Survey of the Use of Predictive Coding in E-Discovery

Julie King

CSC 570

May 4, 2014

ABSTRACT

Predictive coding is the latest and most advanced technology to be accepted by the legal community for use in e-discovery. It is an iterative process involving relevance feedback and supervised machine learning methods to rank documents' potential responsiveness, allowing review to proceed much more quickly, yet with quality equal to or better than traditional e-discovery methods. There are several hurdles predictive coding must overcome to be fully accepted as the new "gold standard" of electronic discovery, but recent events indicate this will happen soon due to the reduction in costs and improvement in accuracy predictive coding offers.

STATEMENT OF THE PROBLEM

E-discovery, or electronic discovery, is the process of discovering and analyzing in electronic form documents and other items (videos, recorded phone calls, etc.) responsive to a request for information during the discovery phase of litigation. The electronic files for these items are referred to as Electronically Stored Information ("ESI") [Baron and Thompson 2007] or generically as "documents." Initial collections of potentially responsive ESI can contain multiple millions of files, each of which needs to be evaluated. Courts know that evaluating millions of files will never lead to production of 100% accurate results, so the standard for production is a "reasonable inquiry" rather than a "perfect" outcome [Roitblat et al. 2010; The Sedona Conference 2013b].

With e-discovery, unlike some information retrieval tasks (e.g., a web search), the task is usually to find ALL responsive ESI, not just the most responsive ESI. Also, litigation requests are designed to be very broad, such as "all documents reflecting or referencing the development of the drug [drug name]" [Roitblat et al. 2010]. Thus, search and analysis methods tailored to the e-discovery industry are desirable. It is not surprising, then, that e-discovery software is a booming business. Estimates put 2013 revenues of e-discovery vendors at $1.5 billion [Katz 2013]. Costs for e-discovery for just one case can be in the millions of dollars and are often the largest part of litigation costs. The bulk of e-discovery costs is made up of the cost of the time it takes attorneys to review and analyze the ESI [Roitblat et al. 2010].

Consequently, lawyers and their clients are keen on finding ways to reduce the cost of attorney review. Technology-assisted review ("TAR," also known as computer-assisted review) is a process in which computers are used to assist humans with the review of ESI to reduce review time yet ensure quality results [Grossman and Cormack 2011].

The effectiveness of TAR is measured by looking at precision and recall [The Sedona Conference 2013a]. Precision is a measure of the number of responsive documents retrieved compared to the total number of documents retrieved (how well does the system retrieve relevant documents without retrieving irrelevant documents?), and recall is a measure of the number of responsive documents retrieved compared to the total number of known responsive documents (how many of the documents known to be responsive were retrieved?) [Blair and Maron 1985; The Sedona Conference 2013a]. There is usually a trade-off between precision and recall, with precision decreasing as recall increases [Blair and Maron 1985; The Sedona Conference 2013a]. In e-discovery, recall is more important than precision, as sanctions can be imposed if a review misses too many responsive documents [Baron and Thompson 2007; Roitblat et al. 2010].

Traditionally, TAR has been based on keyword searching. Courts are accepting of this practice because it is easy to show exactly what searches were run to identify responsive or privileged ESI [Smith 2010; The Sedona Conference 2013a]. Unfortunately, keyword searches have not produced very good results, often identifying only 20% to 60% of the relevant ESI [Blair and Maron 1985; Krause 2009; The Sedona Conference 2013a; Smith 2010]. Also, keyword searches can be very over- or under-inclusive, due to intricacies of language [The Sedona Conference 2013a].

Recently a change to using predictive coding, based on multiple artificial intelligence aspects, is pointing to a real potential for significant reductions in the costs of review, with results superior to keyword-search-based methods.

REVIEW OF EXISTING APPROACHES TO PREDICTIVE CODING

Courts have begun to notice predictive coding as a possible far better alternative to traditional keyword searching. Predictive coding (also known as adaptive coding, and intelligent review, among other things) goes far beyond keyword searching and incorporates techniques such as clustering, classification, and machine learning to produce more accurate results [Barry 2013; Pace and Zakaras 2012; The Sedona Conference 2013a]. Most importantly, and distinctively, predictive coding is an iterative process that involves relevance feedback and supervised machine learning to assign a relevance score to documents based on the likelihood that the item is relevant [Pace and Zakaras 2012; The Sedona Conference 2013a; Yablon and Landsman-Roos 2013; Zhao et al. 2009].

There is no single method of predictive coding, however. Predictive coding is not one technological process but is instead a function that can incorporate many different technologies in a variety of ways [Katz 2013]. The machine learning techniques can include latent semantic

analysis, naïve Bayesian classifiers, support vector machines, logistic regression, genetic algorithms, and neural networks, among others [Katz 2013; The Sedona Conference 2013a].

Among predictive coding systems currently in commercial use, latent semantic analysis and support vector machines, both with integration of Bayesian probability, seem to be the most heavily used methods. If anyone is using genetic algorithms or neural networks, they are not disclosing their methods, so it is difficult to know if those methods are indeed being used (this is not surprising, as many vendors are reluctant to divulge serious technical information about their processes [Baron and Thompson 2007]). In predictive coding, latent semantic analysis, Bayesian classification, and support vector machines all involve assigning weights to terms, then using those weights to help determine likelihood of a document's responsiveness. Latent semantic analysis looks at frequency or user-identified relevance of terms [Sigler 2009]. The idea is that latent significance of various terms (such as code words not previously identified), and of relationships among terms, can be discovered in this manner, and that the weights can help show whether an ESI item is likely to be responsive, or related to a particular concept [Barnett et al. 2009; Sigler 2009; The Sedona Conference 2013a]. This method is particularly helpful when attorneys are not certain what they are looking for [The Sedona Conference 2013a]. Bayesian classification considers such things as frequency and proximity of terms, or where they appear in an ESI item, and assigns probabilities to documents based on how likely they are to be similar to other items with the same terms in the same areas [Sigler 2009; Tingen 2012]. Support vector machines appear to be what is being used instead of neural networks, and involve classification of items such that a line distinguishing one group of similar examples is as far as possible from that group as it is from another group of similar examples [Aurangabadkar and Potey 2014**].**

Regardless of analytical method, most systems currently in use involve supervised, rather than unsupervised, machine learning [Katz 2013].

Generally, predictive coding involves the following steps:

1. Creation of a set of "seed" documents, usually through a keyword search, but possibly also through clustering, and sometimes randomly generated.
2. Review of the seed set, often by a lead attorney for the case, who codes the documents in the set for responsiveness, privilege, issues, etc. (whatever the focus of the review is).
3. Training of the software using the coded seed set.
4. Automatic coding of some documents by the software based on their similarity to documents used in training.
5. Review of that coding, or samples of it, by the attorney, who makes changes to the coding if necessary.
6. Iterations of the software learning from relevance feedback provided by manually-coded or manually-reviewed samples.
7. Once the attorney or an algorithm decides the software is sufficiently trained (no additional samples are likely to improve effectiveness), automatic coding of the remaining documents based on their similarity to documents used in training.
8. Final output of a ranked list of documents, with the ranking based on the documents' likelihood of being responsive, being privileged, or pertaining to particular issues.
9. Most systems also include a validation process at the end that samples documents marked non-responsive, or with a low percentage of likely responsiveness, at certain intervals to see if they that determination is accurate.

[Acosta 2012; Auttonberry 2014; Byram 2012; Gallo and Kim 2013; Losey 2013; Morgan 2013; Smith 2010; Wiener 2012; Yablon and Landsman-Roos 2013]. Some programs include, in addition to relevance feedback, the system selecting new training examples based on knowledge it has already gained in training, usually selecting documents most likely to be incorrectly classified based on current knowledge [Losey 2013; The Sedona Conference 2013a]. Because no system is perfect, there are inevitably going to be some responsive documents coded as non-responsive by the program. Thus, effectiveness is often judged by the "error rate" found during the validation process, which is used as a confidence level that shows that a certain percentage of the non-responsive documents are likely actually responsive [Yablon and Landsman-Roos 2013]. For example, a confidence level of 90% means that 10% of the non-responsive documents are likely actually responsive. Figure 1 shows the general predictive coding process.

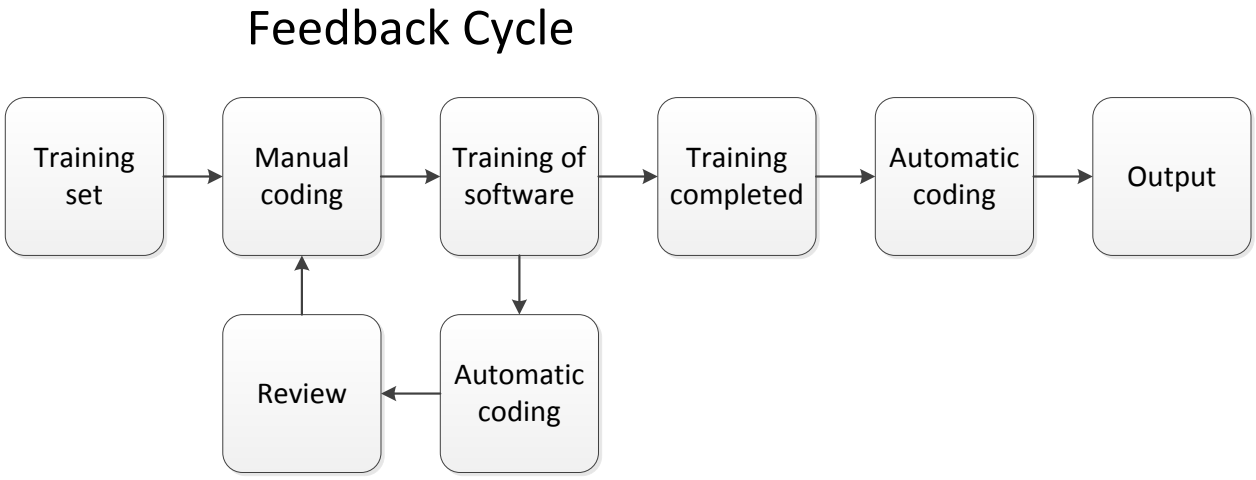

Fig. 1. The predictive coding process.

Researchers and vendors have shown the accuracy rate of predictive coding based on supervised machine learning, measured by a combination of recall and precision, is as good as or better than manual or keyword-based processes, and often the time savings of predictive coding

is a vast improvement [Byram 2012; eDiscovery Institute 2010; Gallo and Kim 2013; Grossman and Cormack 2011; Lee et al. 2011; Roitblat et al. 2010; The Sedona Conference 2013a]. Furthermore, results can be duplicated on multiple runs of the process using the same training material (assuming experts make the same coding calls for training), which is not at all the case with manual or keyword-based processes [eDiscovery Institute 2010; Lewis and Hagen 2012].

Naturally, as with most programs, the concept of garbage in, garbage out applies. If supervised learning is used, the results will not be of high quality if the people doing the training are not as well versed in the intricacies of the case as possible [The Sedona Conference 2013a]. It is unlikely that anyone would be able to make flawless or 100% consistent responsiveness calls at the beginning of a case, but a well-informed expert with a vested interest in the case, such as the lead attorney, or experts from the client company, would certainly make higher quality calls than junior or contract attorneys or the vendor's technical staff. Indeed, an investigation by Webber and Pickens [2013] confirmed a 14% decline in effectiveness of predictive coding when training was based on decisions made by people other than the "authoritative assessor, whose conception of relevance is to be used to evaluate the classifier's effectiveness."

Webber and Pickens [2013] also found that the type of training material is significant to the program's success. They found that using a greater volume of noisier documents rather than fewer cleaner examples produced higher effectiveness, particularly if there was more than one expert making responsiveness decisions for the training set.

Predictive coding, despite its effectiveness, does not wholly eliminate the need for manual review of documents by attorneys. Attorneys still need to examine the relevant documents to learn their actual content to develop their arguments for the case, and the predictive coding programs do not really provide that information [The Sedona Conference 2013b]. Rather,

predictive coding significantly reduces the number of documents that must be manually reviewed, as documents below a certain likelihood threshold are not manually reviewed at all [Auttonberry 2014]. There is much debate currently, however, about how much manual review must be done before production to opposing counsel, with some arguing that no manual review is necessary for documents meeting an agreed-upon threshold of responsiveness, and others arguing that documents falling within a particular range of responsiveness (say, 30% to 50%) should be manually reviewed to determine actual responsiveness [Auttonberry 2014; Gallo and Kim 2013; Yablon and Landsman-Roos 2013].

Current Commercial Use Of Predictive Coding

There are currently dozens of vendors offering predictive coding programs [ComplexDiscovery 2013; eDiscovery Institute 2010; Murphy 2013]. With the significant savings in time and costs predictive coding can provide over manual or more traditional TAR methods, with equal or superior results, it is not surprising that vendors are moving into this area in the hopes that those savings will attract clients. Indeed, predictive coding is said to be the future of e-discovery, much to the detriment of hordes of contract attorneys who are being replaced by the technology [Yablon and Landsman-Roos 2013].

While there is a basic pattern to the predictive coding programs currently commercially available, the actual mechanics vary significantly among vendors [Barry 2013]. Most importantly, there are differences in classification methodology. Another one of the differences among vendors is how the training set is created: some vendors create it from information provided by an expert on the case, using this information to run keyword searches, or implement other methods, to find a sample of documents to use; other vendors think that method results in bias and prefer to use a purely randomly generated training set [Pickens 2013; Yablon and

Landsman-Roos 2013]. Both methods appear to produce ultimately similar results once the entire predictive coding process is through, so it is not necessarily a handicap to not have thorough knowledge of the case or the materials before doing predictive coding [Pickens 2013]. Another significant difference among vendors relates to validation of results [Pickens 2013]. Validation is necessary for court acceptance of an e-discovery method [Pickens 2013]. Some vendors validate using a random sample of documents the program coded as nonresponsive, while others use a modified random sample of such documents [Pickens 2013].

Following is an examination of the more technical aspects of several currently popular predictive coding programs for which information is publicly available.

Catalyst

Catalyst uses a list of search terms counsel thinks are likely to produce responsive documents, and collects a sample of documents with hits for those words and without [eDiscovery Institute 2010]. Those documents are then coded by counsel, and the program then looks for additional terms in those documents [eDiscovery Institute 2010]. Terms are then given weights based on their likelihood of producing responsive documents [eDiscovery Institute 2010]. The program produces a predictive weight for each document, based on the search terms present in that document and the weights of those terms, which counsel can use to decide whether the document should be manually reviewed and to prioritize documents for manual review (e.g., review the highest-scoring documents first) [eDiscovery Institute 2010].

Catalyst uses a "statistically valid sample with 95% confidence level" of non-selected documents for validation [eDiscovery Institute 2010]. The sample is taken from documents at equal intervals along the spectrum of documents scoring below a certain level of likely responsiveness [Pickens 2013]. For example, if the cutoff for responsiveness is a 30% likelihood,

samples might be taken from documents at 25%, 20%, 15%, etc. likely responsiveness. Catalyst believes this allows for the responsiveness cutoff to be easily adjusted if the validation results show it might be necessary [Pickens 2013]. Catalyst's audit trail includes "virtually everything that is done," including all searches, changes to all records, and notes about SQL procedures [eDiscovery Institute 2010]. Catalyst claims to offer savings over linear review of 25% to 60% with an average savings of 40% [eDiscovery Institute 2010].

CategoriX (Xerox Litigation Services)

CategoriX initially develops its model from manually coded documents [Barnett et al. 2009; LaRosa 2012]. The software is trained using the model as follows: each document is represented by a word-frequency vector, which is compared to the list of words (the vector space model) generated from the documents used to create the model and the conditional probability vector assigned to each of those words based on its likelihood to be in a relevant document (probabilistic latent semantic analysis) [Barnett et al. 2009; eDiscovery Institute 2010; LaRosa 2012; Privault et al. 2010]. Results are evaluated by attorneys who give feedback [LaRosa 2012]. This process is repeated multiple times to improve accuracy [eDiscovery Institute 2010].

Barnett et al. [2009] evaluated CategoriX and found that the models generally performed best when a threshold for the probability of being responsive was set at 0.75. For 4 of the 5 models tested, the higher the threshold, the better the recall, but precision was lower. At a threshold of 0.75, average responsiveness recall of the 4 models was 90.5% and average responsiveness precision was 75.25%. The authors used a formula for overall performance of (2*P*R)/(P+R), where P represents precision and R represents recall. Performance was consistently in the low 80% range at the 0.75 threshold for the 4 models, with an average range slightly below that at a threshold of 0.5 and an even lower average when the threshold was 0.95.

For all models, performance was better than manual review, with higher recall at similar precision rates. CategoriX's audit trail includes "every aspect" of its process [eDiscovery Institute 2010]. CategoriX claims to offer savings over linear review of 30% to 77% with an average savings of 55% [eDiscovery Institute 2010].

Equivio>Relevance

This program is an active learning expert-guided system using support vectors that works as follows [Groom 2012]. A sample of documents is generated by the program (not via expert-led searches), then tagged for responsiveness by the expert [eDiscovery Institute 2010; Groom 2012]. The tagged sample is used for training, and multiple iterations of samples and training are used [eDiscovery Institute 2010]. Using a support vector machine, weights are assigned to terms in the documents and are used to predict responsiveness [Groom 2012]. When training is optimized, the program calculates a relevance score for each document [eDiscovery Institute 2010]. Equivio's audit trail includes "all actions, including the entire training process" [eDiscovery Institute 2010]. This includes the set of weighted terms [Groom 2012]. Equivio claims to offer savings over linear review of 50% to 80% with an average savings of 65% [eDiscovery Institute 2010].

Recommind

Recommind is based on probabilistic machine learning, involving support vector machines for responsiveness coding and "probabilistic topic extraction" based on probabilistic latent semantic analysis for classification into subcategories of relevance based on issues in the case [Puzicha 2009]. Recommind has experts generate a sample set through keyword searches, "category grouping, and more than 40 other automatically populated filters" to produce a high-quality sample less likely than ones generated by only one method to produce false hits or false

negatives [Recommind 2013]. The system is trained by having the system use the coded sample to suggest documents that might be relevant; the reviewers then accept or reject the determination, and further iterations of suggestions and review are done [Recommind 2013]. Recommind uses a random sample of 10,000 non-selected documents for validation and claims this results in a "95 to 99 percent confidence that relevant documents have been identified" [eDiscovery Institute 2010, Recommind 2013]. Recommind claims to offer savings over linear review of 20% to 95%, with an average savings of 40% [eDiscovery Institute 2010].

Relativity

Relativity is a supervised learning system that uses latent semantic techniques [Groom 2012]. Like Equivio, it does not use an expert-cultivated seed set and instead generates its own training set automatically. It uses latent semantic indexing to find documents with similar conceptual content [Groom 2012]. The program indexes documents, but only those above a certain conceptual threshold, then asks the trainer to give a confidence level and confidence interval [Groom 2012]. Documents are then given to reviewers to code for relevance [Groom 2012]. The system does not generate a set of weighted terms that can be exported [Groom 2012]. This could be problematic because courts prefer transparency of process, but so far this does not seem, in my experience, to have stopped firms from using Relativity (it seems to be a very popular choice in the Chicago e-discovery field).

Issues with adoption of predictive coding

Despite research and practical application showing predictive coding provides superior results to the standard keyword-based review, attorneys have not been quick to embrace this technology. There are several major reasons for this, including not understanding the technology; not believing computers can produce better results than manual review, despite significant

evidence to the contrary; being too comfortable with manual and keyword-based reviews, their results, their long-standing acceptance by the courts; and worrying that the process involves too much of a "black box" that can't be adequately explained to the courts or opposing counsel to convince them of accuracy and validity. [Barnett et al. 2009; Byram 2012; eDiscovery Institute 2010; Krause 2009; Losey 2013; Murphy 2013; The Sedona Conference 2013a; Tingen 2012].

Very recently, however, courts have begun to accept and even embrace predictive coding, which is leading to more use of predictive coding [Auttonberry 2014; Gallo and Kim 2013; Murphy 2013; Smith 2010; Yablon and Landsman-Roos 2013]. It is likely that courts will require disclosure of the training sets and training and validation processes, and perhaps information about the program's methodology, just as they now often require disclosure of the specific keyword searches used [Gallo and Kim 2013; The Sedona Conference 2013a; The Sedona Conference 2013b; Yablon and Landsman-Roos 2013,]. It appears that vendors are learning to become forthcoming with disclosure of their processes, and they are even proud of reporting their accuracy rates [Yablon and Landsman-Roos 2013]. This data is actually more prevalent and reliable than such data for keyword-based searches [Yablon and Landsman-Roos 2013]. Vendors will need to accept that they might have to sacrifice some secrecy about proprietary methods in order to gain court approval (and thus more clients) [Byram 2012].

Bagdouri et al. [2013] suggest that it is not problematic that certain parts of the process, particularly creation of the classifier, are in a "black box." They propose a model for court certification of predictive coding processes that involves focusing on the validity results of using the classifier on a predetermined, agreed-upon certification test set. Focusing on the validity of results of the classifier rather than the creation process allows any process(es) the vendor or attorneys desire to be used to create the classifier [Bagdouri et al. 2013].

Clients may drive the change as they are demand more cost-effective discovery. With recent approval from courts, and ever-increasing pressure to contain ever-increasing discovery costs, attorneys are beginning to be more accepting predictive coding [Barnett et al. 2009].

CONCLUSIONS

Thanks to recent court acceptance, and the proven cost savings and accuracy of predictive coding, both of which are superior to manual and keyword-based review, predictive coding is likely to be the immediate future of the e-discovery industry. Given the pace of technology adoption by the legal industry and the courts, it is likely to be the main technology for quite some time to come. As costs and reliability drive attorneys to make e-discovery vendor decisions, it is likely that serious research into making predictive coding even more accurate, efficient, and cost-effective will occur. Katz [2013] suggests that the most important aspect of this research will be finding the best methods for defining similarity among documents.

The machine learning aspect of predictive coding is based on the quality of information received by the program. Thus, better information will produce better results. As attorneys become more familiar with and more trusting of the predictive coding process, they are likely to understand how to give better information that will improve training.

I think predictive coding has a long way to go before it can virtually guarantee discovery of all "smoking gun" documents. Such a document might be very brief and bear more of a resemblance to documents with lower responsiveness rankings than those with higher responsiveness rankings due to the nature of the training documents. Sometimes a small phrase or one sentence makes a document the most important one in a case. Nonetheless, predictive coding seems to have much better odds of producing such a document than manual review or keyword-based searches and their proven limitations.

Also, it will be interesting to see if predictive coding can be used to make distinctions between responsiveness, relevance, and importance, thus further reducing the time attorneys must spend on reviewing documents. Responsiveness is based on whether or not a document falls under the category of documents requested, relevance is based on the probative value of the document (whether it helps or hurts the case, or doesn't make much difference either way), and importance is based on whether or not the document will be of actual use (similar to relevance, but to a higher degree) [Yablon and Landsman-Roos 2013]. Perhaps predictive coding can be applied to the results of an initial production produced with predictive coding to generate such results. Out of a collection of documents deemed responsive, a different training set could be devised to allow prediction of relevance, for example.

I am confident that the usefulness and profitability of predictive coding will lead to continuous improvements to the processes, leading to widespread adoption of this new high-quality and cost-effective means of handling e-discovery the legal system has been in need of.

REFERENCES

1. Adam M. Acosta. 2012. Predictive coding: the beginning of a new e-discovery era. *Res Gestae* (Oct. 2012), 8-14.
2. Sumedha Aurangabadkar and M.A. Potey. 2014. Support vector machine based classification system for classification of sport articles. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICICT)*. IEEE. DOI:http://dx.doi.org/10.1109/ICICICT.2014.6781268
3. L. Casey Auttonberry. 2014. Predictive coding: Taking the devil out of the details. *La. Law Rev.* 74, 613-648.

4. Mossaab Bagdouri, William Webber, David D. Lewis, and Douglas W. Oard. 2013. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (CIKM '13). ACM, New York, NY, USA, 989-998. DOI:http://doi.acm.org/10.1145/2505515.2505708
5. Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider and Robert Wickstrom. 2009. Machine learning classification for document review. In *Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop)*. DESI Press, Barcelona.
6. Jason R. Baron and Paul Thompson. 2007. The search problem posed by large heterogeneous data sets in litigation: possible future approaches to research. In *Proceedings of the 11th international conference on Artificial intelligence and law* (ICAIL '07). ACM, New York, NY, USA, 141-147. DOI: http://dx.doi.org/10.1145/1276318.1276344 http://doi.acm.org/10.1145/1276318.1276344
7. Nicholas Barry. 2013. Man versus machine review: the showdown between hordes of discovery lawyers and a computer-utilizing predictive-coding technology. *Vand. J. Ent. & Tech. L.* 15, 2, 343-373.
8. David C. Blair and M. E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM* 28, 3 (March 1985), 289-299. DOI:http://doi.acm.org/10.1145/3166.3197

9. Elle Byram. (2013). The collision of the courts and predictive coding: defining best practices and guidelines in predictive coding for electronic discovery. *Santa Clara Computer & High Technology Law Journal 29*, 4, 675-701.
10. ComplexDiscovery. 2013. Got technology-assisted review? A short list of providers and terms. Retrieved May 1, 2014 from http://www.complexdiscovery.com/info/2013/01/26/got-technology-assisted-review-a-short-list-of-providers-and-terms/
11. eDiscovery Institute. 2010. eDiscovery Institute survey on predictive coding. Retrieved Apr. 28, 2014 from http://www.discovia.com/wp-content/uploads/2012/07/2010_EDI_PredictiveCodingSurvey.pdf
12. Andrew Gallo and Sarah Kim. 2013. Practice tips: predictive coding: process and protocol. *B.B.J.* 57, 22-24.
13. Tom Groom. 2012. Three methods for ediscovery document prioritization: Comparing and contrasting keyword search with concept based and support vector based "technology assisted review-predictive coding" platforms. Retrieved Apr. 28, 2014 from http://www.d4discovery.com/2012/04/three-methods-for-ediscovery-document-prioritization/
14. Maura R. Grossman and Gordon V. Cormack. 2011. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. J.L. & Tech.* 17, 11-16.
15. Daniel Martin Katz. 2013. Quantitative legal prediction--or--how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory L.J.* 62, 909-966.
16. Jason Krause. 2009. In search of the perfect search. *ABA Journal* 95, 38-43.

17. Stuart LaRosa. 2012. Establishing a defensible approach to technology-assisted review. Retrieved Apr. 28, 2014 from http://www.metrocorpcounsel.com/articles/21729/establishing-defensible-approach-technology-assisted-review
18. Tae Rim Lee, Bon Min Goo, Hun Kim, and Sang Uk Shin. 2011. Efficient e-Discovery Process Utilizing Combination Method of Machine Learning Algorithms. In *Proceedings of the 2011 Seventh International Conference on Computational Intelligence and Security (CIS '11)*. IEEE Computer Society, Washington, DC, USA, 1109-1113. DOI:http://dx.doi.org/10.1109/CIS.2011.246
19. David D. Lewis and Regina Jytyla Hagen. 2012. Intelligent review technology: improving the practice of document review in legal discovery. Retrieved Apr. 28, 2014 from http://www.krollontrack.com/library/irtwhitepaper_krollontrack2012.pdf
20. Ralph Losey. 2013. The many types of legal search software in the CAR market today. Retrieved Apr. 27, 2014 from http://e-discoveryteam.com/2013/03/03/the-many-types-of-legal-search-software-in-the-car-market-today/
21. Henry Coke Morgan Jr. 2013. A survey of emerging issues in electronic discovery: predictive coding: a trial court judge's perspective. *Regent U.L. Rev.* 26, 71-80.
22. Jeremy Pickens. 2013. Predictive ranking: technology assisted review designed for the real world. Retrieved Apr. 28, 2014 from http://www.catalystsecure.com/predictive-ranking-technology-article.html
23. Caroline Privault, Jacki O'Neill, Victor Ciriza, and Jean-Michel Renders. 2010. A new tangible user interface for machine learning document review. *Artif. Intell. Law* 18, 459-479. DOI:http://dx.doi.org/10.1007/s10506-010-9090-z

24. Nicholas Pace and Laura Zakaras. 2012. Rand Institute for Civil Justice, where the money goes: understanding litigant expenditures for producing electronic discovery. Retrieved Apr. 28, 2014 from http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf
25. Jan Puzicha. 2009. Defensible predictive coding. In *Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop)*. DESI Press, Barcelona.
26. Recommind. 2013. Predictive coding for dummies, Recommind special edition. Retrieved Apr. 14, 2014 from http://www.dummies.com/Section/id-813917.html
27. Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. 2010. Document categorization in legal electronic discovery: computer classification vs. manual review. *J. Am. Soc. Inf. Sci. Technol.* 61, 1 (January 2010), 70-80. DOI:http://dx.doi.org/10.1002/asi.v61:1
28. Johannese S. Scholtes. 2009. Text-mining: the next step in search technology. In *Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop)*. DESI Press, Barcelona.
29. The Sedona Conference. 2013. The Sedona Conference best practices commentary on the use of search and information retrieval methods in e-discovery. Retrieved Apr. 14, 2014 from https://thesedonaconference.org/download-pub/3569
30. The Sedona Conference. 2013. The Sedona Conference commentary on achieving quality in the e-discovery process. Retrieved Apr. 14, 2014 from https://thesedonaconference.org/download-pub/3641

31. Sonya L. Sigler. 2009. Are lawyers being replaced by artificial intelligence?. In *Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop)*. DESI Press, Barcelona.
32. Jennifer M. Smith. 2010. Electronic discovery and the constitution: inaccessible justice. *J. Legal Tech. Risk Mgmt.* 6, 122-172.
33. Jacob Tingen. 2012. Technologies-that-must-not-be-named: understanding and implementing advanced search technologies in e-discovery. *Rich. J.L. & Tech.* 19, 2-48.
34. William Webber and Jeremy Pickens. 2013. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 929-932. DOI: http://doi.acm.org/10.1145/2484028.2484156
35. Gary Wiener. 2012. Technology-assisted review: what is it and why should you care? *Hous. Law.* 50, 24-28.
36. Charles Yablon and Nick Landsman-Roos. 2013. Predictive coding: emerging questions and concerns. *S.C. L. Rev*. 64, 633-679.
37. Feng C. Zhao, Douglas W. Oard, and Jason R. Baron. 2009. Improving search effectiveness in the legal e-discovery process using relevance feedback. In *Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop)*. DESI Press, Barcelona.