TechnoFeature™

Practice management and technology articles written by experts.

Review: Equivio: Near-Duplicate Ediscovery Technology

By Brett Burney

Equivio

www.equivio.com

TechnoScore: 4.8

1 = Lowest Possible Score; 5 = Highest Possible Score

INTRODUCTION

De-duplication is a fancy-sounding term that's trumpeted around the water coolers of eDiscovery.

In theory, the concept is simple — take out all the duplicate documents and you'll significantly reduce the corpus of information that you'll have to review during discovery.

In reality, de-duplication poses some awkward hurdles when you dig down into the practical bowels of the process. Which is why Equivio and their "near-duplicate" approach has struck a resounding chord in the litigation technology industry.

DE-DUPE, DE-DUPE, IT'S OFF TO WORK WE GO

With the overwhelming amount of email and electronic information that must be reviewed in today's civil litigation arena, any tool or process that can cull and winnow down the universe of data is most welcome. Common methods of filtering by date ranges or keywords are effective, but those methods instinctively drag in every file consistent with the search criteria. If ten exact email messages (i.e. sitting in the Inboxes of ten different recipients) match the keyword terms, then someone is probably going to have to look at all ten of those documents.

De-duplication attempts to alleviate the burden of unnecessary duplicate files showing up in a review platform. Fortunately, each electronic file is unique, which means each file can be assigned a unique alphanumeric value called a "hash." Hash values can be generated quickly and then compared to other hash values. If two hash values are the same, that means the two electronic files are duplicates.

The problem with hash technology (at least in the legal discovery world), is that it is so literal. If one comma is added in a Word file, or the name of a file is changed, the hash value will be completely different. Those minor changes may not be relevant for the issues the reviewers are looking for, but both the before and after files would appear in the review platform because they are not considered duplicates any longer.

Equivio and their "near-duplicate" approach has struck a resounding chord in the litigation technology industry.

DON'T COME NEAR MY DUPLICATOR

Equivio addresses this problem by finding what it calls "near-duplicates." A strict duplicate is an exact copy of a file, but two near-duplicates are merely "mostly similar" to each other. For example, near-duplicates would only have slight changes between them like a few different words, different fonts, or different file types (i.e. a Microsoft Word document that was converted to a PDF).

It may take some time to get familiar with the concept of near-duplicates, but the hypothesis makes

(Continued on next page)

a lot of sense in the digital world we all now live in. For example, if one person adds quotation marks to a phrase in a Microsoft Word document and then saves that file with a slightly different filename, most of us would view those two drafts as extremely similar. We would probably consider them duplicates.

Similarly, we're all familiar with an email "conversation." When we pass email messages back and forth, most of us include the text bodies of previous messages so that everyone involved can go back and read the past conversation if necessary. If we needed to review this email conversation, it would make sense to only read the entire conversation from the last email of the series. We don't need to read each and every email volley along the way.

And one last example — if a document exists as a Microsoft Word file, a converted PDF file, and a scanned TIFF image, all of those iterations of the document would be considered duplicates. However, they would all have unique hash values and thus they would not be considered duplicates according to the traditional hash function.

My testing suggests that Equivio lives up to the hype, which explains its rapid adoption rate in the eDiscovery marketplace.

These are the problems Equivio seeks to solve. My testing suggests that Equivio lives up to the hype, which explains its rapid adoption rate in the eDiscovery marketplace.

WARNING: TECHNICAL SOFTWARE AHEAD

Equivio is not considered an end-user product, although law firms and corporate law departments are certainly free to use the software. Equivio considers itself a software company — it does not offer a service. Its clientele consists of eDiscovery service providers who successfully integrate Equivio into various processing and review tools.

Some of these vendors include kCura, iConect, CaseLogistix, and Ringtail. The idea is that while you process your eDiscovery data using your preferred application, you can also apply Equivio's near-duplicate technology.

Most vendors can already provide strict de-duplication. But the number of duplicates discovered may be very low as described above. Straight de-duplication will nevertheless weed out some of the files after which Equivio can take care of the rest. Equivio boldly estimates that its technology can usually eliminate an additional 20-50% of the document population.

Equivio requires either Window XP or Server 2003, although the company just released a version for Vista. Equivio works on top of either a MySQL, Oracle, or Microsoft SQL Server. If you're not comfortable installing and managing a database, you'll need to recruit some help. You will also need to set up an ODBC connection for each database that's created.

Once the database is set up properly, you can launch the Equivio Near Duplicates software which is otherwise known as the "Equivio Core" — this is where the grunt work gets done.

Next you simply point Equivio to your folder of native files, text files, or PDFs. Equivio cannot read images (such as scanned TIFFs), but it most certainly can analyze the accompanying OCR text files.

Equivio also works incrementally, so if you receive subsequent document collections, you can always process the new batch and add it to the documents already done.

In the Equivio Core process, you can set several options including the EquiLevel which is the minimal level of resemblance for two documents to be considered as near-duplicates. Equivio recommends starting with a 60% EquiLevel, but that can easily be changed if the results are not helpful. Once you click "Go," the documents are analyzed and ready for the next step called Equivio Extract.

Equivio Extract has an almost wizard-like interface that walks you through a vast array of options for the grouping and extraction process. The Extract

(Continued on next page)

mode is the underpinning of Equivio's technology. Instead of a document-centric view of comparing two or more files together, Equivio gives you a set-centric view by grouping documents into EquiSets.

Each EquiSet has a "Pivot Document" that best represents all the documents in that set. You can set some preferences for how Equivio determines a Pivot Document, although it's a good practice to select "Maximum Word Count" so that you know the Pivot Document will be as comprehensive as possible.

The ultimate goal of the Equivio Extract utility is to create a CSV delimited text file that you can load into a document review platform such as CT Summation, Concordance, or any number of platforms offered by eDiscovery vendors. The CSV file lists each document and indicates whether or not it is a Pivot Document. If it's not a Pivot Document, then the CSV file indicates whether it is a straight duplicate, or it gives a similarity percentage as compared to the set's Pivot Document.

Each document is also assigned an EquiSet number. All near-duplicate documents (as determined by the EquiLevel set in the Equivio Core process) are given the same EquiSet number.

Lastly, the CSV file also includes an EquiSort column that assigns a unique, sortable ID to each document so that reviewers can use that column to sort the near-duplicates properly in their review platform.

REVIEWING YOUR DUPES

So how does all this help a reviewer when the documents are finally loaded into a review tool? With Equivio's "set-centric" review theory, a reviewer can sort the documents according to the EquiSort field, and then look at the Pivot Document of each EquiSet. If that Pivot Document is clearly irrelevant, then you can safely assume that the other documents in the EquiSet are irrelevant without having to read through each one of them.

If the Pivot Document is relevant, then the reviewer can immediately skim through the remaining documents in the EquiSet, focusing only on the specific differences in each document. You don't have to jump around the database looking for similar documents, you can focus all your energy on comparing the similar documents right in front of you.

One of the best ways to compare documents is to use a document comparison tool. You're free to use one of your own, but Equivio does provide a bare-bones tool that you can use directly from the software.

HIGHS AND LOWS

There is a lot going on with Equivio, and this review only scratches the surface of the features. On the other hand, I like that Equivio only concentrates on the one major problem of grouping similar documents together — it doesn't offer a review platform or anything else.

Equivio boldly estimates that its technology can usually eliminate an additional 20-50% of the document population.

Equivio is not for the technically faint of heart — if you're not comfortable finding your way around MySQL databases or CSV files, you should definitely leave the Equivio implementation to someone else. I wish Equivio could offer a simpler tool for law firms and lawyers to use, but perhaps the company prefers to focus on the large eDiscovery vendors in the marketplace.

I also found the Equivio Core process to be brutal on my computer's processor. Granted, I was not using a very high-powered machine to run the software, but I would highly recommend a powerful computer for the software that can sit over on the side to do its work.

TECHNICAL SUPPORT

If you do decide to take the plunge with Equivio yourself, you will fortunately have a solid amount of documentation and tech support available to you. I

(Continued on next page)

was very impressed with the User Guide and other support documents provided, and tech support was absolutely wonderful in walking me through some questions I had.

CONCLUSION

Duplicate files aren't going anywhere — they will continue to haunt our eDiscovery projects for many years to come. Fortunately, Equivio (and its partners) can help guide us through the confusion.

ABOUT THE AUTHOR

Brett Burney is the Principal of Burney Consultants LLC where he focuses his time on bridging the chasm between the legal and technical frontiers of electronic discovery. Burney Consultants also provides exceptional support for litigation databases, document review projects, and trial technology. Visit his blog.

Contact Brett:
E: burney@burneyconsultants.com

About TechnoFeature
Published on Tuesdays, *TechnoFeature* is a weekly newsletter containing in-depth articles written by leading legal technology and practice management experts, many of whom have become "household names" in the legal profession. Most of these articles are TechnoLawyer exclusives, but we also scour regional legal publications for superb articles that you probably missed the first time around.