

Modern search tools have improved things somewhat. Computerised thesauruses allow us to search for synonyms and homonyms without having to explicitly set out every possible variation. Other tools allow for 'stemming' - for example, in Lexis Nexis putting in the term 'run+' will cause the engine to search for 'run', 'runs', 'running', 'runner', and so on. The disadvantage is, of course, that these techniques increase the number of effective search terms and, as a consequence, the number of false positives. Other techniques allow for logical searching, for example, you might be able to set the search program to look for documents containing the words 'John and 'rides' within 5 words of 'mustang' in all documents written during 2008. This kind of logical searching can be very effective once the user has had some practice with it.

These tools have simplified the process somewhat, but they still have not solved the problem of computers' inability to deal with tacit and context-based information. Researchers are making efforts to deal with this problem, however.

Some advances are being made in applying artificial intelligence systems to the problem. Some sophisticated systems now contain 'neural networks' that allow the user to 'teach' a computer to recognise simple tacit linkages between words and expressions. So, for example, where the investigation involves a fraud, the search engine can be fed a number of fraud manuals and books to allow it to better recognise obvious linkages between words and concepts. Furthermore, as more documents are analysed the artificial intelligence will (in theory) become better and better at identifying linkages. Unfortunately teaching a neural net takes time, so at present the technique is only practical for the largest cases. Additionally, at present even the most sophisticated artificial intelligence has only a small fraction of the analytical ability of the most inexperienced legal trainee.

At present, we must conclude that text analysis technology may be better at data reduction than actual data analysis. If keyword searching does nothing else, it can identify files that definitely do not contain relevant information, which is at least a start. Advances are being made, and more sophisticated systems will eventually become available, but it will be some years before we can expect systems that will be able to substantially reduce the analytical burden. Until then, our best tool will remain the Mark 1 Human Brain.

Non-textual data analysis

Where data is in non textual form, the problem for the computer usually begins with converting it into a form that can be searched at all.² Again, the human brain is a computer that handles this problem very well, with a 'subprocessor' (the Visual Cortex) optimised for this purpose. Again, its defect is that it cannot handle large amounts of this data.

Perhaps the most basic conversion problem is bringing text stored in 'graphical' formats, such as TIF or PDF files, or even printed documents, into a form that can be searched on a computer. In fact, most large legal firms now possess optical character recognition (OCR) systems, that can perform this activity reliably. Tests on modern OCR systems

² Of course, once the data is converted to textual format, the analyst still faces the problems with textual data discussed already.

have shown that they can convert printed text to computer format as well as or better than any human. OCR systems still have problems with handwriting, however, particularly ‘cursive’ or script lettering. But, then again, how many times have you been unable to read someone else’s handwriting?

Sound files are also becoming less of a problem. Modern voice and language recognition systems are becoming more and more capable of converting voice recordings to text. We have reached the point that clearly modulated speech can be reliably converted to text by electronic means. Additionally, because many of the language recognition systems are based on neural net artificial intelligences, they tend to become better at recognising individual’s voices as they process the data. Of course, the systems available now tend to be designed for languages common in Western business circles. While you will find analytical tools for English, German and Japanese easily enough, Thai or Croatian language systems could be more of a problem. **Language recognition systems can also have problems with heavy accents or unusual grammar, but so too can humans.** Sound analysis systems still have problems, however, particularly with inflection – anger, distress, doubt, irony. As with written material, the human brain relies on cues which computers are incapable of detecting unless specifically programmed to do so and preparing such programs is far from straightforward. We know all too well that humans cannot always detect anger or distress in the voices they hear. It is hardly surprising, then, that computers too have problems.

Dealing with graphics or movie files where there is no text or language is a more complicated problem. Processing a graphical image is an enormously complicated task, that, in our own brains, requires a significant proportion of the overall ‘processing power’ – and still doesn’t always work. Hence the existence of optical illusions. Perhaps the area where most progress is being made is in the area of illegal or pornographic photographic imagery. Most systems for identifying such imagery are artificial intelligence based, and some of the market leading systems are highly effective in this role.

The problem with these systems is that they tend to be ‘idiots savants’. They can excel within their narrow specialisation, but are practically useless outside it. Within it they can also still provide ‘false positive’ and ‘false negative’ findings and, as ever with computers, they do not have the ability to deal with information outside their own narrow programming. For instance, **the manufacturers of a widely used and otherwise reliable pornography scanning application sheepishly admit that, despite their best efforts, the system still identifies pictures of hedgehogs as hard-core pornography.** The neural net built into the application misinterprets the curves and tones in the photos as being consistent with pornographic material. Despite the fact that it is one of the most sophisticated artificial intelligences commercially available, the system simply does not have the depth of interpretation to spot a hedgehog when it ‘sees’ one.

A problem for litigators alone?

Of course, the problems associated with unstructured data are not merely an issue for litigators. Data protection legislation in most European countries require that organisations disclose all information they hold concerning individuals on request in response to requests from those individuals. It is difficult to do this if you don’t know

what information is held in the first place, nor where such data is likely to be held. For example, pursuant to Section 4 of the Irish Data Protection Acts 1988 and 2003 (DPA), an individual has a right to obtain a copy of any information relating to them kept on a computer or in a structured manual filing system, by any person or organisation, regardless of when the data was created. This right is subject to certain narrowly defined exemptions.

The key term for understanding access to manual data under the DPA is “relevant filing system”, defined in the DPA to mean "any set of information relating to individuals to the extent that, although the information is not processed by means of equipment operating automatically in response to instructions given for that purpose, the set is structured, either by reference to individuals or by reference to criteria relating to individuals, in such a way that specific information relating to a particular individual is readily accessible".

The Irish Data Protection Commissioner has recommended the following criteria in determining if manual data is part of a "relevant filing system" and therefore subject to the DPA.

- 1 Is the personal data part of a set (i.e. a regular filing system within a particular organisation which the organisation maintains for conducting its business)?
- 2 Is the set structured by reference to individuals or by reference to criteria relating to individuals. Guidance from the Commissioner indicates that "... if a file exists with a person's name or ID number on it this meets the criterion. If the file does not have a name on it but has sub-divisions with a name or ID, and the file title indicates that it contains personal data e.g. record of sick absences then this would also meet the criterion. If the file has a subject matter on its title, rather than a person's name, and it is known that the subject matter relates to individuals, then it meets the criterion - e.g. a file concerning a competition for promotion within a workplace".
- 3 Is the data readily accessible? Again guidance from the Commissioner indicates that "... if files are archived and are not used for decision-making as part of the day to day operations of the organisation, and retrieval involves disproportionate effort (or perhaps even cost where a storage company is used), then, the data could be said to be not readily accessible. In such a circumstance, the data subject would need to be able to identify particular data by file reference or date so that on a reasonable view of things the data could be said to be readily accessible".

In addition, the Commissioner has indicated that where in the course of searching for electronically stored documents by reference to an individual a data controller finds a reference number for a manual file, that manual file should be considered to form part of a relevant filing system.

Similarly, organisations subject to freedom of information type statutory regimes can find it extremely difficult, disruptive and costly to meet their obligations in a reasonable time frame. In addition, other regulations such as the Basel II rules on corporate liquidity, and the US Sarbanes Oxley Act, have similar type data disclosure requirements.

The problem does not stop there. Many US states and nation states require companies to advise customers when their personal information has been compromised in a data breach, such as when laptop computers go missing, or hackers access computer systems. When structured information, such as credit card databases or client lists are stolen, identifying the people affected is typically straightforward. But when other types of information is stolen, assembling such a list can be highly problematic.

The solution?

The most effective current solution for problems associated with analysis of data, both structured and unstructured is proper document management. In the US most large companies have implemented, or are beginning to implement, effective document management systems, largely as a response to concerns over litigation costs, security, and the other issues discussed in this article. Law firms have begun to provide services in litigation preparedness, typically ensuring that key documents are readily recoverable if the need arises (and that unnecessary documents which might cause unforeseen problems later are properly and appropriately disposed of).

A detailed discussion of document management is outside the scope of this article, but you can expect it to become a critical topic in many European board rooms in the near future. It is worth pointing out however, that most common computer applications, such as Microsoft Office, already contain a great deal of embedded management data (called Metadata) as standard, and they allow users to add yet more if necessary. These can provide the first blocks in an effective document management infrastructure.

It is obvious that it generally takes less time to complete a discovery based on a well-organised paper-filing system than it does from a disorganised one. The same is true of the computerised filing system. The primary difference is that computerised filing systems are orders of magnitude larger, meaning that corresponding costs and potential cost savings are proportionally larger. The key to speedy electronic discovery is having the documents properly filed and sorted in the first place, essentially to add some structure to the mass of unstructured data. If, however, this has not been done, given the state of the tools available today, the lawyer's best hope is to use computers to reduce the amount of data that will have to be analysed in the old fashioned way.

Pearse Ryan is a partner in the Technology & Life Sciences Group at Arthur Cox, Dublin.
www.arthurcox.com

Andrew Harbison is Director in the Forensic & Investigation Services Group at Grant Thornton, Dublin.
This article was originally published in *Computers and Law* (www.scl.org)