NAVIGANT

GLOBAL CONSTRUCTION

DATA ANALYTICS IN CONSTRUCTION INDUSTRY LITIGATION

Robert Neary, Associate Director,
Global Legal Technology Solutions

Amy Tsang, Associate Director,
Global Legal Technology Solutions

Experts estimate that more data was created in the last two years than throughout all of human history. By the end of 2019, 246 billion emails will be created each day.[1] With the adoption of the internet of things (IoT), telematics, or machine-to-machine communication, this data explosion is poised to accelerate at an even greater pace. Against the backdrop of this massive data growth, businesses facing litigation, regulatory inquiries, or enforcement agency investigations find it difficult — if not impossible — to manually review the millions of documents that fall within the scope of these matters, on reasonable budgets and in time to meet stringent deadlines.

In addition, industry-specific data types present unique challenges when faced with adversarial legal proceedings and the construction industry is no different. With separate industries relying on differing data sources and legal events involving data drawn from any source or record types with potentially relevant materials; from structured timekeeping, shipping, and computer-aided design (CAD) or building information modeling (BIM) files to unstructured email, document management systems, and handwritten reporting (e.g., daily construction reports), costs quickly become significant when clients and their outside counsel are faced with reviewing these materials in a traditional linear manner. Keyword searches can begin to cull the data that requires legal review, but risk reducing defensibility and make it harder to answer the question: "Did we find everything?"

To assist with these challenges, corporations and legal teams are utilizing advanced technology, including predictive modeling, near-duplicate detection, and concept clustering, to minimize review, limit costs, and maintain defensibility. As the data evolves, new tools are being explored such as sentiment analysis to categorize and gain understanding into these evolving data sets.

1. The Radicati Group, Inc., "Email Statistics Report, 2015-2019," The Radicati Group, Inc. (2015), https://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf.

PREDICTIVE MODELING

Predictive modeling, also known as predictive coding, is a form of supervised machine learning technology that seeks to identify targeted documents in a population. The software makes predictions regarding the overall data population by utilizing training data derived from the population at issue and can be used to identify relevant material, perform quality control of production, or privileged document populations, and assist with categorizing data by matter-specific issues. This process allows for significant time saving as well as potential cost saving of upward of 50 percent related to document research and review when large volumes of data are at play.

These processes rely on the text present in the data, not the data type or source, thus eliminating the hurdle of conforming data types to a single standard and empowering a legal team to analyze not only client data but the adversarial party's production data as well. When text is not present, (e.g., certain CAD or BIM files), it must be noted that modeling cannot be successfully performed and will require legal team review.

Predictive modeling requires limited human interaction by the legal team's subject matter experts (SMEs). The SMEs will review a subset of the entire population for relevance (training set) and this subset will be used to teach the program, which will then make predictions about each document in the corpus based on the teaching. Once the process is complete, the SMEs validate the model's accuracy by reviewing a statistically valid sample to determine the model's accuracy (validation set). The SME's interaction with the training and validation sets contrasts with the traditional linear review approach of reviewing each document in the corpus at great cost in both time and expense. This approach provides the legal team with more time to focus on the merits of the matter, while not being bogged down by extensive reviewing of multiple irrelevant documents.

As an example of this technology in use, a prominent utility provider had hired a company to build its new flagship manufacturing facility. Upon taking possession of the factory, the provider identified a long list of deficiencies with its construction that hampered its ability to operate efficiently. The client entered arbitration seeking monetary relief for the cost of correcting the deficiencies and for the burden of operating at reduced output. An expert was engaged to assist with the identification of delays, completeness of project turnover, damages from the deficient facility, and provide expertise regarding the discovery process.

The client had a repository of over 1.3 million documents drawn from multiple sources (e.g., email, schedules, CAD files, reports) and was required to produce relevant documents from the repository within a 10-week time frame. With limited legal team resources (four or five attorney reviewers) to conduct a full review of the repository, the review was projected to take over eight months to complete.

The expert deployed predictive modeling to analyze the text of the documents and create multiple predictive models to identify relevant documents and potentially privileged documents. The models classified the material and assisted the legal team's review by identifying relevant material, facilitating the removal of irrelevant material from the review workflow, as well as providing quality control of the ultimate production population to verify that potentially privileged material would not be inadvertently produced. This resulted in the legal team meeting the arbitration panel's deadline without the need to add large numbers of additional legal staffing.

Predictive models were also created to categorize the opposition's production. These models were set up to supplement traditional keyword searching to categorize the produced data and identified key documents related to the various deficiencies at issue. In very short order, the team had a deep understanding of the material produced by the opposition, which allowed for greater time to formulate their legal arguments.

Having faced the prospect of hiring additional legal staff to complete review of all material by the arbitrator's deadlines, the legal team leveraged predictive modeling to save the client from hiring additional legal staff, as well as saving hundreds of thousands of dollars in review costs.

NEAR-DUPLICATE DETECTION

Near-duplicate detection technology is a form of machine learning that identifies near-duplicate or duplicate documents based on the textual content of the documents. By evaluating and comparing the text present in each document in a population against one another, near-duplicate detection groups documents based on textual similarity. Contrary to predictive modeling's supervised learning, near-duplicate detection utilizes unsupervised learning, which does not require human intervention or training. This process allows for analysis as soon as data is available.

The technology enables greater quality control over a legal team's review of documents. In very little time, a team can determine if the relevance or privilege of a document is in line with similar documents. Issue research is improved with the ability to quickly jump to documents similar in content to investigate related information. The technology enables legal teams to validate privilege determinations across groups, find documents related to specific issues, and perform first-pass review based on similarity to speed up workflows.

Recently, a legal firm utilized this technology to research version history of documents (e.g., contracts, meeting minutes, agendas). The client was faced with searching across 39 million records in an effort to research the evolution of certain contracts. Near-duplicate detection scanned the entire population and successfully created groupings of near-duplicate documents. This allowed a single member of the legal team the ability to quickly drill down into similar documents and explore the history of the documents in question.

In addition to utilizing the technology to research and identify similar documents, it was leveraged to cull down the data population requiring review. The near-duplication detection identified textual duplicates in the overall population. By reviewing one copy of the document, all other copies deemed as duplicative did not require review and were eliminated from the review population. This workflow also ensured that consistent legal determinations were applied across duplicative copies.

CONCEPT CLUSTERING

Another form of unsupervised machine learning is concept clustering. This analytics method utilizes algorithmic processes to analyze the text of a document population and categorize, or cluster, the population into conceptually similar groupings. In contrast to keyword searching, this approach compiles and analyzes not just single words but the surrounding words in the text as well. This enables the software to conceptualize what terms relate to others to form more accurate clusters.

Concept clustering attempts to account for changes in meaning and context as well. For example, the term "column" may be used in an email to reference information in an attached Excel chart, or it may be used to reference the support structure of a building. The term "foundation" may reference a corporate goodwill initiative or the support base of a structure. In both cases, concept clusters aim to segregate documents based on meaning.

This can be beneficial in the early stages of a matter, when legal teams may not yet have a deep understanding of the subject matter of the documents at issue. Concept clustering can quickly and efficiently analyze a population and cluster the data for the legal team to investigate. Members of the legal team are then able to quickly drill down into documents associated with each concept, discard irrelevant material, and prioritize research and review.

A recent example of this technology in use revolves around allegations of potential violations of the Foreign Corrupt Practices Act by a construction equipment manufacturer. The client initiated an internal investigation as soon as the allegations arose and the data from more than 100 employees was collected. In this time-sensitive matter, the client needed to quickly gain insight into the data at issue to formulate a review strategy. Concept clusters were created using the data of each employee and counsel was able to quickly identify the issues and topics the employees were involved in. Using these concept clusters to prioritize their review and sample the data, the legal team was able to identify seven potential bad actors and prioritize their data for review, remove the data of over 30 employees from the review population, and identify key information in the investigation in the first week the data became available.

EMERGING TECHNOLOGY

As business innovation continues, new systems for collecting and storing information evolve, and so must the methods of triaging this data. One method being explored is sentiment analysis, or opinion mining, which has the potential to help identify relevant materials and categorize data populations as well as connect the dots for legal teams.

Sentiment analysis is the process of computationally identifying and categorizing language used in a piece of text as positive, negative, or neutral. Employed extensively in the marketing industry to determine subjective opinions of current and potential product buyers, its adaptation to the legal community is being explored as a means of gaining greater insight into employees' opinions surrounding the time period of a litigious event.

With key opinions regarding project status, interruptions, shipping delays, or safety issues expressed across a corpus of potentially hundreds of thousands or millions of data points, the ability of multiple members of a legal team, reviewing separate documents at separate times, to connect the dots is low. Sentiment analysis may be able to scan large sets of issue-specific documents (e.g., daily foreman reports, key individuals' emails, or chats) to gain an understanding into how a project progressed and the subject opinions of key stakeholders.

CONTACTS

ROBERT NEARY

Associate Director,
Global Legal Technology Solutions
+1.202.481.8385
robert.neary@navigant.com

AMY TSANG

Associate Director,
Global Legal Technology Solutions
+1.646.227.4728
amy.tsang@navigant.com

navigant.com

About Navigant

Navigant Consulting, Inc. (NYSE: NCI) is a specialized, global professional services firm that helps clients take control of their future. Navigant's professionals apply deep industry knowledge, substantive technical expertise, and an enterprising approach to help clients build, manage, and/or protect their business interests. With a focus on markets and clients facing transformational change and significant regulatory or legal pressures, the firm primarily serves clients in the healthcare, energy, and financial services industries. Across a range of advisory, consulting, outsourcing, and technology/analytics services, Navigant's practitioners bring sharp insight that pinpoints opportunities and delivers powerful results. More information about Navigant can be found at navigant.com.

linkedin.com/showcase/NavigantGlobalConstruction

twitter.com/navigant

CONCLUSION

As data volumes continue to increase at an incredible rate, and industry-specific data types continue to diverge, new tools are needed to stem the tide of increasing costs associated with investigating, analyzing, and reviewing these populations in response to a legal event.

Data analytics tools empower corporations and their outside counsel with greater insight into massive data populations, while allowing significant savings in cost and time. They aid in categorization, relevance determination, privilege consideration, issue identification, and quality control. Predictive modeling enables legal teams to tackle very large data sets by removing irrelevant documents quickly, conducting large-scale reviews without the cost of hiring legal reviewers, performing quality control, pushing the potentially most important documents to the front of the review, and analyzing opposition party data at a fraction of the cost. Near-duplicate detection provides quick reference for similar documents and bolsters quality-control methods. Concept clustering offers the ability for legal teams to investigate the data surrounding a matter quickly and gain an understanding of the types of material in play.

Used together or apart, these machine-learning technologies can analyze large amounts of data in a fraction of the time it would take using traditional approaches, and strengthen a legal teams' ability to quickly find the information most relevant to a matter.

Note: Most of these examples above come from actual Navigant experience. Navigant has a dedicated group of individuals who are experts in advanced data analytics that service satisfied clients.