



White Paper

Choosing A Near-Duplicate
Identification Solution

Introduction

Legal teams are increasingly aware of ways that document review projects can be made more efficient through better identification of “near-duplicate” documents. These documents, which share significant content with one another but which still require individual analysis, appear particularly frequently in collections of electronically stored information (“ESI”) that include substantial numbers of e-mail messages or that include snapshots of data taken over time, such as backup tapes.

Near-duplicate documents may contain important differences, such as the addition of the word “not” to key sentences, but because of their close degree of similarity, even documents that differ in substance still require joint analysis for purposes of relevance, privilege, and many other subjective coding projects. As a result, using near-duplicate document groupings in a document review increases review efficiency and ensures consistent treatment of closely related documents—all without forcing attorneys to ignore document differences.

A number of solution providers in the litigation support industry have developed technology for identifying and grouping near-duplicate documents. These systems differ from each other in underlying logic and in implementation, but it’s not easy to understand how these differences make a difference when applying these solutions to a given document population. Rather than debate the merits of one algorithm over another, though, legal teams should focus on the functionality required of a near-duplicate identification system, as not all solutions provide some of the baseline capabilities.

Can the near-duplicate system identify documents with a level of similarity below 90%?

Some of the near-duplicate solutions emerging in the market use re-purposed conceptual search or classification technologies. Such technologies, when used for grouping near-duplicates, allow only a very high level of similarity as the near-duplicate threshold. Experience shows that near-duplicate groupings expand dramatically when using levels of similarity of 75% and lower. While it’s certainly helpful to identify documents with an extremely high degree of internal similarity, review teams can almost always get even more benefit from seeing a broader range of closely related documents. Early drafts of a contract may look quite different from the final executed document, but having these documents next to one another in review permits much faster review of the second document and invites comparison of differences between the two, rather than start-to-finish analysis of both documents. If these documents remain ungrouped (sometimes referred to as “false negatives”), they remain available for search and review, of course, but their separation from their near-duplicate cohort reduces the benefits of using near-

duplicate identification in the first place.

When the system misses large volumes of near-duplicates, due to a near-duplicate threshold that is fixed too high, the number of discrete documents that need to be reviewed increases; a document must be freshly evaluated every time it occurs outside a near-duplicate grouping, consuming additional time and energy. In addition, as happens in virtually every document review, the failure to capture near-duplicates within a collection increases the likelihood that similar documents will be classified differently.

Does the system allow the user to set the “near-duplicate threshold” based on case-specific criteria?

Ideally, a near-duplicate identification system can adjust its output to provide greatest value to the document review team. Some systems, however, use a fixed approach in defining the relationship between documents in a collection. Such “one size fits all” approaches may not meet the needs of all projects, because documents that are valid near-duplicates within the context of a document collection may fall outside the generic definition used by the system.

A better near-duplicate identification system will permit users to select the degree of overlap between documents that is required. In particular, the near-duplicate threshold must be adjustable for different data profiles. Substantially similar documents will have a different level of commonality depending on whether the source of the document is paper or electronic. OCR text in particular requires lower similarity thresholds due to OCR errors. Stating this more directly, two near-duplicate documents that differ only in a few words, will have much higher commonality if they are compared in their native format, rather than OCR derived from scanning a paper printout of the same documents. For OCR documents, high levels of resemblance are not useful because they fail to identify most of the near-duplicates. This also applies in scenarios when the user is matching documents across different collections, such as collections exchanged between plaintiff and defendant, between multiple defendants, or OCR and electronic collections generated at different points in time. In all these cases, to the extent that OCR documents are involved, experience shows that lower near-duplicate levels yield the most effective matching results. These examples illustrate the need for flexibility in the near-duplicate identification solution so that the user can determine the near-duplicate threshold, and adjust it to the specific data set and business objective at hand.

Does the system eliminate false positives?

When they occur, false positives—unrelated documents that have been included in a near-duplicate grouping—greatly limit the power of near-duplicate-based document analysis.

After all, grouping near-duplicate documents should permit a reviewer to categorize all the documents in that group based on a review of only a few documents. The possibility of false positives, however, forces reviewers to carefully review all documents within the grouping to ensure that no documents are incorrectly categorized. While such focused review is still much faster than reviewing a mixed and unorganized collection of documents, it still does not provide the legal team with the full benefits of a valid near-duplicate analysis.

Potential users of near-duplicate technology should ask how each solution provider minimizes the risk of false positives. Ideally, the solution should verify that the documents within each set of near-duplicates have a level of similarity exceeding the user-defined similarity threshold.

Can near-duplicate identification be applied on an incremental basis?

As a rule, with occasional exceptions, document collections grow over time, with small (or large) numbers of documents added to an initial set of materials. Some near-duplicate solutions require that the entire collection be processed (or re-processed) as a whole in order to identify near-duplicates on a collection-wide basis. Others permit additions to be processed separately, with the results merged into the existing near-duplicate analysis. If available, this second approach offers significant benefit to the legal team, both in terms of speed (smaller amounts of data process faster than large collections) and in cost (smaller amounts of data cost less to analyze than repeated processing of large collections).

Does the near-duplicate system support the presentation of differences between grouped documents?

Near-duplicate identification is designed to increase the speed and efficiency of document review. In an ideal implementation, efficiency is maximized when a document must be read in full only once, and only differences between it and its near-duplicates reviewed thereafter. While all review platforms using near-duplicate identification can easily inform a reviewer of whether a document is an exact duplicate of another in the context of the review, some platforms offer further functionality by highlighting changes between the first document in the near-duplicate grouping and others within the cohort so that reviewers have strong visual cues as to the document language that will require close scrutiny. Near-duplicate analysis is often performed separately from the actual document review tool used by the legal team, so it is possible that different review platforms may display the same near-duplicate analytical results differently. Seeing a real-time implementation of review solution will quickly reveal the depth of the integration between review platform and near-duplicate identification technology.

Are the near-duplicate groupings mutually exclusive?

In the most efficient document review paradigm, each document is reviewed only once for baseline analysis, with the potential of second-pass review for additional criteria not considered in the first pass review. Near-duplicate identification adds further efficiency by permitting some review decisions to be applied to multiple documents after review of only sample documents within a near-duplicate grouping. A number of near-duplicate technologies, however, permit a document to be included in more than one grouping. While this is intended to support different points of entry to the document collection, this approach also invites conflicting classification to be applied to a document, as different near-duplicate groupings may be treated differently. To the extent that 100% review is conducted of all materials within a near-duplicate grouping (a potentially important strategy if the degree of duplication is set fairly low), permitting a document to exist in multiple near-duplicate groupings also means that it will be reviewed multiple times, wasting time and money and reducing a major benefit of near-duplicate identification.

Can near-duplicate analysis be applied on a collection-wide and subset basis?

Duplicate and near-duplicate documents appear throughout a discovery document collection. While it is helpful to see the near-duplicate groupings that occur within a review subset or search results, near-duplicate identification provides greatest benefit when it is able to find the largest groupings possible, which would typically be across the entire document collection. Universal near-duping, across the entire data population, permits reviewers to find and classify closely related documents, even if they lack the specific search terms found in some but not all of the near-duplicate documents.

Collection-wide near-duplicate identification also helps the review team better understand the context of a document. Often, the author of a document might have several drafts stored on a computer hard drive, with additional copies attached to email messages or stored on external hard drives or thumb drives. It can be very helpful to group near-duplicates within a custodian's materials (groupings within a subset) to see how a document changed through its creation process. Separately, it may also be valuable to understand the distribution of a document through an organization by identifying all places throughout the collection where a near-duplicate copy can be found (groupings within the entire collection). Ideally, a near-duplicate identification solution permits the legal team to use both analyses as appropriate.

Does near-duplicate analysis slow processing and delay the start of the review?

No technology should be deployed without understanding its impact on work flow and the professionals using the technology. In looking at near-duplicate identification systems, it's important to test whether near-duplicate identification slows the speed with which document reviewers can first access the collection. Clearly, throughput of the near-duplicate identification technology directly impacts the time it takes to begin substantive document review. Experience shows that a near-duplicate system that can process a million documents in 1-2 days does not represent a bottleneck in the processing and review work flow. In order to optimize operational flexibility, while minimizing implementation costs, market expectations are that the system can be deployed on standard PCs.

Another consideration is scalability. Solutions that are optimized for single-processor operations may require an extended amount of time to identify relationships in even relatively modest document collections, delaying the start of review. Near-duplicate identification solutions that can be distributed across multiple CPUs or processing stations, in contrast, usually process documents significantly faster, minimizing time needed to get the collection ready for review. Given the growing document volumes in discovery, the ability of the target system to support small, large and mega cases from the operational point of view, is a key criteria in selecting a near-duplicate identification system.

As noted earlier, some systems require that the entire collection be processed in one continuous operation, which means that processing cannot begin until all materials have been collected—and review cannot start until after the materials have been processed. Conversely, systems that can support incremental processing (point 4, discussed previously) permit document review to start as soon as a first batch of material has been analyzed and loaded into the legal team's review platform. In this instance, too, an early start doesn't prevent near-duplicate identification from eventually extending across the entire collection as more documents are added over time.

Are the near-duplicate groupings pre-processed, or searched for on-the-fly?

Best-of-breed solutions will pre-process the near-duplicate groupings, optimizing response time in the review environment, while also allowing the near-duplicate groupings to be utilized in the organization and assignment of documents across the review teams. This contrasts with solutions that are searching for near-duplicates on-the-fly. Another problem with on-the-fly solutions is that they often require the user to manually launch a search for near-duplicates on each reviewed document. From the

user’s point of view, this can be tedious– in typical collections, we can expect to find 20-30% near-duplicates, but this means that for 70-80% of documents, the search for near-duplicates will yield no results. The result is that on-the-fly solutions can frustrate users, and rapidly run into disuse.

A closely related issue is how near-duplicate identification has been integrated into the review tool. Does the integration slow down the speed with which the platform operates? How quickly can the system find and display near-duplicates of an interesting document? A responsive system will encourage reviewers to experiment with near-duplicate analysis; a sluggish system will discourage document reviewers from invoking this functionality, reducing return on investment made in these additional processing costs.

Conclusion

Near-duplicate document identification is a powerful tool that can offer significant benefit to legal teams reviewing discovery documents. As with any tool, though, its implementation, including both underlying technology and review team workflow, will determine the overall benefit that the team receives. The following checklist summarizes requirements for a near-duplicate identification capability:

#	Criteria	Best-of-Breed
1	Near-duplicate thresholds not limited to 90% and above	√
2	Allows user to set near-duplicate threshold	√
3	No false positives	√
4	Supports incremental processing	√
5	Supports red-line comparison	√
6	Near-duplicate sets are mutually exclusive	√
7	Near-duplicates are calculated across the entire collection	√
8	Throughput and scalability	√
9	Pre-processing of near-duplicates	√

Proper due diligence as to tools and an understanding of how the near-duplicate grouping information changes the logistics of traditional document reviews will help make the use of this technology a success and a valuable return on investment for clients.