CLUSTER CLEAR

Are clustering tools the solution to tedious identification and reduction processes in e-discovery?

by George Socha, Adam Strayer, and Heena Shaikh

VOLUME 101 NUMBER 4 WINTER 2017
JUDICATURE
Published by the Duke Law Center for Judicial Studies. Reprinted with permission.

www.law.duke.edu/judicature

FRONTRUNNERS IN THE COSTLY GAME OF E-DISCOVERY HAVE BEGUN TO DISTINGUISH THEMSELVES BY USING DATA ANALYTICS in creative and effective ways to tackle the critical tasks of identifying key evidence, unearthing relevant information, and eliminating irrelevant materials. Historically these tasks have been tedious, expensive, and time-consuming. Traditional approaches to these challenges, such as linear manual review of documents one by one, haven't worked well, and shortcuts, such as developing search terms without first scrutinizing available documents, don't tend to yield accurate or useful results.

But newer e-discovery tools are much better at recognizing patterns and grouping similar documents together quickly, often with virtually no initial input from users. These tools offer hope for even a rookie data analyst. This article examines one class of these tools: concept clustering. Clustering tools quickly group documents by concept, identifying documents of greatest interest and making it easier for a human reviewer to make consistent decisions. Clustering is typically used in combination with other e-discovery tools.

Here's how it works: Consider an attorney who needs to respond to 15 document requests and is looking at 100,000 documents (which is actually a small set of documents in these days of email, Word, and Excel). For each request, the attorney can use concept clustering to identify groups of documents that contain concepts related to the request. From there, the attorney might

decide one cluster of conceptually related documents clearly is unrelated to the first request and mass-tag the documents in that cluster as nonresponsive to Request 1. The attorney might next examine the documents in a second cluster, determine they all clearly are responsive, and mass-tag them accordingly. The attorney then might look at the documents in a third cluster and conclude they warrant closer examination, perhaps by re-clustering the documents or maybe by going through them one at a time. And in looking at the top concepts as well as some of the documents in a fourth cluster, the attorney might realize those concepts reveal a fruitful line of inquiry not previously considered.

CLUSTERING TOOLS AUTOMATICALLY GROUP TOGETHER DOCUMENTS THAT ARE CONCEPTUALLY RELATED. ONE OR TWO REVIEWERS CAN THEN EVALUATE CLUSTERS OF RELATED DOCUMENTS, MAKING QUICK AND CONSISTENT DECISIONS ABOUT THE TREATMENT OF THOSE DOCUMENTS.

WHAT IS CONCEPT CLUSTERING?

More traditional e-discovery systems conduct a linear review of documents; documents are 'batched' together in ways that make sense for data processing but not for review purposes. This approach requires large numbers of reviewers to look at documents one at a time, usually with no easy way to group documents in a greater context.

In contrast, new clustering tools automatically group together documents that are conceptually related. One or two reviewers can then evaluate clusters of related documents, making quick and consistent decisions about the treatment of those documents.

Clustering also helps with organizing and exploring data. Most clustering tools identify the prevailing concepts in a group of documents based on the words and semantics – the meanings of the words – found in the documents. The tools then cluster, or organize into groups, documents containing related concepts and provide users with a few reference terms for each cluster.

For example, suppose we wanted to find documents about food – not just ones containing the word "food," but ones conceptually related to food. A keyword search for the term "food" would identify documents containing the string of characters "f-o-o-d." A keyword search that also used wildcards would return, in addition to "f-o-o-d," documents containing words like "f-o-o-d-i-e."

Searching for concepts delivers a much broader set of results. With a conceptual search, a clustering tool groups documents based on specific concepts, such as "grocery, market, produce" and "recipe, tablespoon, ingredient." Examining the contents of the clusters, or looking at related clusters, allows the user to more quickly and efficiently identify documents about types of food purchased, restaurants visited, grocery stores frequented, and so on.

To be effective, clustering tools should be able to recognize different types of documents; separate noise from meaningful content; and identify outliers, concepts, and documents that are very different from whatever has been defined as a core concept or document. Effective clustering tools also should be able to interpret and present results in ways that are meaningful to users, for example, by showing relationships between clusters visually (see Figures 1, 2, and 3 on the next page), and by allowing users to quickly drill down for greater detail. And, the order in which documents are added to the system should not affect results.

Two primary goals of clustering are data identification and data reduction. By working with conceptual clusters, users can quickly segregate documents into "relevant" or "responsive" and "nonrelevant" or "nonresponsive" groups. In an investigation about a pharmaceutical product, for example, an attorney might start by using concept clustering to examine documents. Clustering tools might create and label groups "March Madness," "football," "lunch orders," or "LinkedIn notifications." The attorney could assess a sampling of the documents in those clusters and potentially set aside the documents in all those clus-

ters as nonresponsive. The attorney could continue the exercise, identifying other clusters of potentially nonresponsive documents, sampling the contents, and putting the clusters into broad groups, responsive and nonresponsive.

INTO THE WEEDS

Many concept clustering tools are used for e-discovery; all aim to achieve similar ends but do so in significantly different ways. Most are licensed to users as part of a larger set of tools rather than as standalone clustering tools.

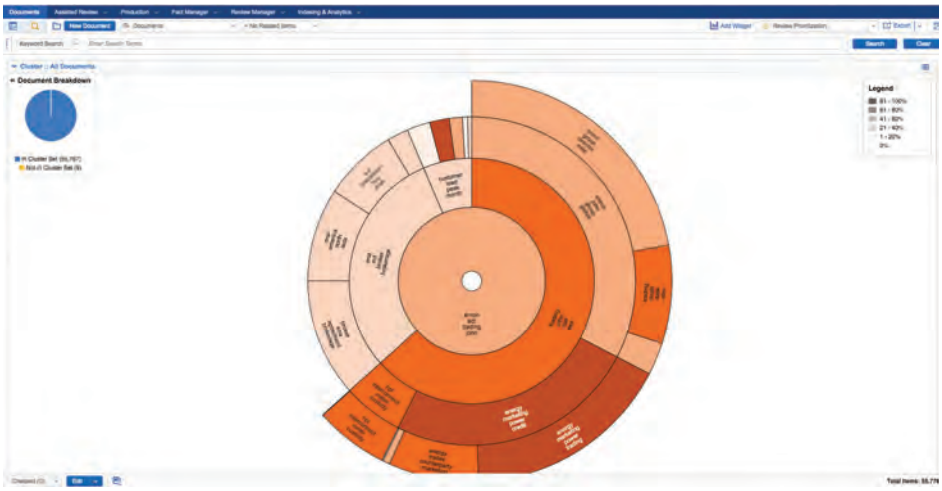
Comparing the clustering tools offered by two companies, Relativity and Brainspace, helps to explain the clustering process and its complexities. Both companies offer clustering tools as part of larger e-discovery platforms.

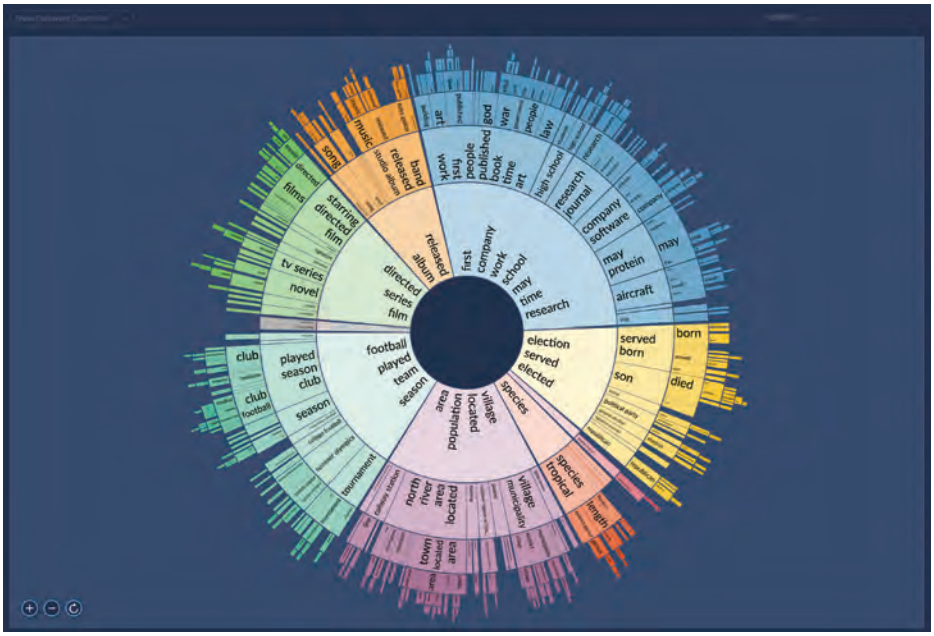
Clustering with Relativity

Relativity Analytics supports conceptual searching of unstructured data, such as the text of email messages. (See Figures 1 and 2.) Users identify which documents are to be included. As documents are added to the system, Relativity Analytics creates an index to identify similar ideas or concepts within that collection of documents. Relativity Analytics then groups similar documents into clusters and sub-clusters and ranks each based on how similar the documents within the cluster are to one another.

To accomplish this, Relativity Analytics uses Latent Semantic Indexing ("LSI"). LSI is a mathematical approach to sorting data

TOP TO BOTTOM: FIGURE 1 (RELATIVITY), FIGURE 2 (RELATIVITY), FIGURE 3 (BRAINSPACE)





based on patterns and term frequency. It is language-agnostic, meaning it can index text from any language. It generally, however, can cluster only one language at a time; so, if you have documents containing both English and Spanish, you would need two separate sets of clusters.

LSI leverages sophisticated mathematics to discover correlations among terms and concepts within a set of documents. It does not rely on outside lists of words, such as lists from dictionaries or thesauri; instead it looks at semantic relationships and the co-occurrence of certain words – how words appear together in documents.

Relativity organizes its clusters in trees, with levels. Users can set three values to control how cluster trees are created:

Users can specify the "generality" of clusters, using a range that goes from 0 for most specific to 1 for most general. The closer to 0, the broader the range of concepts in a cluster, the fewer clusters, and the more documents per cluster; the closer to 1, the narrower the range of concepts in a cluster, the more clusters, and the fewer documents per cluster.

Users can set a "minimum coherence score" to determine how closely related documents need to be in order to be in the cluster.

And users can set a "maximum hierarchy depth," which is a limit on the number of levels in a cluster tree.

Clustering with Brainspace

Brainspace takes a somewhat different tack. Brainspace uses both LSI and additional capabilities to identify multiple concepts in each document. Brainspace examines sole terms and word order, as well as combinations of words and phrases. Its output looks different. (See Figure 3.) And it yields different results than Relativity. In fact, the search results of every concept clustering tool will differ from one another, even when using the same set of documents. They cannot help but differ because each system uses different algorithms to search content, meaning that the systems search data in different ways; often systems search somewhat different content or organize the same content in different ways for searching, which again leads to differing results. Furthermore, each system is designed with somewhat different goals in mind.

USES

Clustering can be and is used in various ways in lawsuits. Here are three examples:

To remove irrelevant documents. Identifying and setting aside irrelevant documents is one of the most powerful uses of clustering. In a products case where the functioning of a third-party immersion heater is at issue, clustering can be used to find and set aside documents that, although they include references to the third-party immersion heater, discuss the heater only in ways that are totally unrelated to the issues in the lawsuit.

To find similar documents. A user can start with a key document and then have the tool generate a cluster with that document at the middle, just like the bullseye at the center of a target. The user can examine the contents of clusters nearest the center, looking for materials of interest, and move out from the center until there is nothing more of interest or time or other resources run out.

To improve the efficiency and effectiveness of review. Clustering arranges documents in groups by similarity of contents. Documents can be distributed to reviewers in ways that take advantage of those groupings, such as sending similar documents together or grouping documents by topic and sending those groups to reviewers, or sending certain groups of documents to specific reviewers.

Ultimately, concept clustering tools are best used in conjunction with other tools, such a predictive coding software and review platform. Clustering is then one part of an overall system used to identify and evaluate the content the user is looking for, find important documents the user did not even know to look for, and effectively and efficiently review the documents for relevance, privilege, and other concerns.

*– **GEORGE SOCHA** is managing director at BDO and cofounder of EDRM at Duke Law; **ADAM STRAYER** is a consulting director at BDO; and **HEENA SHAIKH** is litigation manager at BDO. BDO is a sponsor of EDRM, which is now part of the Duke Law Center for Judicial Studies. Learn more at **EDRM.net**.*