Emmanuel Alvarez, ea354
Law and Technology

Predictive Coding: The New E-Discovery

Introduction

Predictive coding is the new innovation in electronic discovery. The discovery process has evolved greatly in the last two decades. Previously, discovery was limited to paper documents such as company accounting books, employment rosters, financial disclosure documents, and contracts. Since then, the progression of communication and information has reached a complex level. The creation of the Internet, innovations in telecommunications, and the cloud based computing produces potential evidence in many forms. Evidence is no longer solely in paper format. Electronic information is also not exclusively in the form of email communications, word documents, etc. Electronic evidence takes the form of internet protocol addresses, cookies, and hidden code within emails and other forms of electronic stored media.

The search for evidence in discovery has gone from a needle in a haystack to a photon sized electric charge in space. It is difficult for firms and investigatory institutions to find documents efficiently and accurately. There has been a large increase in the number of man-hours necessary for document review and discovery. The increase in man-hours has caused the cost of discovery to increase as well. The discovery process has responded by creating an efficient, cost effective method to keep up with the ever-evolving technological innovations of today.

Predictive coding is the most recent and highly publicized method of electronic discovery. Predictive coding is a computer-assisted review of electronic data where electronic learning technology classifies and categorizes electronic data. This process is a much more sophisticated method evolving from keyword and concept searching previously used for electronic discovery. Predictive coding is argued to be cost-effective in reducing the man-hours by increasing the "recall and precision"[1] produced by electronic discovery software. Yet, there is hesitation in relying on electronic software to analyze the relevance of documents or categorizing them as reliable "hot documents."[2]

This Paper will attempt to explain the recent innovation of predictive coding. It will detail the cost effectiveness and profitability of such technology for the investigatory process of firms and entities, as well as the potential drawbacks. The Paper will compare predictive coding to the older methods of electronic discovery, and will then conclude by discussing potential legal issues predictive coding may face.

[1] William A. Ruskin, *Predictive Coding: Will E-Discovery swallow the Judicial System?*, LexisNexis (Feb. 5, 2013, 04:53pm), http://www.lexisnexis.com/legalnewsroom/litigation/b/ediscovery/archive/2013/02/05/william-a-ruskin-predictive-coding-will-e-discovery-swallow-the-judicial-system.aspx.

[2] Hot documents are the essential that firms look to find as evidentiary documents that could be further in investigation or litigation. Allison Nadel, *E-discovery: The value of Predictive Coding in Internal Investigations*, Inside Counsel (Aug. 13, 2013) http://www.insidecounsel.com/2013/08/13/e-discovery-the-value-of-predictive-coding-in-inte.

Background

Discovery is a costly and expensive process, costing 50 to 70 percent of the litigation budget.[3] Prior to electronic discovery, attorneys and individuals used an "eyes on"[4] review. This "eyes on" approach is still required today, and electronic investigations cannot be left completely to software and computers alone.[5] Evidence must be reviewed for relevance, privilege, and other such challenges arising during the legal proceedings.

There is an endless amount of electronic stored data to be reviewed. Approximately 93% of data available today is created electronically, and only about 30% of the data is ever printed to paper.[6] The amount of data out there has exploded because of the ease with which documents may be composed electronically. The ability to find specific data amongst the endless terabytes requires man hours mixed with electronic assistance. Discovery has not reached the level of artificial intelligence where it may be solely left to computers to judge relevance, privilege, and other such complex legal issues.

Prior to predictive coding, the legal industry used various processes for electronic discovery. The first prominent method used is key word searching. Key word searching is similar to using a search engine. This method is used to find relevant key words and phrases attorneys deem relevant to the case, or they speculate to be in potential evidence. Attorneys or investigators come up with a list of relevant key terms or phrases they believe would be found in potential evidence. Key word searching is argued to not be efficient, because only about 20% of relevant data on average is extracted using this process.[7]

Key word electronic discovery has various issues it faces. Keyword searching is based on an exhaustive collection of phrases and terms. This makes it difficult when such terms produce multiple suffixes or can be expressed in various manners. For example, I previously worked with the FBI Regional Computer Forensic Laboratory in Silicon Valley where they used key word searches. There is a struggle with this process when modifications to names, bank names, or even account numbers in financial investigations are made. Specifically, we had trouble with a case where money was laundered through Switzerland. We found we needed to include terms such as "Switzerland, swiss, Geneva, and GN." The recall on such searches was endless, and pulled up irrelevant information when searching through twenty plus computers seized from the defendants. The hefty recall the searches brought back requires man-hours to sort through to identify and preserve the documents. A mere ability to search electronics for proposed terms and phrases is not efficient.

A second method of electronic discovery is concept searching which expanded from key word searching.[8] Concept searching is based on concepts from a list of terms and phrases.

[3] Eric Seggebruch, *Electronic Discovery Utilizing Predictive Coding* 15, *available at* http://www.toxictortlitigationblog.com/Disco.pdf.

[4] Human individuals assessing the documents relevance.

[5] William A. Ruskin, *Predictive Coding: Will E-Discovery swallow the Judicial System?*, LexisNexis (Feb. 5, 2013, 04:53pm), http://www.lexisnexis.com/legalnewsroom/litigation/b/ediscovery/archive/2013/02/05/william-a-ruskin-predictive-coding-will-e-discovery-swallow-the-judicial-system.aspx.

[6] Eric Seggebruch, *Electronic Discovery Utilizing Predictive Coding* 3, *available at* http://www.toxictortlitigationblog.com/Disco.pdf.

[7] *Id. at 4.*

[8] Allison Nadel, *E-discovery: The value of Predictive Coding in Internal Investigations*, Inside Counsel (Aug. 13, 2013) http://www.insidecounsel.com/2013/08/13/e-discovery-the-value-of-predictive-coding-in-inte.

Unlike key word searching, this list need not be exhaustive. The software then uses concept similarities from a developed series of algorithms. [9] This highly technical method relies heavily on properly set up algorithms and a well set up list of terms and phrases. This is strongly related to clustering and discussion threading, which also strongly relies on algorithms to group together, documents based on document type (eg. Emails), language, file size, etc.

The newest process of electronic discovery is predictive coding. Predictive coding uses a human element to produce a set of seed documents the software uses to categorize and analyze electronic data.[10] The software relies on human decision making to begin a set of classifications brought together into concepts similar to concept searching. The software searches are based on human judgments of coding used for subsequent rounds of predictive coding.[11] The software produces a set of responsive documents. The human component either accepts or rejects the responsive documents to optimize the algorithm for future searches. This process can be compared to liking or disliking a song on Pandora Radio. Yet, responsive documents are given a confidence score to further perfect production along the line. The program explains why responsive documents are returned. The seed set continues to grow and expand with the selected responsive documents. Unlike simple concept searching, it does not auto-code, but uses a better balance of human judgment with technological assistance.

Predictive coding: the ups and downs

Predictive coding is very organized and efficient. Predictive coding organizes responsive documents into topic areas. Attorneys can then sift through documents by topics rather than just a stack of responsive documents. Topic sorting proves profitable to law firms in the organization of litigation. Litigation is not just an in court process, but a series of investigatory steps to put a case together. Predictive coding allows attorneys to set up the litigation process by efficiently organizing electronic discovery to specifically set up for individual depositions, financial analysis, and any other topics relevant to the case at hand. Predictive coding also reduces the amount of man-hours necessary to sort through, organize, and select documents as relevant. Predictive coding overall provides better recall and precision, more quickly than the traditional methods.[12] Unlike, the previous methods, it does not require exhaustive lists, or extensive preparation prior to the program producing responsive documents.

[9] Concept similarities may include grouping together finance documents, documents relating to banks, or any such information that would be relevant to a particular term or idea. For example, if the phrase "money transfer" was introduced, the program would feed back documents such as bank documents, money transfer related communications, etc. Allison Nadel, *E-discovery: The value of Predictive Coding in Internal Investigations*, Inside Counsel (Aug. 13, 2013) http://www.insidecounsel.com/2013/08/13/e-discovery-the-value-of-predictive-coding-in-inte.
; *Predictive Coding 101 & The Litigator's Toolbelt*, e-discovery 2.0 blog, (Dec. 5, 2012), http://www.clearwellsystems.com/e-discovery-blog/tag/concept-search/.

[10] *Predictive Coding 101 & The Litigator's Toolbelt*, e-discovery 2.0 blog, (Dec. 5, 2012), http://www.clearwellsystems.com/e-discovery-blog/tag/concept-search/.

[11] Eric Seggebruch, *Electronic Discovery Utilizing Predictive Coding* 8, *available at* http://www.toxictortlitigationblog.com/Disco.pdf.

[12] *See* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. &. TECH. ll (201l). http://jolt.richmond.edu/vl7i3larticlel l.pdf.

The cost effectiveness is very appealing to firms. Predictive coding still follows the same electronic discovery steps used in key word and concept searching: Identification, preservation, collection, review, analysis, and production.[13] The Keyword searching I performed while working with the FBI took months to format and search electronic data based on the terms and phrases. The number man-hours necessary to sort through the large amount of responsive data to identify relevant documents to be collected, preserved, and produced to opposing counsel through key word searching was costly. Key word searching is usually not effective, and takes an ongoing game of "go-fish" using words.[14] Predictive coding reduces the "eyes on" on review of electronic stored information. Although the efficiency of predictive coding compared to the previous methods of electronic discovery is not given an exact figure, it has been shown to produce recall with better precision.[15]

Courts have faced legal issues with predictive coding. Predictive coding must meet the standard of reasonableness in the production of discovery.[16] Predictive coding may be cost effective, but it must produce reasonable results under a request for discovery.[17] Rule 26(g) requires counsel to make a good faith reasonable inquiry in discovery production.[18] The burden is on the party challenging the method of electronic discovery to prove such a method is unreasonable under Rule 26 of the Federal Rules of Civil Procedure.[19]

Predictive coding is reliable in so far as the human component is reliable and reasonable. Improper coding and wrongful selection of responsive documents will undermine predictive coding from seeding to the final output. The same way a mistaken like on your Pandora station may play a Justin Bieber song on your Wu-Tang station, a failure by the human component in predictive coding may produce responsive electronic data completely irrelevant to the case. In a current court case the Court stated electronic discovery without testing and refinement may not be "reasonably calculated to uncover all responsive material".[20] Therefore, predictive coding is only reasonable when there is a good faith, reasonable human component testing and refining the predictive coding process. However, the legal challenge to reasonableness usually comes from the counsel not using predictive coding.

[13] Allison Nadel, *E-discovery: The value of Predictive Coding in Internal Investigations*, Inside Counsel (Aug. 13, 2013) http://www.insidecounsel.com/2013/08/13/e-discovery-the-value-of-predictive-coding-in-inte; Gordon v. Kaleida Health, 2013 U.S. Dist. LEXIS 73334, 2013 WL 2250506 (W.D.N.Y. May 21, 2013).

[14] National Day Laborer Organizing Network et al. v. United States Immigration and Customs Enforcement Agency, et al., 2012 WL 2878130, at *11 (S.D.N.Y. July 13, 20l2).

[15] The efficiency value of predictive coding may not be produced because it is a very modern and technical method of electronic discovery. Although predictive coding may produce more relevant documents on a whole, it could possibly be overlooking potential evidence just as keyword searching does.

[16] Eric Seggebruch, *Electronic Discovery Utilizing Predictive Coding* 20, *available at* http://www.toxictortlitigationblog.com/Disco.pdf.

[17] Sawhorse Enters. v. Church & Dwight Co., 2013 U.S. Dist. LEXIS 48155 (D.N.J. Apr. 3, 2013) (citing *Entertainment Technology, Corp. v. Walt Disney Imagineering,* No. 03-3546, 2003 U.S. Dist. LEXIS 19832, 2003 WL 22519440 at *3 (E.D.Pa. October 2, 2003)); *Gucci America, Inc. v. Daffy's, Inc.,* No. 00-4463, 2000 U.S. Dist. LEXIS 16714, 2000 WL 1720738 (D.N.J. Nov. 14, 2000); *Philadelphia Newspaper Corp. v. Gannett Satellite Information Network, Inc.,* No. 98-CV-27821, 1998 U.S. Dist. LEXIS 10511, 1998 WL 404820 (E.D.Pa. July 15, 1998).

[18] *Id.*

[19] Larsen v. Coldwell Banker Real Estate Corp., 2012 U.S. Dist. LEXIS 12901, 2012 WL 359466 (C.D. Cal. Feb. 2, 2012)

[20] National Day Laborer Organizing Network, at 40-42.

The reasonableness standard is a good standard for predictive coding, but it impinges on the court and opposing counsel a hurdle in judging the reasonableness of predictive coding. Inexperienced opposing counsel has difficulty in predictive coding, to challenge and prove the process to be unreasonable because they do not have the means to be use or be expert in predictive coding. Small firms cannot judge off hand whether the proper software or seed documents were used in the process. Courts, as well as opposing counsel, would have to be expert enough to deem the set of seed documents, the accepted and rejected responsive documents, and the testing and refinement process as unreasonable before challenging them in the production of discovery. To deem the process reasonable then, the court would need to analyze the entire process used, including both the human and computer component. This is counter to judicial economy and requires a certain expertise in the process. Large firms spend endless amounts of money on such expertise, and it is costly to counsel for experts to come in testify in regards to the process. Small firms and courts thus need the funds to be trained in the process and be able to participate in the discovery process. Large firms may also face a challenge if they do not properly assess the seed documents and recall, and the whole process is ruled unreasonable in court. Large firms must reasonably and diligently move through the process of predictive coding. Predictive coding may save money in the discovery process, but it is still very expensive to reach the level of expertise to program and fine-tune the predictive coding process. The courts and small firms may not be able to meet such a level of expertise.

Nevertheless, predictive coding has proven to be more efficient, cost effective, and able to produce better results.[21] The legal realm is often left behind in technological innovation. Predictive coding has been ruled to meet the standard of reasonableness set by the courts, and should not be held back because it is highly technical. The courts may assess the technology industry, but experts in predictive coding may testify and assist the courts and firms in legal innovative progress. This trial process may face a couple of hardships in adopting the innovation, but overall the improvement to litigation in this highly technical era, is cost-effective and efficient in the long run. Furthermore, courts and opposing counsel should not dictate the method of discovery used, unless they can prove such methods fail to produce reasonable results.[22]

Conclusion

Predictive coding is currently the most efficient and cost effective method for electronic discovery. Predictive coding combines a good balance of human and computer components to continuously provide a set of seed documents to better assist and find conceptual relevance between potential discovery documents. The organization and method of topic categories reduces the costs for litigation overall. However, the continuous reasonableness standard used by the courts must still be met. With the more elaborate predictive coding, the court and small firms may have difficulty in judging whether the process was reasonable. This may cause potential hurdles, or even potentially changing the standard from reasonableness to a more expert standard. Nevertheless, the innovation and progress it brings to the discovery process cannot be denied as it makes the process quicker, more efficient, and most cost effective.

[21] Kleen Products LLC v. Packing Corp. of America, No. 10 C 5711 at 287, 290, 303 (NI.D. Ill, Feb 27, 2012).
[22] Id.