



Comparison of Auto-Categorization with Human Review

In litigation, each side must produce to the other the set of potentially responsive documents. The sides have an obligation to produce documents that they know to be relevant. As part of the process, they also often exchange discovery requests, which describe what they believe should be produced. For example, a large company may be "requested" to "produce any and all materials, including but not limited to written documents, e-mail communications, computer databases or any other media, that address, describe, reference, or in any way relate to [Subject X.]" Other discovery requests may list tens of specifications of the form: "All documents constituting or reflecting discussions about unfair or discriminatory allocations of [Brand X] products or the fear of such unfair or discriminatory allocations."

Even after limiting the range of documents that must be considered by custodian or date range, the remaining collection often amounts to many gigabytes or even terabytes of files, including emails and their attachments, instant message texts, Wiki pages, and other electronic documents. This volume creates a tremendous burden for the producing party and litigants are anxious to identify methods that can be used to reduce these burdens.

Corporations want faster, cheaper and more accurate processes and systems to use in large-volume litigation document review. They want systems that will continue to allow them to identify responsive documents and analyze those documents that have been identified as responsive. They need systems that will help them to identify and prevent the inadvertent release of privileged documents.

Systems to support electronic discovery include human reviewers, various information retrieval algorithms, artificial intelligence tools and others. The goal of any of these systems is to effectively determine which documents in a collection are responsive and which are privileged. In the next sections we discuss how to measure the effectiveness of review systems. In the final section of this paper we discuss a research plan to apply these measures to various review systems.

The effectiveness of review systems

All discovery process relies on the same foundation. No matter what system is used for discovery, in order for a document to be produced it must:

- Have been preserved
- Be accessible
- Be retrieved
- Be presented to reviewers for assessment for potential production
- Be judged to be responsive and not privileged



If any of these steps fail, the document will not be produced. All five of these conditions must succeed. It is important to keep these steps in mind, because every discovery method, even those that involve no technology at all, affects one or more of them.

The traditional approach to discovery from the days when paper was the dominant medium for documents was to pour through boxes of papers reading each one in turn to determine whether or not it was responsive and whether it was privileged.

Although some attorneys still prefer to review documents printed on paper, the major innovation in this approach has been the shift to reviewing document images on a computer screen, rather than paper documents from a box.

These emerging trends have substantially improved the speed of electronic discovery review, but they have not addressed the reliability of the process. Scientists recognize that there are two aspects to any measurement. Reliability is the degree to which the measurement is stable and consistent. For example, if the same reliable method is applied more than once under the same conditions, it should yield the same result.

Validity is the degree to which the measurement is true and correct. At one point, people tried to measure intelligence by the size of people's skulls. These measure could be fairly reliable, measuring a skull more than once or by more than one person would yield about the same size, but they were not valid. Skull size did not accurately measure intelligence.

In discovery, we rarely measure the reliability of our reviews, but the available evidence suggests that discovery reviews may not be particularly reliable. For example, studies have found that people disagree frequently about how to categorize documents. In fact, the rate of disagreement can be higher than the rate of agreement. This disagreement means that a document may be categorized as responsive or not depending on who happens to review it. If two reviewers disagree about the relevance of a document, either one (or both) of them is wrong, or there is a lack of systematicity in how they categorize the documents. You cannot count on achieving high accuracy in a review when there is widespread disagreement about which documents are relevant.

For example, in one study from Bell Communications Research, different people assigned the same name to short documents only about 13% of the time. Tonta (1991) examined how books were assigned library catalog subject headings at the Library of Congress and the British Museum. Of 82 titles that were independently cataloged by the two libraries, only 16 percent were assigned the exact same subject headings by the two institutions and 36 percent were assigned partially matching subjects (allowing for synonyms). Even well trained librarians tended to disagree about how to assign books to well-learned subject categories.

Ellen Voorhees at NIST also studied the consistency of experts matching documents with categories as part of the TREC conference. Three assessors, one of whom created the query that was being judged, evaluated a set of documents for their relevance to the query. She found that three well-trained experienced assessors agreed on the relevance of documents to these queries only about 30% of the time. Whether a document would be



assigned to one or another category depended strongly on the person making the assignment.

As we see, there is already a level of unreliability in the traditional review process. Users, attorneys, parties, and the court system, would like a process that would improve both the reliability of electronic review and its validity. Computer systems based on information retrieval technology offer promise as a potential means to increase reliability and validity and, at the same time, decrease the burden of conducting an electronic review.

Several information retrieval technologies are available that have the potential to improve the quality of electronic discovery and reduce the cost. The goal of using any technology should be to increase the efficiency and reliability of human judgment, not to replace it. As physicians use laboratory tests and other diagnostic tools to help them treat their patients, it is becoming increasingly important to make use of information retrieval tools to help attorneys meet their electronic discovery responsibilities. Modern physicians could spend more time examining their patients with their hands and eyes, but it is doubtful that this additional time would improve the diagnostic accuracy or the level of care that is available using the full panoply of tests and other automatic diagnostic tools.

This is not the place to review the variety of information retrieval technologies that could be employed. It is sufficient to note that each of these systems is guaranteed to improve the reliability of the review process. Give the same documents to a computer system and almost all of them will give exactly the same results every time. It remains however to discover whether these systems can provide valid results.

The two most commonly used measures of information retrieval effectiveness are precision and recall. Precision is the proportion of the retrieved documents that are responsive. Recall is the proportion of responsive documents that have been retrieved. Imagine a collection that contains 1,000 documents, 300 of which are responsive. One information retrieval system might retrieve 250 of the responsive documents (Recall = $250/300 = 0.83$). It may also retrieve an additional 100 documents that are nonresponsive (Precision = $250/350 = 0.71$). Another system might retrieve 275 of the responsive documents (Recall = $275/300 = 0.92$) along with an additional 75 nonresponsive documents (Precision = $270/350 = 0.79$). Because both precision and recall are higher for the second system, than for the first, we can conclude that the second system is more accurate.

In practice there is usually a trade off between precision and recall. One can often adjust a system to retrieve more documents, thereby increasing recall, but at the expense of retrieving more irrelevant documents, and thus decreasing precision. A query for “gold or silver,” for example, will usually return more documents about metals than a query just for “gold,” but may also retrieve documents about “gold medals” and “gold standards” as well. Metaphorically, one can cast either a narrow net and retrieve fewer relevant documents along with fewer irrelevant documents, or cast a broader net and retrieve more relevant documents, but at the expense of retrieving more irrelevant documents.

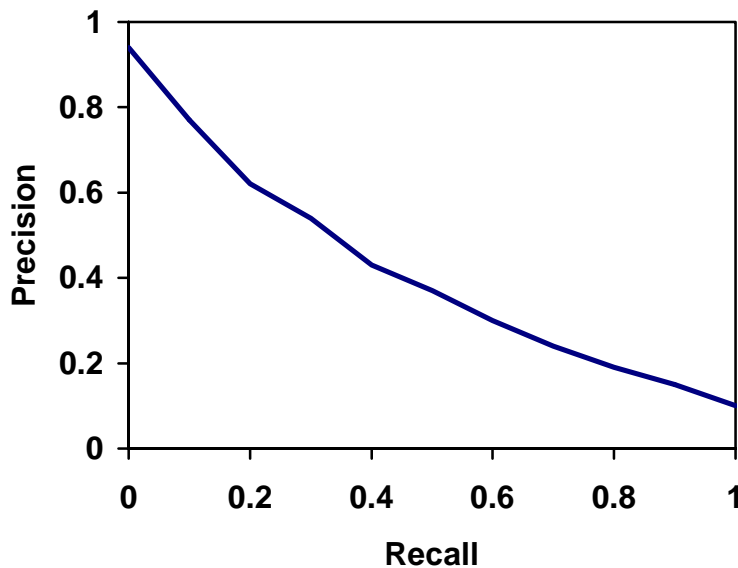


Figure 1. A precision / recall curve showing that as recall increases, precision decreases. Generally, in order to get more relevant documents you have to accept also capturing more irrelevant documents.

An example of this trade off is shown in **Error! Reference source not found.** As more documents are retrieved, recall increases, but precision decreases. At the point at which 40% of the relevant documents had been examined, only about 43% of the examined documents would have been found to be relevant. This kind of tradeoff can be observed when you adjust the system to be more or less choosy. It is also the kind of pattern you would expect if the system returned a ranked list of documents, ranked by the probability, for example, that the document would be considered relevant, and reviewers examined each document in order. Generally, the more documents you retrieve, the higher recall will be and the lower precision will be. Systems with more power are those that yield higher levels of recall for a given level of precision or higher levels of precision for a given level of recall.

In electronic discovery, we generally prefer systems with higher recall, even if it means that we have to tolerate a moderately lower level of precision. Still we prefer systems that can deliver both more precision and recall if available.

Research Plan

The goal of this research is to compare the ability of computational systems with that of human reviewers in identifying documents that are responsive and/or privileged in a collection. This comparison will rely primarily on precision and recall measures, though other measures will be brought into consideration when necessary.



The data are being provided by a Fortune 20 company (Company). This collection or corpus consists of the 1.5 million documents that were considered for production in response to a DOJ HSR 2nd request during the review of the acquisition of a Fortune 100 competitor. The company has graciously provided the data, the results of the production, and statistics on time, cost and quality of the review, which we can use as a comparison basis.

Three processing vendors have volunteered to process the data in this study using their usual methods. They will be provided with the same training material that were originally given to the Company reviewers, but will not have access to the Company judgments. Each vendor will be using their system to make independent judgments about the responsiveness and privilege status of each document.

Each vendor will process the data and produce a table that contains for each document a report as to whether the document is judged by their system to be responsive or nonresponsive, and whether privileged or nonprivileged. Additional information concerning the classification or confidence of that judgment may optionally also be provided. Vendors will not be penalized if their system does not automatically produce confidence scores, but these scores, if available, can provide useful information for assessing the overall performance of the systems.

Document ID	Responsive	Responsive Confidence (optional)	Privilege	Privilege Confidence (optional)	Additional category information (optional)
X1001d	1	0.9	0	0.3	AX, MD
X1002b	0	0.2	0	0.0	

The goal of this project is not to compare automated systems against one another, but to compare their performance with that achieved by more traditional methods. Because the final results of a system depend both on the computational capabilities and on the way in which the system is used, systems may differ from one another for a number of reasons and may require different amounts of professional effort to allow them to work effectively.

The automated systems will also be compared against a re-review of a subset of documents by human reviewers. Ideally, measuring the accuracy of any process, requires a definitive “gold standard” or “ground truth” classification of the documents. The original classification done by Company reviewers cannot be taken as absolutely definitive. Reviewers may have different opinions about which documents are responsive or privileged, their attention varies over the course of their work day, they learn about the issues and the content of the documents during the course of their review, but they generally cannot go back and re-review documents that they may have misclassified. The same person may categorize a document differently at different times.



Because of the variability in human categorization judgments, the automated systems may or may not agree in all cases with the original judgments entered by the Company reviewers. For this reason, it is essential to get an estimate of just how reliable human reviewers are. In addition, there are statistical procedures, such as Latent Class Analysis, that have been used to estimate classification accuracy when there is no “gold standard” by which to measure it. These statistical tools have been used most widely in studies of medical diagnostics (nosology), but the problems are very similar to those found in information retrieval. We will be classifying documents, based on test results (judgments of people or computer systems), as to whether they are responsive or not. Physicians classify patients, based on test results, as to whether they have a specific illness or not.

Table 1. Contingency relations in using two systems to classify documents.

	System 2	
	Judged Responsive	Judged Nonresponsive
Judged Responsive	A	B
Judged Nonresponsive	C	D

When two systems independently categorize documents there are four logical ways that each document could be categorized. Cell A contains documents that are categorized as responsive by both systems. Cell B contains documents that are categorized as Responsive by System 1, but not by System 2. Cell C contains documents that are judged Nonresponsive by System 1, but Responsive by System 2. Finally, cell D contains documents that are judged nonresponsive by both systems. Cells B and C contain documents that are categorized differently by the two systems.

One measure is to simply assess the agreement between the two systems. The proportion of agreement, $(A + D)/(A + B + C + D)$, could be used, but it does not take into account the probability that the two systems agreed by chance. For example, if both systems judged everything to be responsive, then they would agree perfectly, but this agreement would convey no information responsiveness of the documents. It is more common to use the kappa statistic (κ), which adjusts the agreement measure for chance.

Discrepant Analysis is a statistical technique used to evaluate the discrepancy between two systems of judges. In this project, these documents will be reviewed by a team of (say 3) reviewers. They will be charged with coming to a consensus on the documents in these two cells. This consensus may then be used as an estimate of the true value of each document and we can build a contingency table like that shown in Table 2.



Table 2. Contingency table for classifying documents when the true status of each document is known.

	Truly Responsive	Truly Nonresponsive
Judged Responsive	A	B
Judged Nonresponsive	C	D

Precision is then A divided by A + B and recall is A divided by A + C.

A discrepant analysis tends to over-estimate the precision and recall of the tested system because it considers the documents that are judged to be responsive by both systems all to be truly responsive. In the discrepant analysis, there is no opportunity to identify items that were missed by both tests or to those that are false positives on both.

More critically, imagine using a silly test, such as a coin toss, to evaluate the discrepant results. By the coin test, half of the discrepant results would become “true” positives for the new test and half would become “true” negatives. The apparent precision and recall of the new test would then be increased, even though the discrepancy-resolving test has no validity whatsoever. Any test used to evaluate the discrepant results can only increase or leave alone the apparent precision and recall of the new test, it cannot decrease it.

In a discrepant analysis, the standard by which accuracy is measured depends intrinsically on the results of the new test. It would be preferable to have the new test compared to some independent standard.

If we knew the true status of each document, then we could compare the performance of any system or process with this gold standard. Instead, we must deal with systems that are less than 100% accurate. These systems can mistakenly judge documents to be responsive when they are not. If we knew the true status of each document we would place those misclassified documents into Cell B in Table 2, but we don’t know what proportion of the responsive judgments actually belong in Cell B.

Table 3. Contingency table for classifying documents when the true status of each document is unknown.

	Truly Responsive	Truly Nonresponsive	Total
Judged Responsive	Y_1	$a - Y_1$	a
Judged Nonresponsive	Y_2	$b - Y_2$	b
Total	$Y_1 + Y_2$	$N - (Y_1 + Y_2)$	N

Without an infallible gold standard, we know how many documents were judged responsive or nonresponsive (a and b respectively in Table 3), but we do not know what



numbers truly were responsive or nonresponsive (Y_1 and Y_2). We need to estimate the proportion of responsive documents in order to calculate precision and recall or any other statistics.

Y_1 and Y_2 cannot be observed directly, so statisticians call them latent variables in contrast to a and b , which are observed and which are called manifest variables. We need to estimate the unobserved latent variables on the basis of observed or manifest variables. Latent Class Analysis (LCA) is a statistical procedure that allows one to estimate these latent variables.

Latent Class Analysis is widely used in biomedical studies to compare the accuracy of diagnostic tests when the true incidence of the disease are not known and there is no "gold standard" measure that can be applied. The same techniques have been used in other areas as well (e.g., Baker, 1962; Blick & Hagen, 2002).

LCA postulates the existence of an unobserved categorical variable that divides the population into distinct classes (e.g., responsive and nonresponsive). This technique has been applied to problems related to diagnostic testing, such as to estimate the true prevalence of individuals with vs. without a specific disease, based on observations of certain tests or symptoms, none of which is known to provide an absolutely accurate gold standard. LCA could then be applied in an attempt to divide the population into "true" positives and negatives. be no other form of dependence between variables entered into an LCA model.

Applying LCA requires at least three different and independent diagnostic tests. In the present study these correspond to the four systems or processes used to categorize the data (e.g., three computer systems, human judgments). LCA assumes that the association between the results of diagnostic tests arise solely from the underlying class status of the individual documents.

In order to perform the Latent Class Analysis, we will need a random sample of documents. The best way to get this sample under these circumstances is through a stratified random sample. In a random sample, each document has an equal chance of being selected. So in order to be sure that we count each document once, we need to identify the proportion of documents uniquely contributed to the total by each system. It turns out that this breakdown is exactly what is needed to compute the LCA.

Each system will uniquely identify some documents as responsive that the other systems do not. In addition, each system will identify some documents that other systems also identify as responsive. In order to calculate the sample size for each, we need to know the unique contributions of each system and each combination of systems. One illustrative hypothetical example of such a distribution is shown in Table 4. System A in this example, is hypothesized to identify 14,600 unique documents that are not identified by any other system, and 17,057 documents that are jointly identified by System A and System B (but none of the other systems). Just under 1% of the documents were uniquely identified by System A (14,600 / 1,600,047) so just under 1% of our random sample (0.9125%) should be drawn for these documents.



In addition to providing the data for this LCA, the sampled documents can also be reviewed by a panel of three or more expert reviewers. These reviewers will help to ensure that the systems are not only reliable, but also valid. The reviewers' judgments will help to ensure that the documents that are identified as responsive actually are responsive. They will also provide two other interesting pieces of information. First, because three or more reviewers are providing judgments, we can use their responses to separately estimate the underlying frequency of responsive documents, providing convergent validation of the results of the system analysis. Second, the reviewers' independent judgments can also provide an estimate of how reliable human reviewers are as they agree or disagree with each other and with the original Company review. After making independent judgments about the responsiveness of documents, they will be asked to work cooperatively to form a consensus judgment about the responsiveness of the documents on which the three or more reviewers did not agree.

The total number of documents to be reviewed is determined by the available resources. The more documents that are reviewed (when picked according to this sampling method), the more accurate will be the estimates to be derived from the analysis. A total sample size of 2500 documents would select about 300 documents from each of the systems, which would be a good sample size for this kind of analysis.

This analysis will allow us to identify with substantial accuracy the precision and recall of each system. It will also enable the calculation of a number of other statistics for comparing the performance of information retrieval systems.

Table 4. Hypothetical results of each review system and corresponding stratified random sample sizes.

Uniquely identified by system(s)	Number of documents to review					
	Number of documents	Proportion	1000	2500	5000	10000
A	14600	0.009125	10	23	46	92
B	13500	0.008437	9	22	43	85
C	16500	0.010312	11	26	52	104
D	15000	0.009375	10	24	47	94
A&B	17057	0.01066	11	27	54	107
A&C	19600	0.01225	13	31	62	123
A&D	10843	0.006777	7	17	34	68
B&C	12050	0.007531	8	19	38	76
B&D	11700	0.007312	8	19	37	74
C&D	12000	0.0075	8	19	38	75
A&B&C	10200	0.006375	7	16	32	64
A&B&D	7330	0.004581	5	12	23	46
A&C&D	5570	0.003481	4	9	18	35
B&C&D	12497	0.00781	8	20	40	79
A&B&C&D	101500	0.063436	64	159	318	635



Uniquely identified by system(s) Documents not identified as responsive by any system	Number of documents to review					
	Number of documents	Proportion	1000	2500	5000	10000
	1320100	0.825038	826	2063	4126	8251

Finally, part of deciding whether a practice is reasonable is the tradeoff of value received for the cost of engaging in the practice. We will ask each of the participating vendors to provide costs estimates for the work they have performed in this task. We will be able to compare the costs of the various processes with the actual costs of conducting the initial review.

References

Baker, F. B. 1962. Information Retrieval Based upon Latent Class Analysis. *J. ACM* 9, 4 (Oct. 1962), 512-521.

Blick, D. J. & Hagen, P. T. (2002). The use of agreement measures and latent class models to assess the reliability of classifying thermally marked otoliths. *Fish. Bull.* 100: 1-10 (2002).

Joseph, L, Gyorkos, T.W., Coupal, L. (1995). Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard. *Am J. Epidemiol*; 141: 263-72.

Yasar Tonta (1991). A study of indexing consistency between Library of Congress and British Library catalogers, *Library Resources & Technical Services* 35(2): 177-185.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, August, 1998, pp. 315-323.

Zhang, N. L. 2004. Hierarchical Latent Class Models for Cluster Analysis. *J. Mach. Learn. Res.* 5 (Dec. 2004), 697-723.