Artificial Intelligence and Transactional Law: Automated M&A Due Diligence

By Ben Klaber

Introduction

Largely due to the pervasiveness of electronically stored information (ESI) and search and retrieval technologies, discovery has changed rapidly. Whereas attorneys previously waded through thousands of paper documents, intelligent ESI management systems are now commonly used. For litigators, these disruptive technologies are changing the legal services landscape.

Similarly, attorneys currently review hundreds or thousands of documents as part of the due diligence process for mergers and acquisitions (M&A). The disclosed documents are used to identify and allocate risks and to establish subsequent steps in the transaction. Disclosure mistakes can lead to substantial liability from violations of representations and warranties in the parties' agreement. Although M&A due diligence typically involves far fewer documents than e-discovery, these documents – and the provisions within them – may be sufficiently consistent across transactions to warrant the use of automated search and retrieval technologies.

This paper proposes that automated models, particularly Reflective Random Indexing, can add considerable value to the M&A due diligence process. This paper suggests that an integrated human and machine-learning process can identify, classify, organize, prioritize and highlight documents, which must be disclosed pursuant to some type of business combination agreement (e.g., stock purchase agreement), with higher efficacy and speed and lower cost than humans can alone. Such a process should be particularly effective for standard, clear terms. More advanced methods or combinations of models may be necessary to identify documents that are responsive to more specific, uncommon or ambiguous terms. A model could be trained and tested on a corpus of written agreements to identify documents that are responsive to common

M&A representations and warranties. Given the immense risks and legal costs involved in large M&A transactions, a model that leads to even modest improvements in due diligence efficacy and efficiency could prove to be a very valuable product.

Due Diligence Process

The primary objective of M&A due diligence is to identify potential problems with the transaction. The acquiring company (Acquirer) will want sufficient disclosure from the target company (Target) to allocate risk between the parties and to acquire a thorough understanding of the Target before taking over the business. Due diligence also guides the transaction.

Though circumstances vary considerably, the due diligence process for mergers and acquisitions typically proceeds as follows. After the parties sign a confidentiality agreement, the Target gathers potentially relevant company documents. This task can be very difficult because documents may be scattered across many locations, leading to important information being overlooked or difficult to find. The Target typically needs to digitize many physical documents.

Next, the Acquirer sends the Target a disclosure checklist of the types of documents and information that should be uploaded to a virtual data room. In response to the disclosure checklist, the Target sends documents to its outside counsel (Target's Counsel). Because of confidentiality and attorney-client privilege concerns, the Target's Counsel reviews the documents before they are uploaded to the data room. Other than managing such sensitive information, however, the Target's Counsel is not exceedingly cautious about disclosure at this stage in the process because no representations or warranties relating to the documents have been made. In fact, sophisticated clients often independently handle much of early due diligence.

Upon obtaining access, the Acquirer and its outside counsel review and analyze the documents to identify significant risks. There are two general categories of risks. First, there

could be pre-existing business, financial or liability issues. Second, material consequences could flow from the transaction due to anti-assignment, change-of-control or confidentiality provisions, for example. In addition, the business combination could raise antitrust concerns. Once the parties are ready to proceed with the transaction, they sign a letter of intent.

Throughout the due diligence process, the parties negotiate the terms of the business combination agreement (Agreement). In the Agreement, the Target typically makes several representations and warranties, subject to specific exceptions that are disclosed in schedules. Any exception that is not specifically disclosed constitutes a breach of the Agreement. Therefore, the Target's Counsel will insist on carefully drafted, narrow representations to lessen disclosure requirements and reduce potential exposure to substantial liability.

Finally, The Target's Counsel reviews the documents in the data room and classifies them for the disclosure schedules. At this stage of the process, accuracy is very important because of the potential liability mentioned above. In addition, the Agreement typically requires that all documents referred to in the disclosure schedules continue to be in full force and effect.

Tasks Ripe for Automation

There is currently a strong need to reduce the burden on a Target involved in an M&A transaction. Early in the due diligence process, the Target needs to identify potentially relevant documents (Task 1). Because the Target assumes liability for documents that are withheld in violation of the Agreement, and because complete and accurate disclosure informs and expedites negotiations, an automation-supported method for categorizing documents could add substantial value. Moreover, material information may be found in unexpected places such as emails rather than formal written agreements. For Task 1, the Target would likely employ a model with greater emphasis on recall rather than precision. During this early stage in the due diligence

process, the Target generally errs on the side of disclosure because the confidentiality agreement is in effect, and the Target has not yet made affirmative representations. However, the Target should not disclose any unrequested documents with attorney-client privilege implications.

Later in the due diligence process, the Target would benefit from an automated model that would classify documents according to the disclosure schedules. Such an algorithm would be particularly valuable because of its ability to catch relevant provisions in unusual documents and locations. For example, a restrictive covenant, such as an anti-assignment provision, might be contained in an email rather than as part of a formal, integrated written agreement. In addition, an automation-supported process could increase efficiency and effectiveness, which is normally hindered by complexity. Automation-supported highlighting and categorizing of potentially relevant provisions could ease the difficult task of identifying and cataloguing every applicable provision in each and every document. For this more nuanced task of assigning documents to disclosure schedules (Task 2), the Target would likely need a more precise model. Unlike Task 1, the model should be tuned in favor of precision over recall because the highlighting, organizing and prioritizing functionality would significantly boost productivity, because all documents would receive a probabilistic score, and because attorneys would continue to at least manually scan the documents.

Proposed Approach

At the outset, it should be noted that even basic statistical search techniques, combined with intuitive human-machine interaction, could be beneficial for practitioners. For the basic information retrieval task of producing a ranked list of documents in response to a query, there "is no evidence that detailed meaning structures are necessary." [4] For powerful results that could disrupt the legal services industry, however, the following approach is proposed.

Word space models "use distributional statistics to generate high-dimensional vector spaces, in which words are represented by context vectors whose relative directions are assumed to indicate semantic similarity." [3] Random Indexing (RI), a scalable alternative to Latent Semantic Indexing (LSI), is an incremental word space model that involves the allocation of elemental vectors followed by training. [1] RI reverses the traditional dimension reduction process by first accumulating context vectors, typically based on documents and terms, and then by constructing a co-occurrence matrix. Due to greater efficiency and parallelization, RI offers a substantial computational advantage over Singular Value Decomposition, enabling scalable dimension reduction. In addition, the semantic space of document and term vectors can be built incrementally as new information is added. However, RI falls short on making meaningful indirect inferences, particularly when terms do not co-occur in any document. Fortunately, RI can be adapted to employ an iterative, cyclical training method known as Reflective Random Indexing (RRI), where "the system generates new inferences by considering what it has learned from a data set in a previous iteration." [1] RRI is a more efficient and flexible way to achieve the same dimensionality reduction as LSI. An RRI algorithm can be tuned and iterated rapidly; because of its reliability and speed, RRI could be used as "an interactive, exploratory tool during early data analysis, culling, analysis and review phases." [2]

An RRI model "builds a set of semantic vectors, in one of several variations – term-term, term-document and term-locality." [2] A training cycle consists of using document vectors to generate term vectors and vice versa. Each term vector represents a condensed version of the applicable documents, and each document vector summarizes the significant terms in the document. The aggregation of these vectors represents the "semantic nature of related terms and documents." [2] The semantic vector space is organized in clusters, enabling directed searches

to improve efficiency. This paper asserts that compelling results could be obtained by integrating an RRI model into the M&A due diligence process.

Implementation

Task 1 and Task 2 involve identifying and classifying documents that must be disclosed as part of the M&A due diligence process. RRI could be utilized to accomplish both tasks, greatly increasing the efficiency and effectiveness of such document review.

There are several disclosure requirements that are essentially standard for most M&A transactions. It is highly likely that a model could be successfully trained to recognize relevant documents for the more standardized and unambiguous terms in the Agreement. The model could be trained on many examples of these substantially similar provisions. Some common examples are confidentiality, non-competition, infringement, indemnification, most-favored-nation, dispute resolution and change-of-control provisions. In addition, corporate organizational documents and employment, license, settlement, exclusivity, joint venture and distribution agreements may fall into this category. Regulatory filings and communications may also be recognizable. A model that can correctly identify the responsive documents for relatively standard, clear terms would be useful across transactions without the need for much individual transaction customization. For some other provisions, their positions along the ambiguity and standardization spectrums may vary more depending on the circumstances. For example, definitions could differ for "contracts of indebtedness" or "outstanding or pending litigation."

The second category of terms is very similar to the first except that these terms involve some specific information. For example, the Agreement might not require disclosure of distribution agreements unless they are above a minimum threshold value, such as $100,000 per year. Notwithstanding such a provision, however, the Agreement might require disclosure of all

distribution agreements (and many other contracts) involving specific periods of time, particularly those beginning after the Target's most recent financial statements or ending before or soon after the expected closing. Some other common examples are specific employment agreements and insider transactions involving specific related parties. As the number of documents increases, the model's ability to distinguish between similar and responsive documents should increase using an RRI approach. However, these terms are more specific to an individual transaction. Therefore, the model may not have sufficient data to make accurate determinations, or the costs of training the model may outweigh the benefits of automation.

The final, and most difficult to model, category along the spectrums is comprised of ambiguous and not highly standardized terms. For such terms, the model may be able to identify potentially relevant documents but have difficulty analyzing nuanced concepts. For example, the Agreement would most likely include a representation and warranty that the Target is not a party to any "material" contracts other than those specifically disclosed in the corresponding schedule. For such vague terms, the model might need to engage in a more sophisticated analysis to reach responsiveness determinations. Although "material" is a term of art in securities regulation, and although the model could be trained on subsets of materiality based on detailed definitions, the model would need to be iterative because materiality is often highly dependent on context.

Because terms of the Agreement are revised during negotiations and because terms are similar across transactions, a successful model must be iterative, continually accommodating new documents, objectives and responses. For example, the Acquirer might initially request disclosure of all employment agreements. The parties might then revise this term, and the corresponding disclosure schedule, to only senior employees, such as vice presidents and above. RRI offers this flexibility, allowing for fast and easy updated results.

The queries would be derived from a due diligence checklist, the terms of the Agreement and the disclosure schedules. Query formulation is an important factor in the efficacy of IR systems. Queries could be pre-loaded in the software, and users could make example provisions the queries themselves, taking advantage of the standardization of many terms. For each query, the proposed model would score each document as a function of the applicable document vector. Utilizing this score, the model would rank all of the documents for each query to facilitate prioritization. The model would organize all of the queries and results so that some documents would be located in multiple labeled query clusters. Finally, the model would highlight the most relevant words in each document based on the term vectors, enabling faster manual review. The software could utilize different colors for different queries so that users could evaluate multiple queries at the same time. With all of the above functionality, the user could easily manually label a document as responsive or not, and the software would reorganize the document accordingly. Except for highly unusual terms in M&A agreements, the model would continually improve with every transaction.

To maximize efficacy and adoption, the proposed model could be developed, tested and marketed by a virtual data room provider. Virtual data rooms offer transacting parties a secure online location to store and review confidential documents. Data room providers have access to a rich data set of millions of documents, and responsiveness functionality could be integrated to improve users' efficiency. For example, attorneys could mark documents as responsive to certain terms, and this label could attach to the document along with the attorney's identification. With this added functionality, data room providers could utilize attorney responses to train the model and further its development. Data room providers would need to negotiate such testing with customers, however, especially because many of the documents are confidential.

Although partnering with a data room provider is proposed, there are at least two other promising avenues for testing. First, a large law firm that is repeatedly involved in M&A transactions would offer many of the same benefits without the confidentiality problem, because clients depend on outside attorneys to become thoroughly familiar with their confidential documents. The model could be trained on manual attorney responses during real M&A deals. Second, a researcher could utilize the large Bloomberg "DealMaker" database, which capitalizes on the fact that many agreements involving publicly traded companies are publicly disclosed in the EDGAR database. DealMaker has pre-classified agreements and advanced search capabilities. The model could be trained on a subset of this corpus of pre-classified documents to recognize the most common provisions and agreements.

Future research should focus on realistic testing and user-friendly design. To encourage wide acceptance of this new product, the model should be easily integrated into the current due diligence process. Once attorneys and their clients come to depend on the model, solutions could be developed for more advanced due diligence tasks.

Conclusion

Many expensive hours are spent reviewing documents for the vast M&A market. The due diligence process could be largely automated, leading to cheaper, faster transactions with better risk management. Reflective Random Indexing offers an elegant, efficient solution to the challenge of classifying, organizing, prioritizing and highlighting corporate documents. Additional, complementary techniques could also be used. Thanks in large part to advances in e-discovery, M&A due diligence tasks are ripe for automation. Even with modest implementation of this proposed approach, law firms and their clients could realize significant gains.

References

[1] Trevor Cohen, Roger Schvaneveldt, Dominic Widdows, *Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections*, Journal of Biomedical Informatics 43 (2010) 240-56.

[2] Venkat Rangan, *Discovery of Related Terms in a Corpus using Reflective Random Indexing*, DESI IV (2011).

[3] Magnus Sahlgren, *An Introduction to Random Indexing* (2005).

[4] Ellen M. Voorhees, *Natural Language Processing and Information Retrieval* (1999).