

eDiscovery Document Review: Understanding the Key Differences Between Conceptual Searching and Near Duplicate Grouping

White Paper by Kimberlee L. Gunning, Esq



equivio
zoom in. find out

INTRODUCTION

As anyone who has spent time in the document review trenches will attest, duplicate and near-duplicate documents comprise a significant percentage of the electronic information requiring review. Several dozen copies of the same agreement, in different fonts because someone thought Times New Roman was classier than Arial. Seemingly endless email chains, which started with an innocent query from a regional manager to district managers until those district managers forwarded it to their assistant managers, who copied their assistants, who hit “reply all” more often than not. We’ve all been there, and it’s not pretty.

De-duplication technology, which culls out exact duplicates from electronic information to be reviewed, is well known. Near-duplicate grouping should be. Typically, between 30–50 percent of electronic files in a case are near-duplicates. As the term suggests, near-duplicates are not exact copies of an electronic file but files with small difference. Examples include electronic documents with a few different words (different versions of an agreement or letter), electronic files with the same content but different formatting, and files with the same content but a different file type (a document in both Word and PDF format).

Many litigation teams start out with traditional keyword searching and then advance to conceptual word searching. A step beyond traditional keyword searching, conceptual searching retrieves all documents that relate to a given subject. No one will dispute that conceptual searching helps provide perspective and context to a document review by permitting the litigation team to review related documents together.

For example, if you are interested in documents regarding Defendant Corporation’s “Project XYZ,” which seems very similar to your client’s “Project ABC,” and which forms the basis for your client’s trade secrets claim, it is helpful to review all documents regarding Project XYZ in context with one another. Reviewing notes from members of the Project XYZ development team and drafts of Project XYZ marketing materials can help bolster your client’s claim that Defendant Corporation has misappropriated your client’s Project ABC technology, changed the name, and marketed the project as its own.

Although it goes a step beyond traditional keyword searching, conceptual searching has its own set of limitations. First and foremost, it does not solve the problem associated with near-duplicate documents. This white paper discusses the key differences between conceptual searching and near-duplicate grouping, and explains how using near-duplicate grouping, either alone or in conjunction with conceptual searches, can help you cope with a mountain of electronic files in an efficient and systematic manner.

NEAR-DUPLICATE GROUPING RESULTS IN MORE UNIFORM RESULTS THAN CONCEPTUAL SEARCHING

Starting with a uniform near-duplicate group of documents, you can review just one representative document from each set of near-duplicates, without having to analyze each individual document retrieved as a result of a conceptual search. For example, if you are interested in the type of form letters Defendant Corporation is sending to its sales prospects, you can start with a set of near-duplicate form letters containing Defendant Corporation's sales pitch — letters that may differ only by date and name and address of recipient — and quickly determine the relevance of the email's content for your case.

Used in conjunction with a conceptual search tool, near-duplicate grouping provides additional structure to your search results. A reviewer can either skip the near-duplicates, as indicated above, or, depending on the nature of the case, analyze the differences between the near-duplicate documents by using a document comparison utility.

NEAR-DUPLICATE GROUPINGS ARE MUTUALLY EXCLUSIVE, RESULTING IN GREATER EFFICIENCY

A single document can belong to only one near-duplicate group. This mutually exclusive grouping provides for more consistency and efficiency when reviewing documents. Reviewing similar documents at the same time, and reviewing them only once, means your litigation team is more likely to categorize similar documents in the same way. With near-duplicate grouping, a single reviewer is more likely to analyze related drafts of a document and code these documents in the same way. Document review proceeds more quickly, and is less prone to error, when a document turns up only once and is not subject to being coded differently by a different reviewer, or by the same reviewer when their energy level begins to wane in the late afternoon.

The mutually-exclusive nature of near-duplicate grouping enables you to quickly weed out documents with little or no relevance to the case, such as announcements of a company holiday party or multiple employee emails discussing last weekend's baseball game. Such documents are the eDiscovery equivalent of dandelions – ubiquitous and hard to kill. Using near-duplicate grouping ensures you only have to pull these weeds once, thus decreasing costs.

Finally, near-duplicate grouping helps you avoid the unfortunate scenario in which your team characterizes similar documents differently and thus withholds some near-duplicates from production while producing others. Characterizing similar documents differently for purposes of document production may cause the opposing party to challenge the adequacy of your efforts to review and produce all relevant non-privileged

documents, which could lead to a “do-over” of your document production, adding to the overall cost.

NEAR-DUPLICATE GROUPINGS ARE PRE-PROCESSED, NOT LIMITED BY SUBJECTIVE REVIEWER DIFFERENCES

Conceptual searches, which retrieve all documents related to a given subject, are improvised and ad hoc. As such, the results and usefulness of a conceptual search, like all term-based searching, are subject to varying levels of knowledge, judgment, and imagination among the different reviewers, not to mention plain old human error.

By contrast, near-duplicate groupings are pre-processed persistent categories. Moreover, they are objective groupings built by software independent of subjective judgments. The ability to identify the textual differences between near-duplicates provides reviewers with a high level of confidence and assurance that the near-duplicate groupings “make sense” and can be trusted. For this reason, the near-duplicate groupings are a useful vehicle for organizing the review flow. The objectivity of the near-duplicate groupings ensures that the use of near-duplicates as an organizing parameter for the review is intuitive and acceptable to users.

NEAR-DUPLICATE GROUPING SLAYS THE EMAIL CHAIN DEMON

Email is by far the most common type of electronic files in most cases. Near-duplicate grouping has unique advantages when applied to review of email. Anyone who has reviewed seemingly endless email chains or form mail will appreciate the ability to identify email files that are near-duplicates.

Typically, it is useful to review the last email in a chain as it contains the previous email and responses. On email copied to several individuals, many of whom may not hit “reply all,” near-duplicate grouping, combined with email thread technology, helps locate all email in the chain regardless of whether recipients of the initial email were copied on later replies and forwards. In short, near-duplicate grouping makes it easier to focus on documents that contain the complete email chain and not just the initial email or reply.

Moreover, when reviewing near-duplicate groupings of email, you’re not limited by custodian (sender or recipient). Instead, you can review the email exchange as it took place and analyze to whom certain messages and attachments were forwarded and when.

CONCLUSION

Whether proving your claims and defenses requires close comparison of nearly-identical documents or a quick and efficient method of isolating a few key documents, or both,

you should consider adding near-duplicate grouping to your electronic discovery toolbox. Electronic discovery usually begins with an unstructured mass of documents. Near-duplicate grouping provides a way out of this morass.

ABOUT KIMBERLEE L. GUNNING

Kimberlee L. Gunning is an attorney in Seattle, Washington (www.gunninglegal.com). Her practice focuses on employment advice and litigation, consumer class actions, and civil and administrative appeals. She often acts as co-counsel or contract attorney in complex litigation matters and has a special interest in the emerging law of eDiscovery.

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 or visit us at www.equivio.com.

Zoom in. Find out.