

Machine Learning Classification for Document Review

Thomas BARNETT,[◇] Svetlana GODJEVAC,[◇] Jean-Michel RENDERS,[♦] Caroline PRIVAULT,[♦]
John SCHNEIDER,[◇] Robert WICKSTROM[◇]

[♦] Xerox Research Center Europe
6 chemin de Maupertuis,
38240 Meylan, FRANCE
 {Firstname.Lastname}@xrce.xerox.com

[◇] Xerox Litigation Services
485 Lexington Avenue, 22nd Floor,
New York, NY, 10017
 {Firstname.Lastname}@xls.xerox.com

Abstract

Identifying potentially responsive data using keyword searches (and their variants) has become standard operating procedure in large-scale document reviews in litigation, regulatory inquiries and subpoena compliance. At the same time, there is growing skepticism within the legal community as to the adequacy of such an approach. Developments in information retrieval and extraction technologies have led to a number of more sophisticated approaches to meeting the challenges of isolating responsive material in complex litigation and regulatory matters. Initially met with resistance by judges, practicing attorneys and legal professionals, such approaches are garnering more serious analysis as the amount of potential source data expands and the costs of collection, processing and most significantly, review, strain corporate budgets. One of these new approaches, applying machine learning classification to the human decision making process in litigation document reviews is the subject of this paper. The human (or manual) review phase of the e-discovery process typically involves large teams of attorney reviewers analyzing thousands of documents per day to identify and record (or “code”) content responsive to document requests, regulatory subpoenas or related to specific issues in the case. The currently accepted approach to the review process is costly, time-consuming, and prone to error. Accurately and efficiently assigning appropriate coding (e.g., responsive/non-responsive) is essential as the volumes of data continue to increase while parties remain subject to strict deadlines imposed by courts and regulatory bodies. This paper suggests that automatic textual classification can expedite this process in a number of ways. It can pre-designate responsive documents. It can reduce the quantity of documents that require human review by identifying subsets of non-responsive documents, after which the remaining material can be organized based on likelihood of responsiveness. Guided by this ranking, review teams can prioritize manual review on selected subsets of documents most likely to be responsive. Further, machine learning textual classification can augment the ability to assess review accuracy by highlighting inconsistencies in reviewer decisions without the requirement of re-reviewing by more senior level attorneys.

1. Introduction

There is a growing consensus that automated search tools are more accurate and effective than simple keyword searching in managing the increasing amount of data subject to analysis in e-discovery (Kershaw 2005, Paul and Baron 2007). The majority of e-discovery services and software products rely on the use of keywords to identify and organize documents in preparation for human review. The legal community is familiar with keyword searching, which forms the bases of case law and statutory law research in legal databases. In the e-discovery context, the limitations of this approach are well documented and understood (Baron, J. 2009; Blair, D.C. & Maron, M.E. 1985, 1990; The Sedona Conference Working Group 2007; Tomlinson 2008).¹ Specifically, the use of keyword searches in document review provides

¹ See also, *Victor Stanley v. Creative Pipe*, 2008 WL 2221841(D. Md. May 29, 2008) (“... while it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI [electronically stored information], all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search.”)

neither a sufficient nor a defensible mechanism for retrieving an acceptably complete set of potentially responsive documents. This is in part because it leaves the burden on the legal teams to conceive, a priori, all possible search terms that might retrieve responsive material. In response to this deficiency, more advanced search technologies have been introduced. Such technologies have been referred to as “concept (or conceptual) searching,” “concept organization,” “concept grouping,” “clustering,” and “latent semantic indexing,” among others. Concept searching, a term adopted by many software and service providers is vaguely defined in the legal market. A simple definition of concept search might be “advanced search” i.e., augmented keyword searching by broadening queries to include synonyms, using a thesaurus to include terms with similar meanings, or using root expansion to include related forms of the same word. A broader definition, however, would encompass machine learning technologies for text mining, and more specifically, clustering or categorization which is the object of the practical study presented in this paper.

We consider here the document review phase of the e-discovery process primarily in the context of review for responsiveness to a document request or subpoena. Notably, an estimated 70% or more of the overall cost of discovery is attributable to human review of documents for responsiveness and privilege.² Reducing the cost of document review would thus substantially reduce the overall cost of e-discovery. To this end, we have applied CategoriX (the XRCE³ machine learning classifier) to several document collections originating from sample e-discovery cases. The main goal was to apply machine learning to a sample of tagged documents (e.g. responsive/non-responsive) extracted from a larger collection, and then automatically apply the likely determinations to the remainder of the collection. Using this type of automatic document categorization, the overall size of the manual document review can be reduced substantially.

This paper is organized as follows: Section 2 describes the machine learning classification technology used for our experiments. Section 3 depicts the testing scheme with the data that have been processed, and discusses the numerical results. Two specific manually reviewed sets of emails have been studied: the first case offers a unique opportunity for comparing multiple manual reviews of the same documents sets together with a CategoriX automated review; the second case presents a challenging and highly skewed data set, with only 1% deemed potentially responsive by human reviewers. Section 4 concludes the paper and discusses further development and broader use opportunities for the technology.

2. The CategoriX classifier

Based on research from XRCE and originally from PARC,⁴ CategoriX (sometimes referred to herein as “CX”) is a machine learning system that learns from human decisions for building probabilistic models which are then used to automatically infer the probable category(ies) of new documents. CategoriX relies on a probabilistic generative model (Hofmann 1999) which can be seen as a hierarchical extension to a probabilistic latent semantic analysis (“PLSA”) where both documents and words may belong to different categories. The basic assumption of PLSA is that there exists a set of hidden (latent or unobserved) factors which can underlie the co-occurrences among a set of observed words and documents. Given a certain category variable C there is a certain probability $P(W | C)$ to generate a word W , and a certain probability $P(d | C)$ to generate a document d . The aim of PLSA is to find these probabilities and the probability $P(C)$ of choosing a class C . Two statistical assumptions are made: the observed pairs (w, d) are assumed to be generated independently; and given a category C , words are generated independently from the documents. Given this framework, a generative model is built from the probabilities $P(C)$, $P(W | C)$ and $P(d | C)$, computed such that the log-likelihood of the observed document collection is maximized.

² *Business Wire*, Jan 25, 2008; *Digital Discovery & E-Evidence* Vol.8, No.6, pg 2

³ Xerox Research Center Europe

⁴ Palo Alto Research Center

A two-phase process: off-line learning & on-line classifying

CategoriX follows a two-phase process: first the generative model is learned from a collection of documents that have already been manually categorized in a pre-defined taxonomy. This is the off-line learning stage where CX learns a decision of a subject matter expert (SME) (e.g., an attorney familiar with the case), in order to later reproduce these decisions on additional documents. Second, the predictive model is used to infer the category assignment of new –not yet classified– documents. This is the on-line classification stage. The off-line learning stage is done once or periodically, whereas the on-line classification stage can have many cycles in order to process different batches of documents through the same classifier model.

Training a classifier model

During the learning phase, the set of SME categorized documents named the “training set” is processed as follows: each document is automatically transformed into a “bag-of-words”; it is the commonly used representation of a text document with a word-frequency vector, ignoring the position of words in the document. Next, a vocabulary is created from the collection of the different words encountered within the documents. Subsequently, the learning process computes a large matrix (aka as the “model kernel matrix”) representing the “profile” of each category in the taxonomy along the global model vocabulary: at row W_j and column C_i we find $P(W_j | C_i)$, the contribution of word j in the model vocabulary to category i in the taxonomy, expressed in terms of the conditional probability for word W_j to be observed in class C_i within the training set of documents. This matrix is referred to as the “training model.”

Running a classifier model

The training model is further used at runtime for evaluating additional documents against each category profile. A new document is first translated into its bag-of-words. Using that vector and the model kernel matrix, the runtime engine is able to compute the likeliness for the document to belong to each of the model categories. The primary output of CategoriX when a document d is sent for evaluation, is a conditional probability vector (whose length is equal to the number of model categories), where the value encountered at index i indicates the probability of category C_i given document d , namely $P(C_i | d)$, all the values summing up to 1. This results in a “soft assignment” of the document to the model categories. Using simple or more advanced heuristics rules, a “hard assignment” of the document to one or more categories can be easily derived. The easiest strategy consists in assigning the document to the category C_i that maximizes $P(C_i | d)$.

3. Experiments and results

The usual outline of a testing plan for CategoriX is as follows: a collection of coded documents is first identified. It can be a random sample extracted from a database, or a tailored set selected for its representativeness of the overall document population. This set of documents is then divided into two parts: the first part is used for training the classifier, and the second one is for evaluating the performance of the training model. For the training part, several training subsets of different sizes can be derived. The subsets can be selected randomly, which further ensures that the classifier performance is not biased toward a particular set of data. When selecting documents for the training set, it is necessary to ensure that each model category is correctly represented. For instance, if the objective of the classifier is to discriminate between two categories C_1 and C_2 , (as in responsive/non-responsive coding), and if a rough probable distribution of the documents between C_1 and C_2 is known, the composition of the training set should follow the same distribution. If the probable distribution is not known, one can rely on the distribution of the collection of currently available reviewer coded documents.

When the classifier outputs the probabilities for C_1 and C_2 for a document, the document can be assigned to the category whose probability is above 0.5. The classifier can, however, be made more conservative by imposing a higher threshold on the probability for being responsive in order to increase the precision,

but at the expense of a lower recall. Adjusting the threshold value allows fine-tuning the rate of precision and recall. To derive an overall measure of performance, we computed the F1 as $2 * P * R / (P + R)$.

We tested CategoriX on two types of document populations: (a) documents dealing with service industry subject matter (S-data) and (b) documents dealing with manufacturing subject matter (M-data). From each group, we selected a sizeable set of reviewer coded documents and trained CategoriX to recognize two categories of documents: responsive and non-responsive for the respective matters.⁵

3.1 S-data experiments

S-data came from a database of 10,000 documents, all of which had been reviewed and coded for responsiveness by five independent reviewer groups (A1, A2, A3, A4, and A5). A subset of 5,000 email only documents was selected for CategoriX modeling and testing. The training sets put aside from this population were as follows:

Training Sets

T1= 1 set of 500 responsive + 500 non-responsive emails, coded by Reviewer group A1.

T2= 1 set of 500 responsive + 500 non-responsive emails, coded by Reviewer group A2.

T3= 1 set of 500 responsive + 500 non-responsive emails, coded by Reviewer group A3.

T4= 1 set of 500 responsive + 500 non-responsive emails, coded by Reviewer group A4.

T5= 1 set of 500 responsive + 500 non-responsive emails, coded by Reviewer group A5.

In the above conditions, T1-T5, each set was used with the responsive/nonresponsive coding of the appropriate reviewer group and CategoriX was trained on decisions of that group. For example, T1 was trained on decisions of A1, T2 on decisions of A2, etc. Since the entire population of S-data was coded by all five review groups, the documents in T1-T5 also have codes of all other review groups (A1-A5). Judgments of other groups not assigned for the training were ignored in that model.

T6= Ten sets of 1K documents, randomly selected from the population of 5K coded emails.

In condition T6, each 1K training set was used with the coding of each of the five review groups. This means that CategoriX was trained on T6_1 for decisions by group A1, A2, A3, A4, and A5. So, we had the following training sets: T6_1_A1, T6_1_A2, T6_1_A3, T6_1_A4, T6_1_A5 and similarly for T6_2, T6_3 etc. Due to differences in reviewer coding of the same document by the five review groups these training sets did not provide an exact 50/50 balance between the two categories. This is shown in Tables 3 and 4.

S-Test= for each training set above, we used as a test set the remaining 4,000 emails that were not part of the training set.

Responsiveness Rates

Reviewer responsiveness rates for the entire document population across the five reviewer groups varied significantly (39.60% - 57.84%). This is shown in Table 1. Responsiveness rate for the subset of emails used for CategoriX modeling was similar, 40.88% - 56.95%, and it is given in Table 2.

⁵ While the testing described in this paper was limited to email data as well as to binary reviewer determinations (i.e., responsive/non-responsive) the CategoriX technology is not inherently limited by data type or number of reviewer classifications (e.g., issue codes or privilege determinations rather than binary decisions). Accordingly, further testing is being conducted in broadening the scope of the application of the technology.

10K - Reviewed Set			
Review Group	Responsiveness Rate	σ	
		Median	Standard Deviation
A1	39.60%		
A5	42.92%		
A4	44.71%		
A3	52.55%		
A2	57.84%	44.71%	6.69%

Table 1. Responsiveness Rate in the complete set of the S-population (10K documents)

5K - Reviewed Email Set			
Review Group	Responsiveness Rate	σ	
		Median	Standard Deviation
A1	40.88%		
A4	45.54%		
A5	48.04%		
A3	50.93%		
A2	56.95%	48.04%	5.37%

Table 2. Responsiveness Rate in the Email subset of the S-population (5K documents)

The wide range of responsiveness rates among different reviewer groups underscores a generally known but seldom challenged fact about human manual review. Reviewers often do not agree on how documents should be coded. They also, as a group, may not make decisions using the same decision factors when coding documents (Voorhees, E. and Harman, D. 1997).⁶

Out of five groups (A1-A5) of reviewer assigned categories, five balanced training sets (T1-T5) were created in the following way: for each T_i , 500 documents coded responsive and 500 documents coded non-responsive by group A_i were randomly selected from the set of 5,000 emails. Due to the inconsistency of coding across reviewer groups, the balanced training sets did not contain the same documents for each group, although there may be an overlap among the training sets. This difference among the training sets introduced document type as a variable (which we call the “document composition effect”). In other words, it could be argued that a CategoriX model based on one review group had a better fit than the model based on another review group because the documents used for one group were not a representative set of the population.

To avoid this effect, we also created the T6 training set as ten random sets of 1,000 documents. Using coding from reviewer groups A1-A5, CategoriX was trained on each review group’s coding to generate five distinct models for each of the ten sets. This resulted in 50 CategoriX models (ten random sets of 1,000 documents coded by five different review groups).

The responsiveness rate for each of the T6 training sets varied from group to group. In the review protocol, reviewers were given a four-fold option for responsiveness: (a) Responsive; (b) Non-Responsive; (c) Questionably responsive; (d) Blank/Error. Because of the four options above, the non-responsiveness rate in this condition did not result in the complement of the responsive set, as in our balanced training set scenario. The responsiveness rates for all the sets in condition T6 are given below.

Training Set	Reviewer Responsiveness Range for A1-A5	Mean Responsiveness Rate	σ
TS6_1	42.10%, 47.40%, 49.60%, 52.00%, 59.70%	50.16%	5.78%
TS6_2	40.00%, 45.80%, 48.00%, 50.10%, 56.40%	48.06%	5.36%
TS6_3	42.30%, 45.90%, 47.30%, 51.30%, 58.00%	48.96%	5.92%
TS6_4	39.90%, 44.10%, 45.40%, 48.60%, 54.00%	46.60%	5.24%
TS6_5	40.40%, 44.90%, 48.30%, 50.80%, 57.20%	48.32%	5.65%
TS6_6	42.10%, 46.00%, 49.40%, 53.60%, 58.20%	49.86%	5.64%
TS6_7	39.60%, 44.40%, 47.90%, 52.40%, 56.50%	48.16%	6.61%
TS6_8	41.40%, 45.60%, 49.50%, 51.50%, 59.80%	49.56%	6.91%
TS6_9	42.60%, 46.20%, 49.10%, 50.00%, 57.80%	49.14%	5.04%
TS6_10	38.00%, 43.70%, 47.70%, 49.10%, 54.20%	46.54%	6.04%

Table 3 -- Responsiveness Rate for training sets in condition T6.

The responsiveness rate varied as much as 18.4% among the reviewer groups in TS6_8 condition, and the non-responsiveness rate varied even more at 25.2% in TS6_5 condition. This level of disagreement among reviewer groups is common and it is indicative of the variable quality of human review (Voorhees, E. and Harman, D. 1997).

⁶ See e-Discovery Institute 2008 and references therein.

CategoriX Models

Each CategoriX model was assessed against the manual categorization of the test population for the same review group the model was based on. This means, for example, that CategoriX was trained on 1,000 emails coded by group A1, a model was created, and based on that model 4,000 emails were scored by CategoriX for responsiveness. Thus, success of the CategoriX scores was measured with respect to the coding of the 4,000 emails by the same reviewer group CategoriX was trained on, in this case group A1.

Precision and Recall values for the CategoriX models based on the balanced training sets T1-T5 at threshold settings 0.5, 0.75, and 0.95 are given below in Table 4. When both precision and recall measures are taken together (F1), four out of five models perform best at the 0.75 threshold, while the model based on group A1 performs better at the 0.5 cut-off point.

Threshold	Review Group	F1	Responsiveness Average Recall	Responsiveness Average Precision
0.5	A1	0.715294	96%	57%
0.5	A2	0.821796	94%	73%
0.5	A3	0.799515	97%	68%
0.5	A4	0.756478	97%	62%
0.5	A5	0.806228	99%	68%
0.75	A1	0.64	64%	64%
0.75	A2	0.822545	87%	78%
0.75	A3	0.812195	90%	74%
0.75	A4	0.801472	92%	71%
0.75	A5	0.848421	93%	78%
0.95	A1	0.646154	60%	70%
0.95	A2	0.751154	63%	93%
0.95	A3	0.699007	64%	77%
0.95	A4	0.755556	68%	85%
0.95	A5	0.785912	71%	88%

Table 4. S-data Average Precision and Recall for five groups for the 1K balanced training sets.

Table 4 also illustrates the shift in the direction of the slope of the precision and recall values at 0.95. At this threshold, for all review groups, precision becomes higher than recall. At all other cut-off points, recall is higher than precision. This is a typical result, since precision and recall are inversely related. This trade-off between precision and recall values can be exploited in different review scenarios. In reviews such as Hart-Scott-Rodino filings related to corporate mergers and acquisitions, CategoriX threshold settings could be adjusted for maximizing recall. In reviews aimed for high precision rate, such as Securities and Exchange Commission inquiries, the CategoriX threshold could be adjusted for high precision. The flexibility of manipulating precision and recall values through the threshold adjustment without having to re-run CategoriX is anticipated to be of great value to the document review process.

In addition to using CategoriX for the purpose of document classification, CategoriX models could also be used for quality control and evaluation of human review. We used the information retrieval measures of the CategoriX models based on the training sets in condition T6 to evaluate human review for: (a) consistency of reviewer coding and (b) quality of reviewer coding.

Measure of Reviewer Consistency

With the T6 condition experiment, we sought to evaluate the ability of each review group to consistently code the documents. If the manual coding of the data is internally consistent, then the model based on that set will result in better CategoriX performance. Conversely, a better performance of a CategoriX model can be viewed as a measure of consistency of reviewer decisions.

As outlined earlier with the T6 data descriptions, we trained CategoriX on each of the five review group's manually assigned categories using ten sets of 1,000 randomly selected documents. Each reviewer group's CategoriX model was then evaluated against its own coding of the 4,000 documents in the test set. Table 5 shows the precision and recall averages for three thresholds in this test of 50 models.

Based on the F1 score at 0.75 threshold, review group A5 performed the best (most consistent) of all five groups, followed by group A3. So, A5 was the best at predicting the categories of the test set based on its assigned categories of the training set. Group A1, on the other hand, at all six thresholds, had consistently lower scores than any other group. Based on these values, we conclude that group performance is measurable using CategoriX metrics.

Threshold	Review Group	F1	Average Recall	Average Precision	σ Recall	σ Precision
0.5	A1	0.731112532	96.23%	58.95%	0.50%	0.84%
0.5	A2	0.815882533	96.35%	70.75%	0.85%	1.11%
0.5	A3	0.804723525	97.79%	68.36%	0.71%	1.17%
0.5	A4	0.771546878	96.89%	64.10%	0.59%	1.42%
0.5	A5	0.80038708	98.16%	67.57%	0.48%	1.62%
0.75	A1	0.75655038	88.14%	66.27%	1.61%	0.91%
0.75	A2	0.828989707	90.58%	76.42%	1.32%	1.29%
0.75	A3	0.831774555	93.26%	75.06%	1.14%	0.72%
0.75	A4	0.803108368	90.30%	72.31%	1.60%	1.18%
0.75	A5	0.840269034	91.80%	77.46%	1.32%	1.20%
0.95	A1	0.674314961	62.77%	72.84%	2.49%	1.05%
0.95	A2	0.771155671	70.09%	85.71%	1.88%	4.21%
0.95	A3	0.750658893	70.47%	80.30%	2.26%	0.83%
0.95	A4	0.750313948	66.37%	86.29%	1.81%	0.64%
0.95	A5	0.771202312	70.23%	85.51%	1.84%	1.14%

Table 5. S-data Precision and Recall for five review groups based on a 1K training set identical for all groups.

Measure of Review Quality

The results of CategoriX modeling in condition T6 also show that we can use CategoriX to assess the quality (as contrasted with mere consistency) of human review. In this test, each review group is evaluated against the reference categorization of group A5 on the test set, taken as the gold standard for this set of documents. Group A5 was taken as the gold standard because this review group also functioned as a control group in the study due to its familiarity with the subject matter and extensive experience reviewing documents in similar cases. A5 assigned categories were thus used to assess the quality of reviewer groups A1-A4.

Reviewer to Reviewer Comparison

By analyzing the manually assigned categories of all 5,000 emails for each group, precision and recall against the A5 manual categorization was computed. The F1, precision and recall values of the reviewer assigned categories from groups A1-A4 with respect to the reviewer assigned categories of group A5 were as follows:

Review Group	F1	Recall	Precision
A1	0.717347169	66.39%	78.02%
A2	0.739778372	80.84%	68.19%
A3	0.775841103	79.92%	75.38%
A4	0.789281886	76.87%	81.10%

Table 6. - Reviewer to reviewer precision and recall with respect to the gold standard of A5 assigned categories.

In this comparison, given in Table 6, group A4 and A3 have the highest scores: F1 for A4 = 0.78 and F1 for A3 = 0.77. This is an evaluation of the review group manual categorization. We can, however, further compare the groups by using CategoriX assignments of the test set, instead of the manually assigned categories of the test set, and measure the accuracy of these predictions against the gold standard. This next evaluation is called “CategoriX to Reviewer comparison.”

CategoriX to Reviewer comparison

For CategoriX to reviewer comparison, we compared the CategoriX models’ precision and recall values of groups A1-A4 against the judgments of group A5. The compiled values are given in Table 7. Using the same measure, the F1 value, the review quality of groups A1-A4 can be assessed through their respective CategoriX models. The relevance of these results is twofold:

- CategoriX comparison shows that models based on groups A4 and A3 are better than A1 and A2 in comparison to the A5 baseline: A4=0.80 and A3=0.79 (at 0.75 threshold).
- A4-based and A3-based CategoriX prediction is better than A4 and A3 manual categorization: 0.80>0.78 for A4, and 0.79>0.77 for A3.

Threshold	Review Group	F1	Average Recall	Average Precision	σ Recall	σ Precision
0.5	A1	0.782437	93.52%	67.26%	0.51%	0.98%
0.5	A2	0.749574	97.17%	61.01%	1.28%	1.81%
0.5	A3	0.776655	97.78%	64.42%	1.03%	1.08%
0.5	A4	0.780812	95.06%	66.25%	0.52%	1.63%
0.75	A1	0.77898	83.09%	73.32%	1.93%	0.58%
0.75	A2	0.774737	92.04%	66.89%	2.73%	2.33%
0.75	A3	0.797498	92.44%	70.12%	0.95%	0.88%
0.75	A4	0.803397	87.79%	74.05%	1.61%	1.32%
0.95	A1	0.641566	55.63%	75.78%	2.26%	0.86%
0.95	A2	0.69499	67.48%	71.64%	5.92%	1.50%
0.95	A3	0.727391	70.23%	75.44%	1.57%	0.84%
0.95	A4	0.698005	60.38%	82.71%	1.14%	1.35%

Table 7. S-data CategoriX average precision and recall for groups A1, A2, A3, and A4 measured against A5.

CategoriX values for all groups, even A1 and A2, show that the automatic classifier is more consistent in its responsive/non-responsive determination than the human review teams. The fact that the same groups came ahead in both *Reviewer to Reviewer Comparison* and *CategoriX to Reviewer Comparison* suggests that evaluating the quality of the review can also be accomplished using CategoriX.

In addition, we also note that recall values in Table 7 show that CategoriX was able to accurately identify more responsive documents in the test set than the human reviewers, while maintaining the precision rates close to the levels of the human review. CategoriX’ higher recall values with minimal loss of precision demonstrates that overall, CategoriX retrieved more responsive documents than the manual review.

3.2 M-data experiments

The set of experiments related to the M-data investigated the influence of the training set sizes and sub-sampling strategies on categorization performance. The M-data set consisted of 170,000 manually reviewed email documents. Approximately 1% of these documents were coded as responsive. Responsiveness could then be considered here as a rare event, with a highly skewed distribution. The methodology used was to first randomly build ten pairs of training/test sets, with no intersection between a test set i and its corresponding training set T_i . Training sets T_i contained 136,000 documents (80% of the corpus), and test sets contained 34,000 documents (20% of the corpus). From these ten initial pairs of training-test sets, we built final training sets by adopting two different sub-sampling strategies. In a first group of experiments, we built nine sub-training sets T_{i_j} out of each T_i by randomly selecting documents out of T_i with different sampling rates ranging from 10% (13,600 documents) to 90% (122,400 documents) while at the same time respecting the original distribution of responsive/non-responsive documents (proportion around 1% for the responsive class). This resulted in 90 training sets, (ten sets of 13,600 documents, ten sets of 27,200 documents, etc). As a second strategy, represented by another group of experiments, we adopted a biased sub-sampling for building nine sub-training sets T'_{i_j} out of each T_i but here including all responsive documents of each initial training set T_i , and sampling non-responsive documents out of the T_i non-responsive population with different rates ranging from 10% to 90%. For purpose of comparison, we also kept the original ten training sets T_i (full size).

Once all the training sets had been constituted, all documents were pre-processed (tokenization and stopword removal), and features (words) were re-weighted according to the tf-idf⁷ weighting scheme.⁸

Figure 1 provides the categorization performance of the different sub-sampling strategies (and no sampling at all) on the test sets: performance is measured by the average F1 value over the ten test sets (average over the i index, for a given j); error bars indicate two times the standard deviation. From this, we observe that:

- When the class distribution (i.e., reviewer coding) is strongly unbalanced as in the M-data (this is often the case in litigation document reviews), the biased sub-sampling strategy gives better performance.
- If this biased sub-sampling strategy is used, there is minimal drop in performance when working with small training set sizes (typically 10% of the whole corpus); note that this fact should be related to a theoretical result shown in Zadrozny et al. (2003), claiming that rejection sub-sampling provides beneficial results when taking into account the cost of misclassification, especially for very rare classes. For very small sampling rates (in this case, training set size lower than 6% of the collection), performance decreased abruptly.
- The confidence intervals on the F1-value show that the document composition of the training sets was not consequential: results do not vary significantly when two training sets are built with different documents, especially provided that one follows the biased sampling strategy.

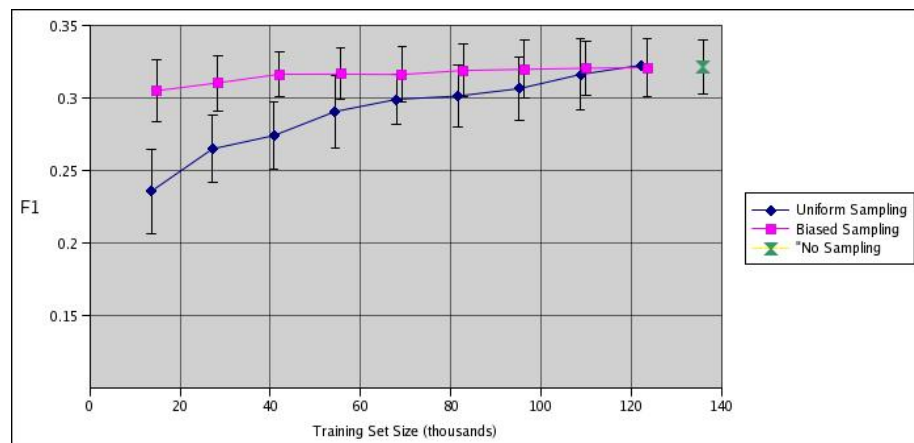


Figure 1. Effect of different sampling strategies on Categorization Performance

Note finally that in this highly skewed case, with very few responsive documents at hand for learning the classifier, CX could achieve a F1 performance around 32% while a random classifier can be shown to give the F1-value not better than 2%.

4 Conclusion

One of the foundational principles in the practice of law is reliance on and deference to precedent. Not surprisingly, change, innovation and experimentation, particularly in technology, are not always accepted with open arms. If an approach to a problem is viewed as established and effective, institutional forces are likely to ossify such an approach. In light of such strong resistance to change, it is noteworthy that a growing number of judges, practicing attorneys and legal commentators have raised doubts about the efficacy and the inherent limitations of the long established practice of relying solely on keyword searches to determine which data is worthy of review. Indeed, the discussion of potentially acceptable methods to analyze large data sets has moved from simple retrieval into the realm of sophisticated

⁷ Tf-idf (term frequency-inverse document frequency) is a statistical measure of word importance in a document collection.

⁸ See Salton et al., 1983.

information extraction technology. Such critical reflection and open debate is the result of more serious analysis of the limitations of traditional approaches (including the limits of human review) as well as the increasing financial burdens associated with the expanding universe of data. As new means of communicating electronically in a less centralized manner proliferate, the challenges faced by parties, the courts and regulatory bodies will only increase.

Applying machine learning classification technology driven by human determinations made by individuals whose knowledge is derived from active involvement in the matter represents a normative change in addressing the challenges of data intensive litigation and regulatory demands. The results of the testing of the Categorix application represent a significant step in the evolution of technology-driven approaches to managing large complex data sets in legal and regulatory actions.

References

- Baron, J. (2009). "How Do You Find Anything When You Have Billion Emails?." e-Discovery Team. <http://ralphlosey.wordpress.com/?s=pipe>
- Blair, D.C. & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Communications of the ACM*, 28(3), 289-299.
- Blair, D.C. & Maron, M.E. (1990). Full-text information retrieval: Further analysis and clarification, *Information Processing & Management*, 26(3), 437-447.
- Business Wire*, Jan 25, 2008.
- Digital Discovery & E-Evidence* Vol.8, No.6, pg 2.
- E-Discovery Institute, Inc. (2008). "Comparison of Auto-Categorization with Human Review." www.ediscoveryinstitute.org
- Hofmann, T. (1999). "Probabilistic Latent Semantic Analysis." In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289-296, Morgan Kaufmann.
- Gaussier, E., C. Goutte, K. Popat, F. Chen (2002). "A hierarchical model for clustering and categorising documents." In *Advances in Information Retrieval -- Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02)*, March 25-27, 2002. Lecture Notes in Computer Science 2291, pp. 229-247, Springer.
- Kershaw, A. (2005). "Automated Document Review Proves Its Reliability." *Digital Discovery & e-Evidence* Vol.5, No.11. November 2005.
- Paul, G. and J. Baron (2007). "Information Inflation: Can the Legal System Adapt?" 13 *Rich.J.L. & Tech.* 10, <http://law.richmond.edu/jolt/v13i3/article10.pdf>
- Salton, G. and M. J. McGill (1983). "Introduction to modern information retrieval." McGraw-Hill. ISBN 0070544840.
- The Sedona Conference Working Group (2007). "Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery" In *The Sedona Conference Journal*, Volume 8 (August 2007), pp. 189—223.
- Victor Stanley, Inc. v. Creative Pipe, Inc., et al.*, 2008 WL 2221841 (D. Md. 2008).
- Tomlinson, S. (2008). "Experiments with the Negotiated Boolean Queries of the TREC 2007 Legal Discovery Track." <http://trec.nist.gov/tracks.html>
- Voorhees, E. M. and Harman, D. (1997). Overview of the Fifth Text Retrieval Conference (TREC 5). <http://trec.nist.gov/tracks.html>
- B. Zadrozny, J. Langford & N. Abe. (2003). "Cost-sensitive learning by cost-proportionate example weighting." IEEE International Conference on Data Mining (ICDM 2003. Third).