



Information retrieval and eDiscovery

Herbert L. Roitblat, Ph.D.

OrcaTec LLC

In a litigation, each side must produce to the other the set of documents that may be relevant to the case. 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 often describe in relatively vague terms, 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. Information-retrieval-like processes are used to then cull these sets down to those that must be produced to the other side.

The attorneys doing the review, must sift through thousands or millions of documents to find the ones that are relevant or privileged (privileged documents, such as certain conversations with an attorney, need to be cataloged, but not turned over). In addition, they may be searching for evidence concerning specific issues. And they may have to be creative in developing queries that can be used to identify documents that might be relevant and distinguish them from those that might not be. Some cases require months of work to develop these queries.

Depending on how carefully the pool of potentially responsive documents was limited by custodian and date, the results of all this review can be that 10 – 30% of the original documents end up produced to the other side.

The characteristics of eDiscovery are different from those of more standard information retrieval tasks.

- There is an obligation to find all responsive documents.
- The percentage of responsive documents may be higher than is typical in IR studies.
- The criteria for determining whether documents are responsive may be subjective or at least vague.
- The people responsible for conducting the process often know very little about information retrieval.

- Little attention has been historically paid to measuring the information retrieval effectiveness of eDiscovery but that situation has begun to change. The most useful measures of this effectiveness have yet to be agreed.
- The consequences of not turning over responsive documents in a timely manner can be substantial penalties running from embarrassment for minor mistakes to the hundreds of millions of dollars for bad-faith failures.

Although the duties inherent in discovery are often couched as absolutes, in fact, there is explicit recognition of the reasonableness of efforts. The recent large sanctions for discovery failure were attributed to unreasonable performance of the party's discovery obligations, not mere inadvertent failure to find a document.

The standard legal approach to discovery has traditionally been to engage lawyers and paralegals to read every document. Information retrieval effectiveness, *per se*, was not traditionally an issue with which lawyers were concerned, because every document was retrieved. As the size of the collections has risen in recent years, however, this approach has been widely recognized as overly burdensome and some lawyers are searching for reasonable alternatives. Information retrieval methods are becoming increasingly important in electronic discovery.

There is not yet any consensus about how to assess information retrieval effectiveness, largely because it has not been considered in any great depth by the legal community. One goal of the present paper is to consider some alternatives that may be informative for this community. They need to know not just how well a particular computer system performed, or how one computer system compared to another, they need an assessment of the reasonableness of their own efforts, which will extend to the quality of the queries they generated as well.

Information retrieval in e-discovery

Blair and Maron (1985) found that attorneys were only about 20% effective at thinking up all of the different ways that the document authors could refer to issues in their case. The case involved a San Francisco Bay Area Rapid Transit accident in which a computerized BART train failed to stop at the end of the line. There were about 350,000 pages in about 40,000 documents for the case (Blair and Maron, *Communications of the ACM*, 28, 1985, 289-299). The attorneys worked with experienced paralegal search specialists to find all of the documents that were relevant to the pertinent issues. The attorneys estimated that they had found more than 75% of the relevant documents, but more detailed analysis found that the number was actually only about 20%. The authors of this study found that the different parties in the case used different words, depending on their role. The parties on the BART side of the case referred to “the unfortunate incident,” but parties on the victim’s side called it an “accident” or a “disaster.” Other documents referred to the “event,” “incident,” “situation,” “problem,” or “difficulty.” Proper names were often not mentioned. The limitation in this study was not the ability of the computer to find documents that met the attorneys’ search criteria, but the inability of the attorneys and paralegals to anticipate all of the possible ways that people could refer to the issues in the case.

Concerning one issue, the attorneys in the case identified three terms that they thought would be adequate to retrieve relevant documents, Blair and Maron found 26 more. The original three words could not by themselves be used effectively to find relevant documents, because they retrieved too many irrelevant documents. Other search terms were needed to limit the range of documents that were returned, but this limitation came at the cost of missing documents that did not happen to have these

additional terms. Coming up with the right combination of terms to yield relevant results and no irrelevant results is nearly impossible.

They found that the terms used to discuss one of the potentially faulty parts varied greatly depending on where in the country the document was written. Some people called it an “air truck,” a “trap correction,” “wire warp,” or “Roman circle method.” After 40 hours of following a “trail of linguistic creativity” and finding many more examples, Blair and Maron gave up trying to identify all of the different ways in which the document authors had identified this particular item. They did not run out of alternatives, they only ran out of time.

Information retrieval measures

Standard information retrieval measures are precision and recall. Precision is the proportion of retrieved documents that are relevant to the query or topic.

$$Precision = \frac{n_{\text{Re sponse} _ \text{Retrieved}}}{n_{\text{Retrieved}}}$$

$$Recall = \frac{n_{\text{Re sponse} _ \text{Retrieved}}}{n_{\text{Re sponse}}}$$

If a collection of documents contains, for example, 1000 documents, 100 of which are relevant to a particular topic and 900 of which are not, then a system that returned only these 100 documents in response to a query would have a precision of 1.0, and recall of 1.0. If the system returned all 100 of these documents, but also returned 50 of the irrelevant documents, then it would have a precision $100/150 = .667$ and still have a recall of $100/100 = 1.0$. If it returned only 90 of the relevant documents along with 50 irrelevant documents, then it would have a precision of $90/140 = 0.64$ and a recall of $90/100 = 0.9$. 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.

An example of this trade off is shown in Figure 1. 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. Systems differ in their levels of precision and recall and in the tradeoff between them.

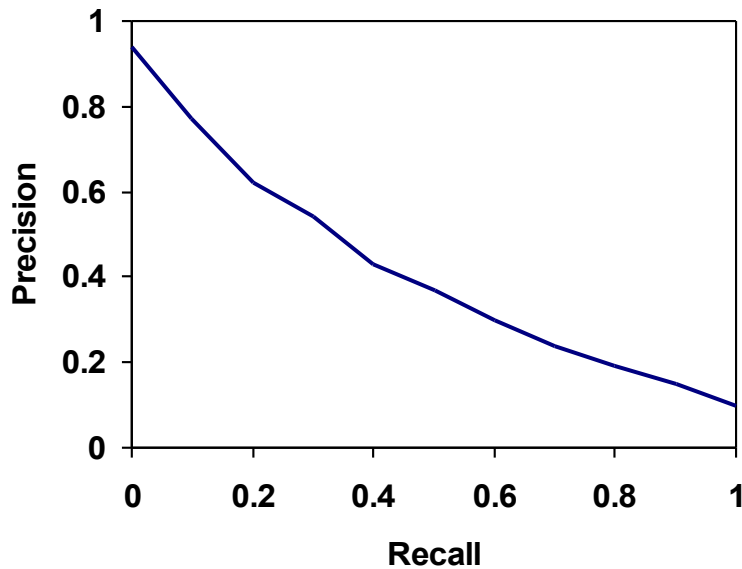


Figure 1. A precision / recall curve showing that as recall increases, precision decreases. Generally, in order to get more relevant documents you have to accept also capturing more irrelevant documents.

Alternatives to precision and recall

Precision and recall are set measures. They assume that the system either returns a document (in the set) or it does not (out of the set). They also assume that a document is either relevant or it is not relevant. Many modern information retrieval systems, in contrast, rank documents by degree of relevance. Some documents are determined by the rules of the system to be more relevant to the query than others, and these are typically returned with higher ranks than the less relevant documents.

Figure 2 shows a useful way to think about the effectiveness of information retrieval tools that rank documents by degree of relevance or by probability of relevance. It measures both the ability to find responsive documents and to rank them in a useful way. Precision and recall only measure the ability to find responsive documents.

The curve, called the retrieval operating characteristic (ROC) curve, represents the cumulative proportion of relevant vs. irrelevant documents retrieved at each rank in the list. An ideal system would rank all of the responsive documents before any of the nonresponsive documents. This would appear in the figure as a vertical line that goes straight up the left axis and then bends to go across the top of the graph, passing through the upper left-hand corner of the graph.. Except on trivial problems, no real system can achieve this level of performance. Real systems tend to mix responsive and nonresponsive documents, but the better system is the one that is more likely to present the relevant ones before the irrelevant ones.

Precision and recall represent the accuracy of the system by two numbers. It is difficult to tell, therefore, whether a system with slightly lower recall but higher precision is more accurate than one with slightly higher recall and lower precision. There are a number of conventional ways to combine these two measures, but none of these is entirely satisfactory. All of them seem to be derived more from convention than from a deep theoretical perspective. To be sure, they are important in comparing one system to another, but they are less useful in assessing the absolute effectiveness of a system. They

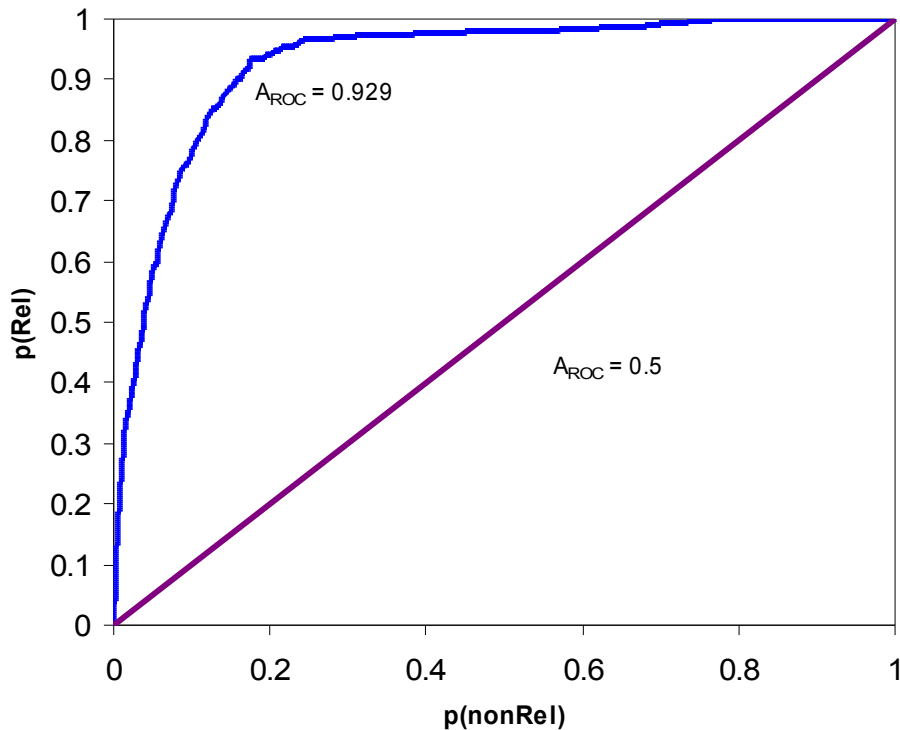


Figure 2. An example retrieval operating characteristic (ROC). An ROC curve shows the effectiveness of a ranked retrieval system. The curve represents the cumulative proportion of responsive and nonresponsive documents retrieved at each rank in the list. The horizontal axis shows the proportion of nonresponsive documents retrieved and the vertical axis shows the proportion of responsive documents retrieved at each rank. An ideal system would rank all of the responsive documents above all of the irresponsible ones. The ROC for this ideal system would be straight up the vertical axis to the upper left hand corner and then straight across to the to the upper right hand corner.

also do not typically take into account the goals of the person using the system. For example, the harmonic mean of precision and recall is one way of summarizing the performance of a system with fixed parameters, but it is difficult to interpret. Similarly, calculating precision at a fixed percentage of recall may be useful if your goal is to examine the top documents until you can answer a specific question. Average precision at 10%-increments in recall (i.e., 0%, 10%, 20% ...90%, 100% recall) may be useful if you are interested in the overall accuracy of the system, but it is difficult to interpret. Is 40% average precision good or poor performance? These measures also depend to varying degrees on the percentage of documents in a collection that are relevant and on the stringency of the person or system making the decision.

Similar questions have arisen in medical diagnosis. There the task is to distinguish between those individuals who are ill with a particular disease and those who are not. Medical tests have been measured in terms of sensitivity (the ability to find ill people) and specificity (the ability to distinguish well from ill). Medical sensitivity corresponds to recall in information retrieval and medical specificity corresponds to precision. One could make the test lax so that more people are recognized as disease holders, or more stringent, so that fewer people are designated disease holders. For example, is there

an epidemic of obesity (depends on how fat a person has to be to be called obese)? Is there an epidemic of autism (depends on the tests used to diagnose autism).

Although sensitivity and specificity measures continue to receive widespread use, some medical investigators use measures based on signal detection theory (see John A. Swets and Ronald M. Pickett, *Evaluation of diagnostic systems: methods from signal detection theory*, Academic Press, New York, 1982). Similar measures have also received attention in information retrieval studies.

As an alternative to using two numbers—precision and recall—to summarize retrieval performance, you can summarize the accuracy of a system by calculating the area under its retrieval operating characteristic curve (ROC). The ROC incorporates the idea that items vary in the degree to which they provide evidence for a document being relevant. Some documents are either more relevant than others, or are simply more likely to be relevant than others.

ROC analysis also recognizes that the performance of a decision system, such as whether or to what degree a document is relevant, depends on two questions. To what degree are the two classes of documents separable from one another and how selective do I want to be in accepting documents as being relevant? If the two classes were identical, then, of course, no system would be able to tell them apart. The better system is the one that is better at distinguishing responsive from nonresponsive documents. On the other hand, you could get higher or lower levels of precision or recall, by changing the stringency of your decision. You could require more evidence for a document to be called relevant or you could be more accepting and call a document relevant based on very little evidence. How strict or lax your criterion is depends on the task you are trying to perform and the costs of various kinds of errors, but it does not change the overall power of the system to distinguish between relevant and irrelevant documents.

In an ROC analysis, a random system that cannot distinguish between relevant and irrelevant documents will produce an ROC that is a straight diagonal line. The probability of a document being relevant or irrelevant is the same. A system that distinguishes perfectly between relevant and irrelevant documents will have an ROC that ranks all of the relevant documents before any of the irrelevant ones. Its ROC would travel up the vertical axis to the top and then travel across the top of the graph.

The area under this curve (A_{ROC}) is a good way to summarize the performance of this system independently of your criterion. An ideal system will have an area of 1.0. A totally ineffective system, one whose ranking is unrelated to the relevance of the documents, will have an area of 0.5.

This A_{ROC} measure has the advantage of characterizing a system's accuracy by a single number. It recognizes explicitly that there is a tradeoff between recall and precision that is not related to the power of the system to discriminate between relevant and irrelevant documents. One can place the cutoff between retrieved and nonretrieved documents anywhere along a system's retrieval operating curve, thereby changing its precision and recall without changing its retrieval effectiveness at all. Finally, the A_{ROC} has a straightforward absolute interpretation. Areas near 0.5 reflect poor performance whereas areas near 1.0 reflect excellent performance.

Precision and recall confound the sensitivity of the system (the ability to distinguish responsive from nonresponsive) with the bias of the user (how much evidence is necessary to count a document as responsive). An ROC analysis does not. An ROC analysis looks like it is measuring recall and precision at specific intervals, but it actually measures them at every level and it provides a natural interpretation of the resulting measure. Both approaches to measuring retrieval effectiveness suffer from the need evaluate large numbers of documents. Sampling may provide a way to approximate these measures with much less effort.

Estimating recall

Measuring system performance by either of these approaches typically requires a rather substantial effort. To measure precision, one has to evaluate all of the retrieved documents to determine whether or not they are responsive to the query. To measure recall, on the other hand, one has to evaluate all of the documents in a collection to be able to determine the proportion of responsive documents that have been retrieved. Similar requirements exist for ROC-type analyses. In any but tiny collections, this is a formidable amount of work. Sampling may provide some relief.

Precision and recall measures are designed to assess the degree to which the retrieval of responsive documents is accurate and complete. Of the two, precision is relatively easy to measure because one has only to assess the responsiveness of those documents that were retrieved. Recall, on the other hand, is much more difficult to measure because one has to know how many documents were actually responsive in order to know the proportion of responsive documents that were retrieved. This is practical only in relatively small data sets because it requires that every document be assessed for responsiveness. Sampling could be used to get an estimate of the total number of responsive documents, but it presents some challenges. Further, it is not immediately obvious what you can do with this information once it has been obtained.

In order to estimate the recall proportion using sampling techniques, we need to have a random sample of responsive documents and then determine the proportion of those that have been retrieved by the system. It is not clear exactly how one can get such a sample without having exhaustively determined which are the responsive documents. In small test sets this can be done with some effort, but in large document sets, especially in a litigation where there is no a priori way to determine what is responsive and what is not, this approach is simply not practical.

One approach to estimating recall is to take a random sample of documents (without regard to whether they have been retrieved or not) and evaluate this random sample for responsiveness. Once a reasonably sized sample of responsive documents has been obtained by this method, it is a simple matter to count the proportion of those documents that have been retrieved. The number of responsive emails that must be found using this random search procedure is

$$n = \frac{Z^2 p(1-p)}{C^2}$$

Where Z is the confidence level in units of the normal distribution and C is the confidence interval. The confidence level is the overall confidence you want to have in the quality of the results. Confidence levels of 95% to 98% are typical for most situations. We will use 98%, meaning that you are 98% confident in the outcome of the measure. Higher levels of confidence can only be achieved with much greater effort. By convention, statisticians often talk about the complement of confidence or α ($\alpha = 1.0 - \text{confidence}$, 0.02 to 0.05 in our examples). The confidence interval is how precise you want your estimate to be. When dealing with sample estimates, there is always some uncertainty about the estimate. The confidence interval is the range of that imprecision. Larger samples are needed to achieve smaller ranges. The true recall percentage will be within plus or minus $C \times 100\%$ of the one estimated from our sample. Another way of saying this is that if we repeated the estimate with a new random sample each time, then 98% of the time, the new sample would have an estimated proportion within $\pm C$ 98% of the time. We will use 0.03 for our desired confidence interval.

The next problem is to choose a level of p , which is the proportion of responsive documents that have been retrieved. Unfortunately, we do not know this proportion before we do the analysis. In fact, it is the very thing we are trying to estimate. The worst case from an estimation point of view is when $p = 0.5$. As a result, statisticians often use this proportion when computing the required sample size.

$$1508 = \frac{2.33^2 \times 0.5(1 - 0.5)}{.03^2}$$

Using these values, we will need at least 1,508 responsive emails to estimate recall with the accuracy we have specified. To get these we will have to review enough randomly selected documents to find 1,508 responsive ones and then count the proportion of those that were detected by the search process. This will be our estimate of the recall proportion over the whole document set. This could be a rather substantial number of documents, especially if responsive documents are rare.

Elusion and elusion sampling

Another measure may be more easily obtained and may be more useful in the discovery context. This alternative measure also has a natural translation into a quality control process. Rather than estimating the proportion of responsive documents that have been retrieved, it may be more practical to determine whether there were significant numbers of documents that were missed by the retrieval process. This measure, called "elusion," is related, but not identical, to recall. One can estimate elusion, the proportion of nonretrieved documents that are responsive and should have been retrieved, but it is more valuable to use this general approach to determine whether significant numbers of responsive documents have been missed. Elusion can be used as a quality check, equivalent to the kind of quality check manufacturers would use to determine whether their manufacturing process meets their standards.

You can use standard exhaustive review procedures to estimate elusion, but it is usually simpler to use a sampling procedure to determine whether the elusion rate exceeds a reasonable criterion. To assess elusion, you evaluate a randomly selected set of nonretrieved documents. If there are any responsive documents among the sample, you can adjust your retrieval criteria to detect these documents and then draw a new random sample. Following industrial standards we apply an "accept on zero" criterion—we only consider ourselves successful if there are no responsive documents in the sample. Elusion assesses what we missed.

Elusion can be estimated from a random sample of nonretrieved documents. The larger the sample, the more accurate is the estimate, but the marginal value of increasing the sample size diminishes as the sample gets larger. The optimal sample size for this measure depends on the confidence level desired and on the desired maximum probability of nonresponsive documents among the nonretrieved set. What is a reasonable effort expended to find responsive documents? How do we know that what we have done is reasonable? Again, we will select a confidence level of 0.98. Unlike our use of recall, we will use elusion to determine that there are no responsive documents that weren't retrieved or if there were, they were less prevalent than some specified maximum acceptable rate. To be absolutely certain that there are no responsive documents that were missed would require an infinite effort. We will have to settle, therefore, for a reasonable but rigorous level of confidence.

Recall that Blair and Maron (1985) found that the lawyers and paralegals in their study retrieved only about 20% of the responsive documents. These lawyers claimed that they would be satisfied with 75% of the responsive documents, and hence, might miss as many as 25% of the responsive documents. Human reviewers are also far from perfect. Many estimates of human relevance judgments put performance in the 20 – 60% range when reading large collections of documents.

In the following example, we have chosen the maximum prevalence of responsive documents in our nonretrieved set to be 2%. No more than 2% of the rejected documents are expected to be responsive (ps). This percentage can actually be set to any desired value. The lower the percentage, the more items have to be sampled.

The number of documents that must be sampled is determined by the formula

$$n = \frac{\log(\alpha)}{\log(1 - p_s)} = \frac{\log(0.02)}{\log(1 - p_s)} = 200$$

Based on these assumptions, 200 documents are randomly selected from those that were not retrieved. These documents are reviewed. If any of those documents are found to be responsive, then the discovery criteria are revised to capture those responsive documents and a new sample of 200 documents is selected. This process is repeated until the sample comes up with 0 responsive documents.

Rather than merely estimating our level of success, as we would do with recall, elusion sampling allows us to assess whether our entire process has succeeded to the level that we require. The biggest problem with using any information retrieval system is knowing what to look for. Using this elusion sampling measure allows us to assess the entire information retrieval process, including the formulation of our queries. If we have inadequately formed queries, then our elusion sample will uncover their existence and allow us to revise our criteria. There will be documents in the elusion sample and the test will have failed.

There are a number of measures of information retrieval effectiveness in addition to precision and recall. In search, these measures depend not only on the ability of the system, but on the ability of the users to derive queries that adequately cover the domain of responsive documents and on the administrators' decisions about how strict or liberal to place the criteria for retrieval. Rarely are tests conducted that would allow one to assess, from a quality perspective, the adequacy of these systems. Elusion provides one possible measure that translates directly into a quality assessment of the entire system, including the users and their queries. Elusion sampling requires only a modest amount of work that does not depend on the size of the collection, only on the specified confidence and minimum probability levels.

Herbert L. Roitblat, Ph.D.
OrcaTec LLC
PO Box 613
Ojai, CA 93024
herb@orcatec.com