



Near-duplicate Detection in Electronic and OCR Collections

Introduction

As the market leading vendor in Canada for litigation support services, Commonwealth Legal Inc. has made it a top priority to research and implement cutting edge litigation support technologies into its daily practice. Not only does litigation support technology have to have a potential and beneficial use in the market, but it also must deliver on the promises that the concept entails. In the last year, Commonwealth Legal has been excited to test and execute a large project utilizing Equivio's revolutionary "near-duplicate" detection technology.

There May be an Elephant in the Room, and You Didn't Even Know

Near-duplicates are files with small differences. They are distinct from exact duplicates which are exact copies of a file. Examples of near-duplicate files include files with a slight textual differences, different formatting (such as bold or italicized fonts), or different file types (such as Word and PDF). Files with textual differences are by far the most common form of near-duplicates, and the most pressing business need. This scenario is also the most challenging from the technology point of view. Near-duplicates are especially common in email, business templates, such as proposals, customer letters, and contracts, and forms, such as purchase or travel requests.



There has been a very significant increase in near-duplicates over recent years. This trend may be attributed to the falling cost of storage and the widespread use of email. Distribution of information is much easier and, within an organization, many more people are involved in document review processes. In two recent cases, we have found 34.4% and 27.9% of near-duplicates. It is very important to remember that these numbers are in addition to exact duplicates. The numbers we have been getting are in line with Equivio's statistics which show that near-duplicate rates are typically in the 25-50% range.

Interestingly, most law firms and corporate legal departments are not even aware that they have a near-duplicates problem. Even those that recognize there is an issue have no idea of the scope of the problem. The main reason for this is that until Equivio arrived on the scene, it was for all intents and purposes impossible to find near-duplicates without a lengthy and costly manual review, or even to assess how many near-duplicates were hidden within a discovery collection. Given the levels of near-duplicates, they represent probably the largest hidden cost factor in the entire discovery and review flow. With the emergence of Equivio's technology, we are now in a position to eliminate this huge and unnecessary cost burden.

Not all that's Similar is a Near-duplicate

Within the industry, there is some confusion as to what exactly constitutes a near-duplicate. In this context, it is important to distinguish near-duplicate detection from classification and conceptual search technology. All these technologies are linked in that they add structure to unstructured data sets. However, beyond this, these are very different technologies. Classification and conceptual search tools group related documents, that is, documents that relate to the same subject. Near-duplicate technology groups similar documents, that is, documents with small differences in content or formatting. For example, consider three very short documents:

- Doc1:** O Romeo, Romeo! wherefore art thou Romeo?
- Doc2:** O Romeo, Romeo! wherefore art thou hiding Romeo?
- Doc3:** Raphael paints wisdom, Shakespeare writes it

Documents 1, 2 and 3 are all related, in that they relate to the works of William Shakespeare. As such, they would all be retrieved by a conceptual search engine with the search parameter "Shakespeare". Similarly, a classification tool would typically group these three documents together.

However, only Docs 1 and 2 qualify as near-duplicates. Docs 1 and 2 are similar in the sense that they differ by just one word. Note that within the search results for "Shakespeare", the near-duplicates add an additional layer of structure by grouping Docs 1 and 2 together.



It should also be noted that search and near-duplicate detection are complementary technologies. The same applies to classification technologies. Equivio's near-duplicate sets provide additional structure to the documents retrieved by a search engine, or grouped under a classification category.

Back at the Office

Equivio is relatively non-intrusive in the litigation work and data flow. Essentially, Equivio creates near-duplicate metadata for each document. This metadata, which is recorded in the Equivio database, indicates the near-duplicate set (aka EquiSet) to which a document belongs, as well as the pivot document in the EquiSet and the percentage similarity of the document vis-à-vis the pivot. The pivot document is the document that should be reviewed first in the EquiSet. The pivot document selection criterion can be set by the user. At Commonwealth, we use the largest document in each EquiSet as the pivot. Equivio provides an extract utility to facilitate loading of the near-duplicate metadata into your review system, whether it be CT Summation, FTI Ringtail, iCONNECT, Concordance, etc.

In the work flow, the fundamental impact of Equivio is that reviewers are able to work with sets of documents rather than individual documents. Each set of near-duplicates is grouped together and can be reviewed together, rather than randomly as is currently the case. This "set-centric paradigm" enables a much more systematic review effort, reducing the cost and time of document review. It also enhances the quality of that review.

Let's see how this fits into the work flow. We start out with an unstructured collection of documents, which often has already been loaded into the review environment. Equivio identifies the near-duplicates and arranges them into sets. The Equivio metadata is imported into the review tool. The near-duplicate metadata allows the attorney to sort documents by EquiSet. The user can then deal with all the documents in an EquiSet in a coherent manner. The user would start by reading the pivot document in the EquiSet. In many cases, after reading the pivot document, and finding it to have no bearing on the case, the individual may decide that the rest of the documents in the EquiSet can be skipped.

If, however, the pivot is relevant and important, the attorney can zoom in to review the remaining documents in the EquiSet. But there is no need to read each document from scratch. Using a compare utility, the attorney can simply review the differences of each document vis-à-vis the pivot document. This is a lot faster than reading each document from beginning to end. It's also a lot more effective because there is no chance of critical differences being missed. Finally, Equivio ensures that near-duplicates can be treated consistently – for example, when coding documents as privileged, responsive and so on. This is a key benefit helping to ensure the quality of the review effort.



One of the things that users like in Equivio is that it's very intuitive. When Equivio groups two documents together in an EquiSet and says that they are 95% similar, the user can know that there will be just a few words different between these documents. This is concrete, clear and objective, and gives people a lot of confidence when they are working with the tool.

The ROI on near-duplicate detection is also very concrete. Industry studies show that the review cost for the law firm is around \$3-4 per document. Equivio provides a very useful ROI calculator which customers can use to quantify the savings that near-duplicate detection can generate on a given case.

Using Equivio in OCR Scenarios

The first use scenario that occurs to most people with Equivio is collections of native electronic documents. This is definitely a "killer app" for Equivio. Interestingly, however, we have also been able to leverage significant value from Equivio in OCR scenarios.

Commonwealth Legal has been involved in a large project spanning rounds of document indexing for the last five years. The project entailed two levels of coding, objective and the much costlier subjective attorney review. As the entire collection was paper based and collected from multiple sources, the challenge has always been to mitigate the volume of duplicate or near-duplicate documents passing on to the subjective review. This process, in the past, involved a long and intensive manual review based on duplicates identified by identical objective coding fields. As the rules for what was to be considered a duplicate document seemed to blur past the exact duplicate criteria, Equivio's "near-duplicate" detection technology seemed to be the perfect solution.

Commonwealth Legal performed a pilot on client data using Equivio's analysis of OCR (Optical Character Recognition) text files with startling results. A manual review revealed that the "EquiSets", Equivio's grouping of "near-duplicate" documents, satisfied our manual criteria for grouping like documents beyond expectation. Given the nature of OCR and its inconsistent interpretation of scanned images, the Equivio results were impressive to say the least. Based on the pilot, the client agreed to proceed with the de-duplication process using the Equivio results as the backbone for the subjective review vetting process. Full processing over the entire collection yielded the same percentage of near-duplicate "EquiSets" that have, historically, been found through the more labour intensive manual reviews.

The benefits of Equivio's technology were twofold. As a vendor, Commonwealth Legal was provided a means to expedite its workflow by isolating documents as unequivocally unique allowing them to pass through both phases of indexing. The client also obtained a valuable tool during the review process as the "EquiSets" generated by Equivio will also be provided



in the final database to assist with the locating, and analysis of all similar document related to a particular issue.

Market Challenges

Equivio offers a unique and innovative solution to a very real pain point. People, though, tend not to be fully aware of this pain. From the marketing point of view, this creates an interesting situation. It's a chicken and the egg problem: until you have the solution, you don't know how much you need it. As a result, customers need to see it in action before they buy in. But from our experience, once they see it, it's very compelling. Despite the challenges that this creates, Commonwealth Legal sees this as part of our role in the market: educating our customers in applying new and innovative technologies that can make them more successful.

About Equivio

Equivio enables the management of data redundancy in content-centric business processes. Equivio's technology zooms in on unique data, allowing you to read less, think more, win big™.

With products for grouping near-duplicates, capturing email threads and automating prioritization, Equivio powers a broad range of business applications, including e-discovery, records management, email archiving, data retention and intelligence. To learn more about winning with Equivio, email info@equivio.com or visit www.equivio.com.

Equivio Inc.
5260G Nicholson Lane, Suite 150
Kensington, MD 20895
USA

www.equivio.com