



White Paper

Using Near-Duplicates:
Applying Near-Duplicate Technology
in Litigation Matters

Introduction

Near-duplicate identification technology, whether deployed on a stand-alone basis to a discovery document collection or in conjunction with other analytical models such as context or concept analysis, has the potential to greatly increase the efficiency of the litigation document review process and significantly reduce discovery costs. However, though it's easy to understand the basic value of grouping together substantially similar documents for common analysis, many practitioners haven't had enough experience with this technology to visualize specific ways they could apply this tool to their cases and projects. This paper offers a number of concrete ways in which near-duplicate identification has been successfully deployed to help a legal team efficiently achieve its discovery objectives.

The discovery document life cycle includes many projects and tasks, depending on the nature of the dispute and the discovery materials at issue. However, at least three activities—initial review of a collection, production (or exchange) of relevant documents, and post-exchange review of documents received from an opponent—are part of virtually every legal matter involving the collection and exchange of factual evidence. Near-duplicate identification can assist a legal team with each of these broad activities

I. Pre-Production Review of Discovery Documents

Preparing discovery materials for production to a requesting party can be a time and budget-consuming process. Legal teams are pressured by the competing objectives of doing the best job possible while also staying within budget. Near-duplicate identification can help, regardless of whether the project is a basic relevance review or a more complex and nuanced subjective coding effort.

A. Increasing Document Review Rates

One challenging development in modern litigation is that many discovery document collections are so voluminous that it is no longer physically possible to review every document for relevance or privilege. Instead, the scope of the document review is defined by the budget allotted for this task, with hopes that the critical documents within the collection will be found through efficient search and review techniques. Near-duplicate identification increases the reliability of time and budget-limited document review by increasing the productivity of individual reviewers.

Document populations inevitably contain many exact duplicate documents as well as

near-duplicate documents, such as multiple versions of a contract containing proposed changes from several reviewers. If a reviewer encounters each of these closely related documents as they occur in the document collection, each document must be individually reviewed and weighed. The reviewer may recall seeing a similar document earlier in the review, but such fuzzy recollections are rarely sufficient basis to make quick, wholly accurate, document classification calls.

In contrast, grouping closely related (i.e., duplicate and near-duplicate) documents from a custodian's materials or, ideally, across the entire document collection, consolidates the number of unique document review decisions that will need to be made. In primarily electronically stored document populations, the rate of near-duplication may exceed 30% of the collection, even after exact duplicate documents are identified and removed via MD5 hash comparison or other methodology. The fewer unique documents that need to be reviewed, the faster the review will be.

Within near-duplicate document groupings, judicious review of a sample document often provides sufficient information for the reviewer to determine whether other documents within a grouping will require individual review. Based on the sample, the reviewer can apply the results of a document categorization across all the near-duplicates in the set. Used this way, near-duplicate groupings can be used to quickly exclude large numbers of similar documents from individual document review and analysis. This set-centric review paradigm, as opposed the traditional linear, document-centric workflow, greatly increases the number of documents that the reviewer can effectively process. The degree of productivity increase depends on the size and number of near-duplicate document groupings within the review population, but even a modest number of small groupings can have a significant impact on the progress of the overall review.

Some groups of near-duplicate documents require more granular categorization decisions, based on the individual document, rather than a bulk decision. However, even in situations where every document in the grouping may require individual review, the total time to review all documents in the grouping will be far less than if these documents were encountered singly and each individually analyzed out of the context of predecessor and subsequent versions.

Moreover, by utilizing the "compare" functionality offered with some near-duplicate solutions, individual document review can be further streamlined, eliminating the need for reading each document, cover-to-cover. The compare function takes the sample as the baseline document, then highlights dissimilar language within each of the near-duplicate documents. Review can then be focused on the unique data in each document, rather than their substantial, and redundant, similarity

B. Increasing the Depth of Review without Increasing Review Budget

Using near-duplicate document identification and grouping to maximize the number of unique documents that are analyzed also meets a different but equally important objective: understanding a document population even when it is not possible to review the entire collection. Here, near-duplicate document groupings serve as an additional filter to prevent reviewers from repetitively analyzing closely related documents instead of new materials, permitting reviewers to work through a greater number of unique documents in the time available for the project. Used for this purpose, near-duplicate identification provides the legal team with greater confidence in the statistics they derive from a limited document review project. In addition, even the earliest assessment reports will be based on a greater diversity of unique documents, leading to greater accuracy.

C. Achieving Consistency in Review

No document review takes place today without some preliminary organization of the document collection. Most often, document collections are organized by source or custodian, by creation date (especially when reviewing e-mail messages and other electronic documents with embedded creation date metadata), or by search queries. Each of these organizational systems adds context to documents so that their relevance can be better assessed. However, each of these approaches is also likely to separate many closely-related documents, which can lead to inconsistent document triage and categorization caused by multiple reviewers and dissimilar contexts in which the documents are analyzed.

Near-duplicate document identification and grouping helps a document review team achieve greater consistency in its overall analysis by bringing together closely related documents that would not be connected by traditional document review paradigms because of distance from one another in creation date, custodian, or source. Viewing these documents together greatly increases the likelihood that similar documents will be similarly categorized. For example, with near-duplicate document grouping, closely related drafts of a document would be analyzed by a single reviewer, regardless of the time that might have passed between versions or whether every version contained key search terms. Indeed, if the reviewer determines that there is sufficient commonality between all documents in the grouping, the reviewer could apply common categorization across the entire grouping, guaranteeing that these documents will be similarly treated, regardless of where they appear in the document collection.

Legal teams should not be concerned that deploying near-duplicate document

identification inevitably disrupts the way in which the legal team has already structured its review; near-duplicate identification technology easily works in conjunction with a legal team's existing search and document batching processes, not as a replacement. For example, sorting documents by search query or by custodian can still provide a very useful point of entry for document reviewers by giving them an opportunity to find initial context for a document. Near-duplicate identification simply offers the opportunity to consistently categorize other iterations of the document when relevant (or irrelevant) documents are found within an assigned review batch.

Similarly, within a chronological sort-based review, near-duplicate identification offers significant enhancement without disturbing the philosophy of the review paradigm. As the reviewer works through the chronology, near-duplicates of each document can be retrieved and reviewed. Reviewing the entire grouping at that point permits reviewers to not only analyze the specific document within their review batch, but also consistently categorize the other iterations of that document that appear elsewhere in the collection.

When these documents are subsequently encountered, they will already have been reviewed, saving time. In addition, a reviewer is likely to gain additional context about the document collection from reviewing an entire near-document grouping and the changes within it.

Near-duplicate document identification can also serve a more active role as the primary review methodology for legal teams that have not yet finalized their workflow for a review project. Instead of organizing review assignments based on custodians or chronological slices, the legal team can use near-duplicate groupings as a basis for review.

In assigning the near-duplicate groupings for review, assignment of the largest groupings first serves two key objectives. First, categorizing the largest groupings will have the largest immediate impact on the document collection. Second, large groupings often contain high-value documents (i.e., documents that went through many revisions or were circulated to many individuals for comment) or low-value documents (e.g., press releases, holiday party announcements).

II. Cost-Efficient Quality Control Prior to Discovery Document Production

Any human-based document review will contain some inconsistent judgment calls. Understanding of a case naturally evolves over time, and even the same reviewer may classify similar documents quite differently, depending on whether they were analyzed at the start or at the end of a document review project. Such inconsistencies weaken the

integrity of a review and may permit the production of materials that should have been withheld. Conversely, a requesting party could use inconsistent document categorization to challenge the adequacy of the review effort, arguing in a worst-case scenario that inadvertently produced privileged documents were negligently produced and should form the basis for subject matter waiver of related privileged communications.

Near-duplicate groupings enhance any existing Quality Control (“QC”) measures used by the legal team to check work done by the reviewers. A simple database query can check whether all documents within a near-duplicate grouping have been classified similarly for analytical purposes such as relevance, attorney-client privilege, or any other subjective categories that were made part of the review. Inconsistent documents can quickly be isolated, examined, and their categorization validated or corrected.

This QC process can be further optimized by working from largest to smallest groupings, permitting QC of the greatest number of documents in the shortest amount of time. As an alternative, a random number of near-duplicate groupings can be identified via statistical model, with each selected grouping tested for inconsistent document categorization.

Using near-duplicate groupings as part of a quality control or quality assurance process is a highly defensible, cutting-edge process. Especially for law firms have only limited staff available for QC functions, incorporating this functionality into the discovery document management process should provide significant confidence, both in the quality of the review effort and in explaining to a court about the adequacy of the measures taken to ensure production quality and to prevent inadvertent production of non-responsive or privileged materials.

III. Mastery of Documents Received in Discovery

Commentators and practitioners often focus on a producing party’s need to understand its discovery documents before they are produced, but it is often equally important to understand the discovery materials that have been received from legal opponents and co-plaintiffs or co-defendants. How many unique documents are in the production? Does the client already have copies of any of these materials? How were specific relevant documents distributed throughout the organization, and with what consequences? What documents should be identified as potential trial exhibits? Near-duplicate identification can help with each of these tasks.

As noted earlier, virtually all document productions—produced or received—contain substantial numbers of duplicate or near-duplicate documents. The same strategies used to quickly prepare materials for production to requesting counsel are equally helpful here,

when legal teams may be under substantial time pressure to digest materials received from opposing counsel and to prepare for depositions or court filings. Near-duplicate groupings can expose the number of unique documents that require analysis, and judicious sampling of representative documents within a grouping will often provide sufficient information for a reviewer to determine whether any other document within the set will require additional review.

Near-duplicate identification can also be used for the more subtle task of identifying whether substantially similar documents are owned or retained by both sides. For example, in legal disputes such as business-to-business litigation or contract negotiation disputes, both sides may have many documents in common, and it makes little sense to review duplicative materials simply because they were produced by an opponent. Near-duplicate identification technology can be applied to competing document collections to determine the degree of overlap between the collections.

Near-duplicate identification can also be applied to determine the degree of overlap between materials received (or harvested) as native files and hardcopy documents from which OCR text has been generated. This is particularly helpful, as hardcopy documents are not as well-suited for automated analysis and are often less-rigorously reviewed by legal teams operating with a tight budget or time constraints. Linking electronic and hardcopy documents provides another opportunity to better understand how key documents were distributed within an organization, the number of unique documents actually provided by a responding party, and many other questions that may be of interest to the legal team. Of course, OCR is only as good as the quality of the source documents and the processing engine that generated the text, but best-of-breed near-duplicate technology permits lowering of the threshold for creating a near-duplicate relationship so that imperfections within OCR have less of a negative impact on linking related documents.

Conclusion

Near-duplicate identification is a valuable tool that assists human reviewers in their analysis of document collections. This technology does not replace the human ability to identify subjective relationships between disparate documents. However, by reducing the effort required to find unique relationships within collections, it frees resources within the legal team to develop the case in other substantive ways. Properly deployed, near-duplicate identification helps a document review proceed faster and at lower cost to clients and improves consistency within human-based document review. It is a valuable and highly cost-efficient way for legal teams to achieve their objectives.