

# **Proposed Guidelines for Measuring the Benefit of Technology for Managing Redundant Data in E-Discovery Review**

---

*Position Paper*



**equivio**  
zoom in. find out

## INTRODUCTION

With document volumes continuing to balloon, litigation costs have increased significantly. This places a significant, sometimes inhibiting, and in many cases unreasonable, burden on litigating parties. In the context of the EDRM model, document review represents a key element within the overall cost of litigation. Various industry studies have shown that document review can cost about \$3 per document. With many reviews exceeding a million documents, cost of the document review can be considerable, even prohibitive. Even more problematic is the difficulty of ensuring proportionality between review costs and the intrinsic stake of the litigation.

## DEFINITION OF REDUNDANT DATA

The reduction of litigation costs, or even just restraining the upward costs spiral, has become a key challenge for practitioners, consultants and technologists in this industry. The industry has been witness to a wide variety of approaches, including off-shore outsourced review, clustering, and a variety of search techniques, including fuzzy search and conceptual search. Among these approaches has been the attempt, using various technologies, to isolate redundant data.

An important manifestation of redundant data is data that appears in more than one document. An obvious, if trivial, example of redundant data is duplicate documents. Redundant data is also apparent in near-duplicate documents and email threads. Near-duplicates and email threads are non-trivial instances of redundant data, and they will be the focus of this paper.

Near-duplicates are documents that differ from other documents by a few words. Even a difference of one character will ensure that two documents will not qualify as exact duplicates, and will be considered near-duplicates. In longer documents, the differences may extend to a few paragraphs. Clearly, the differences in a near-duplicate document represent unique content, but the remainder of the content is redundant. Near-duplicates harbor significant volumes of redundant data. Tests conducted on the TREC 2008 data set, comprising 6,910,192 documents, and using a threshold resemblance of 60%, show near-duplicate levels of 31.7%. The near-duplicate level is in addition to exact duplicates.

Email threads also include significant volumes of redundant data. When responding to an email, most users include the original email in the replay or forward email. The last email in the thread typically contains all the previous layers in the chain. Viewing this from the perspective of litigation review, it can be sufficient to read just the last email in the thread. Stating this generically, the identification of email containment renders the contained email redundant. Email containment analysis of the emails in the Enron data set shows that 60.8% of the emails are contained in other emails. Software that can

identify these superset emails has the potential to reduce the review effort by up to 60% because the contained emails are redundant.

## THE CHALLENGE

Given the potential value in terms of cost savings for document review, it is hardly surprising that a number of vendors and technologies have emerged to deal with the problem of redundant data. These technologies have generated a lot of interest in the electronic discovery arena, and are widely used. While vendors vigorously tout the value of using near-duplicate and email thread technologies, a key challenge is to quantify the benefit.

Quantification and measurability are requirements for most new technologies, but they represent a particular challenge for redundant data. From the user's point of view, redundant data represents a classic chicken and the egg scenario. In the absence of the solution, it is difficult, if not impossible, to quantify the problem. While users typically profess to being aware of redundant data in their litigation collections, they are universally unable to measure the extent of the problem. As a hidden cost, many users are reluctant to invest in a remedy. The resultant cost to litigants is profound.

In the following sections, we will discuss possible techniques for measuring the benefits of near-duplicate and email thread technologies.

## QUANTIFYING THE BENEFITS OF NEAR-DUPLICATES TECHNOLOGY

A key factor in quantifying the benefits of near-duplicates technology is the proportion of near-duplicates. This may seem trivial, but the following example reveals a potential complexity. Let us assume we have 5 documents in our collection – A, B, C, D and E. Using a near-duplicates grouping technology, we discovered that C, D and E are near-duplicates. The question is – what is the percentage of near-duplicates in this collection? Clearly, the 3 of the 5 documents are near-duplicates, so the percentage of near-duplicates is 60%. In order to isolate the effect of near-duplicate groupings, it is important to separate near-duplicates and exact duplicates. For example, if documents A and B were exact duplicates, the near-duplicate level would remain 60%. Such calculations, to determine the level of near-duplicates, are relatively trivial. However, calculation of the review savings is more complex. In order to explain this, we need to describe the review process for near-duplicates in more depth.

The common practice for review of near-duplicates proceeds as follows:

- 1** Assign the near-duplicate sets to reviewers, ensuring that one reviewer handles a document and its near-duplicates in a systematic manner.

- 2 Start review of a near-duplicate set with one of the documents – the base document.
- 3 After reading the base document, the reviewer decides whether the rest of the documents in the near-duplicate set can be skipped or require granular review.
- 4 Using a compare facility, the user zooms in on the differences in each near-duplicate vis-à-vis the base document.
- 5 The user can bulk handle a near-duplicate set.

From step 2 in this procedure, it is clear that, at the minimum, the reviewer needs to read at least one document from the near-duplicate set. Returning to our example, the reviewer will, at best, have to read documents A, B and one of C, D or E. That is, the review savings will be, at best, 40%. As we shall see below, this needs to be even further refined.

One factor that needs to be account into account is the extent of resemblance between the near-duplicate documents. In order to illustrate the point, we can take this to an extreme. Let's assume that we set the threshold level of resemblance to 0%. All the documents in the collection would be grouped together, and our near-duplicates percentage would be 100%. However, this is obviously of no value. For most data sets, a near-duplicate level of 60-65% is optimal. If the threshold is set above 70%, you will tend to miss many documents that human beings consider to be near-duplicates. Similarly, if the threshold is set below 50%, the differences tend become too substantive. The only exception is poor quality OCR data where a threshold of as low as 40% will return useful results. The sensitivity of near-duplicate groupings to the resemblance threshold can be illustrated, once again, using the TREC data. The results are shown in Table 1.

*Table 1 Table 1: Percentage of Near-Duplicates as a Function of Resemblance Threshold\**

<b>Resemblance Threshold</b>	<b>Percentage of Near-Duplicates</b>
50%	39.6%
60%	31.7%
75%	20.8%

\* Results based on tests of TREC 2008 data

Another important factor in determining the value of near-duplicate groupings is the size of the near-duplicate sets. Let's assume we have two collections, each of 10 documents. In the first collection, we find one near-duplicate set comprising 6 documents. In the second collection, we find 3 near-duplicate sets, each of 2 documents. From the reviewer's point of view, it is much quicker and easier to review 1 set of 6 documents, rather than 3 sets of 2 documents. In other words, to summarize the discussion to date,

the review savings generated by near-duplicate groupings will maximize the higher are near-duplicate percentages and the lower the number of near-duplicate groupings.

The final factor to take account is the extent of internal review within the near-duplicate sets. As described above, the standard review procedure for near-duplicate sets is to read one of the documents. In many near-duplicate sets, it's possible to make a valid responsiveness call based on that initial document ("base review only"). In other sets, the reviewer needs to make a granular review of each of the other documents in the set. As noted, the review of the other documents in the set is conducted using a compare capability that highlights the differences ("internal review"). As such, the extent of internal review within near-duplicate sets is a function of two parameters: the percentage of near-duplicate sets in which internal review is required, and, where internal review is required, the extent of differences between documents. In respect to this second factor, if a very low resemblance threshold was used, we can expect that the differences between near-duplicates will be greater in scope. As such, the review effort is a function of the similarity threshold used.

We can summarize the review savings generated by near-duplicate detection technology as follows:

$$RR(\text{near-duplicates}) = C * \{(B * [S - 1]) + (I * [S - 1] * R)\}$$

Where:

- RR: Reduction in review costs due to near-duplicate groupings
- C: Review cost per document
- B: Number of base review only sets
- S: Average size of near-duplicate sets
- I: Number of internal review sets
- R: Mean percentage resemblance between near-duplicates and the base document in the internal review near-duplicate sets

## QUANTIFYING THE BENEFITS OF EMAIL THREADS TECHNOLOGY

As noted above, the identification of containment emails harbors the potential for significant review savings. Intuitively, we know that the last email in a thread typically contains all the previous layers of that chain of emails. However, in some cases, the thread is truncated, for example, by users that do not want to expose specific content to subsequent addressees, or earlier layers are subsequently modified. As such, identification of containment emails is dependent on analysis of email content. Such a

capability becomes more pertinent as the mix between documents and emails continues to shift towards email.

At the basic level, the savings generated by email containment analysis is a function of the number of contained emails. Standard review practice is to review only the containment emails. To quantify this saving, let us assume that our collection contains 10 emails. The email analysis technology identifies that 7 of these emails are contained in the other 3 emails. The review saving will be 70%. However, this needs to be refined to take account of attachments.

Let us assume that we have three emails:

<p>To: Bob          From: Jim          Date: Tues 1:03          Subject: Meeting          -----          Bob and Howard,          We'll meet at 7 pm to          discuss attached          document, OK?</p>	<p>To: Jim          From: Bob          Date: Tues 5:30          Subject: Re: Meeting          Jim,          I won't be able to make it.          Sorry, Bob            ----included message          follows----          To: Bob          From: Jim          Date: Tues 1:03          Subject: Meeting          -----          Bob and Howard,          We'll meet at 7 pm to          discuss attached          document, OK?</p>	<p>To: Bob          From: Jim          Date: Tues 5:40          Subject: Re: Meeting            BOB, OK, FORGET IT, JIM          ----included message          follows-----          To: Jim          From: Bob          Date: Tues 5:30          Subject: Re: Meeting            Jim,          I won't be able to make it.          Sorry, Bob            ----included message          follows----          To: Bob          From: Jim          Date: Tues 1:03          Subject: Meeting          -----          Bob and Howard,          We'll meet at 7 pm to          discuss attached document,          OK?</p>
--	---	---

Attachment: East Contract May2.doc	-	-
Email-1	Email-1.1	Email-1.1.1

In terms of content of the email text, we can see Email-1.1.1 contains both 1.1 and 1. On the face of it, the reviewer could read just 1.1.1 and will have covered all the data. However, email 1 includes an attachment. Attachments are not usually carried forward (as we see in this case, emails 1.1 and 1.1.1 do not include the attachment. From the reviewer's point of view, it is important to review both email-1.1.1, as well as email-1, in order to review the attachment.

Based on this analysis, it can be noted that the review cost savings generated by email containment analysis are a function of the number of contained emails, less any emails that include attachments. This can be expressed mathematically as follows:

$$RE(\text{email threads}) = C * (N - NA)$$

Where:

RE: Reduction in review costs due to email thread groupings

C: Review cost per email

N: Number of contained emails

NA: Number of contained emails that include an attachment

This proposed formula assumes that the reviewer will read each of the containment emails. However, some threads contain multiple sub-threads which share a common route, and, often, a substantially common thread. In such situations, the efficiency of the review can be further enhanced by invoking a compare function to highlight the differences between the containment emails within a given thread. As such, the proposed formula may be a conservative estimate of the email thread cost savings.

## LIMITATIONS AND QUALIFICATIONS

The foregoing paper has proposed possible techniques for quantifying the benefits of redundant identification technology, specifically for near-duplicates and email threads. However, in conclusion of this paper, it is posited that the analysis of quantifiable benefits to the exclusion of non-quantifiable benefits, is liable to understate the potential value proposition, and distort use methodologies.

Accordingly, in addition to the quantifiable benefits discussed above, the identification of redundant data has strategic implications that would need to be taken account of in

analyzing the potential benefits of the technology. One of the key non-quantifiable benefits of such technologies is the reduced risk of inadvertently missing key data. The corollary of identifying redundant data is the identification of unique data – in other words, unique documents or emails, and the unique data within those documents. The suppression of redundant data brings the unique data to the surface, and the concomitant risk of the reviewer missing key non-redundant data is reduced accordingly. Similarly, near-duplicate and email thread groupings, which are the pre-requisite to identifying redundant data, also enable bulk handling, and so contribute to the reduced risk of inconsistent coding of documents. When used post-review, the near-duplicate and email thread groupings can be utilized to fulfill a quality assurance function, identifying sets of similar documents in which responsive or privileged codes are not consistent. While the development of accepted techniques for quantifying the benefits of redundant identification technologies is undoubtedly an important step towards enabling widespread adoption of near-duplicate and email threads technology, the analysis of the technology's benefits must also take into account other parameters that do not readily translate into quantifiable variables.

## ABOUT EQUIVIO

Equivio develops text analysis software for e-discovery. Users include the DoJ, the FTC, KPMG, Deloitte, plus hundreds of law firms and corporations. Equivio offers Zoom, an integrated web platform for analytics and predictive coding. Zoom organizes collections of documents in meaningful ways. So you can zoom right in and find out what's interesting, notable and unique. Request a demo at [info@equivio.com](mailto:info@equivio.com) or visit us at [www.equivio.com](http://www.equivio.com).

**Zoom in. Find out.**