The Sedona Conference Journal

Volume 8 2007

Search and Information Retrieval Science

Herbert L. Roitblat

Recommended Citation: Herbert L. Roitblat, *Search and Information Retrieval Science,* 8 SEDONA CONF. J. 225 (2007).

For this and additional publications see:
https://thesedonaconference.org/publications

SEARCH AND INFORMATION RETRIEVAL SCIENCE

Herbert L. Roitblat, Ph.D.
Orcatec LLC
Ojai, CA

I. A BRIEF HISTORY OF INFORMATION RETRIEVAL

The problem of how to use automated systems to find stored information is at least as old as computers themselves. In 1945, Vannevar Bush, Director of the Office of Scientific Research and Development, noted:

> The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships (The Atlantic, June 1945[1]).

Bush pointed to electronic, photographic and other emerging "mechanical aids with which to effect a transformation in scientific records." Even Bush's prophetic notions of how machines would make information storage and retrieval more effective grossly underestimated the difficulty of indexing and retrieving that information. The development of general purpose computers whose capabilities far surpassed those of the special purpose machines that Bush had in mind and the proliferation of machine readable information caused the information retrieval industry to explode.

Several mechanical systems, like those that Bush described, were in use in the 1940s and 1950s, some even later. Some systems used punch cards and rods. Each of the collection's topics was represented by a hole around the edge of a card. If a document was relevant to that topic, then the edge of its hole would be torn out. A knitting needle or other rod would be inserted into the deck and the whole collection would then be shaken. The relevant cards would fall out of the deck. Obviously, systems like these were very limited in the number of documents and number of topics they could handle.

The "Rapid Selector"[2] was another device that stored information on reels of microfilm along with machine-readable index marks. It was capable of storing 72,000 frames of information on 2000-foot film reels. Each frame consisted of a picture of a document page and an array of dots indicating the topics for which the page was relevant. At least this is the way it was supposed to work. According to some reports, the mechanics of the system were never quite up to the promise.
Later information retrieval systems continued to use similar index systems, in which each document is indexed by one or a few specific terms derived from preconstructed classification schemes and subject heading catalogs (The Library of Congress cataloging scheme is an example of such a catalog system for books). Thesauruses and other hierarchical vocabulary schemes made it possible to expand the

1 http://www.theatlantic.com/doc/194507/bush
2 http://www.ischool.berkeley.edu/~buckland/goldbush.html

vocabulary by adding more specific or more generic concepts than those originally specified. These systems are called controlled vocabulary systems because documents were only indexed by a few terms drawn from the hierarchy.

By the early 1950s information retrieval systems were moving away from the hierarchical catalogs to schemes based on single term descriptors or key words, coded without context. This innovation freed the searcher somewhat from the tyranny of the cataloger and allowed one to search for combinations of these single terms to produce complex queries. Searchers could use phrases and chains of synonyms. Researchers found that these coordinate keyword indexing methods were much easier to use than the hierarchical catalogs, and were no less accurate. They still relied on a controlled vocabulary of key terms, assigned by trained experts, but they could be combined in novel, unanticipated ways when searching the catalog.

Even after the introduction of commercial computers for information retrieval in the mid 1950s, indexers continued to rely on document descriptors that were drawn from a controlled vocabulary and were assigned by domain-experts. Computer memory (both internal core memory and external memory devices) was expensive, and indexing every word in the document would take up more storage than the documents themselves. Further, the documents to be retrieved were generally not machine readable, and, before OCR (optical character recognition), would have to be typed in by hand. These factors conspired to limit the range of terms that could be included in the search vocabulary. As a consequence, expertise was needed to catalog documents and further expertise was needed to find them again. Document ranking was not required to be too precise because the available computer resources at the time limited most searches to batch processing. A query would be submitted and some time later (from minutes to days) a printout of all the search results would be returned to the user.

Text analysis

It was generally believed in the 1950s that indexing documents, that is, assigning keywords to them, was a job that could only be done well by professionals. In this context, H. P. Luhn[3] introduced the idea that computers could not only handle the keyword matching and sorting task, but they could actually be used to analyze the content of written texts. He proposed automatic indexing and term weighting techniques based on the frequency and location of the words in the text. For example, a term that appeared twice in a paragraph or in two succeeding paragraphs could be considered a major concept in the document. His system would then use these terms to index the document. This idea was a major milestone in computerized information retrieval. Luhn is also credited with the idea that documents could be represented mathematically by term vectors.

Typically, each position on the list represents one word. If the word is present in a text, the corresponding element in the vector representing that text is set to be nonzero. If the word is absent, the corresponding element is set to 0. So, for example, if the 518th element of a vector represents the word "collusion" it is set to be 0 unless the word "collusion" actually happens to be in the text.

In a 1960 paper, Calvin Mooers[4] recognized that manually assigning descriptors to large volumes of documents was impractical. The job would have to be done by machine. He proposed an "inductive inference" machine, which takes as input a set of solved examples, such as a set of documents and the corresponding set of index terms and then uses an inference mechanism to derive the rules by which those index terms were assigned.

> An inductive inference machine is one that can be taught a series of correctly solved examples of problems so that it can proceed on its own (with some supervision and corrective intervention, probably) to the solution of other problems in the same class… It is capable of learning a great variety of tasks. (Mooers, 1960, p. 232)

3 *E.g.,* Luhn, H.P. (1957). "A statistical approach to mechanized encoding and searching of library information." IBM J. Res. Dev., 1, no.4, 309--17.
4 C.N. Mooers, "The Next Twenty Years in Information Retrieval," American Documentation, 11:3, July 1960, 229-236.

"To do more than merely pick out words by a frequency count, one would have to build into the method the capability of handling the equivalence classes of words and phrases" (Mooers, 1960, 231-232). In this way, Mooers not only anticipated the usefulness of Luhn's methods of automatic text indexing, but he predicted the use of neural network or other inference mechanisms to implement such methods. (Mooers is quoted by Salton, http://historical.ncstrl.org/tr/temp/ocr/cs-tr.cs.cornell.edu/TR87-827).

As recently as 1987, when neural networks and similar soft-computing solutions were in their infancy, and the most common machine inference systems were based on expert systems, Gerard Salton argued:

> It is now generally accepted that to obtain semantic interpretations of texts, a knowledge base is needed that classifies the main relationship between entities. Many theories of knowledge representation have been proposed over the years, and the evidence shows that when the topic under consideration is narrowly specified and when the text processing task is of limited scope, useful knowledge structures can in fact be prepared intellectually for use in practical systems.

Since that time, work with neural networks has proliferated and their ability to address aspects of semantic representations has been validated to a substantial degree. Nevertheless, most information retrieval systems seem to be based either directly on versions of Luhn's original proposal or on Mooers' and Salton's view that expert intervention is necessary to build intelligent indexing and retrieval systems.

II. THE VECTOR SPACE MODEL

The vector space model[5] is arguably one of the most important inventions in the field of information retrieval. It allows texts to be represented by numbers. A vector is an ordered list of numbers. Each position in the list corresponds to a specific word. As a result, there is one cell or value for each word in the vocabulary. Two examples are shown in Figure 1 on this page and Figure 2 shown on the following page. In each vector, the cells that correspond to words in the text are set to be nonzero. The other cells represent words that do not happen to be in this particular text and they are set to be 0. Word order in the text is not represented in the vector. For this reason, this approach is also called a "bag of words." The system represents whether the word is present or not, but not the word's position in the original text.

A primary advantage of using the vector space model is that the tools of matrix algebra can be applied to making decisions about texts. Two texts are similar if they contain similar words. One way to compute this comparison is to multiply these two vectors together according to the rules of matrix algebra. Comparing texts is much more efficient for a computer using matrix algebra than it is trying to compare them letter by letter.

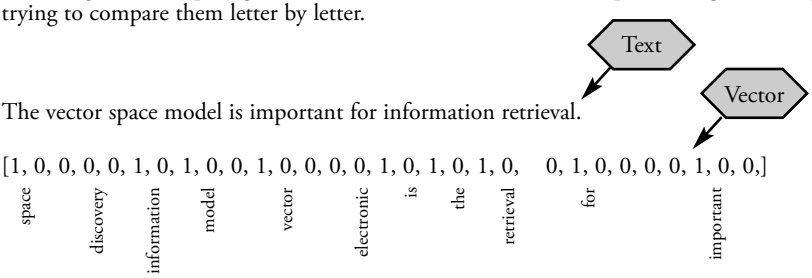

Figure 1. A text and a victor representation of it.

5 *See also* Salton, G., Wong, A., and Yang, C. S. (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613-620.

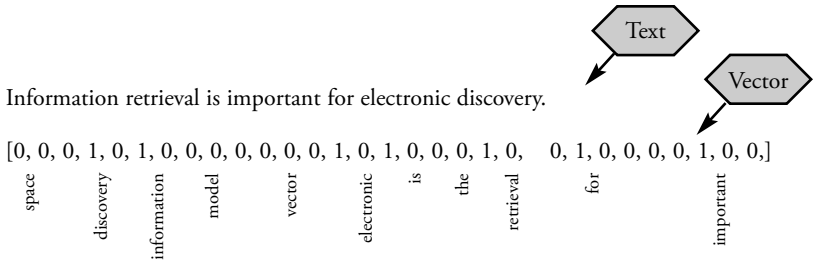

Figure 2. A text and a victor representation of it.

III. TERM WEIGHTING

The vector notation makes working with representations of documents more mathematically convenient, but it does not solve the problem of how to make those representation vectors. Two important problems remain. First, what content units are to be included in the representation vectors, and, second, how are the values in the vectors to be determined so as to distinguish the important terms from the less important ones? The two examples in Figure 1 on the previous page and Figure 2 above set the vector cell to 1 when the word is present in the text. But, not all words are equally important. It is usually more effective to weight the vector values by some estimate of how important the word is. For example, words that occur more frequently in the text might be more indicative of what the text is about. Words that are more rare in the document collection may also be more informative than words that are more common.

For example, one common approach is called *TFIDF*. A word is considered more important, and, thus, receives a higher value if it occurs more often in the text being represented (*TF*, Term Frequency), and it is considered less important if it occurs in many documents (*IDF*, Inverse Document Frequency).

In most of the early automated information retrieval systems, context-free index terms were used, but these ultimately proved inadequate. Context-free means that the word gets the same value no matter where it appears. The problem is that the word "insurance" could be very important in documents about a warehouse operation, but very unimportant in documents about an insurance company.

Every word in the examples in Figure 1 and 2 is indexed in the vector. In contrast, many systems distinguish between content words, which convey information about the content of the text and function words, which do not. These function words are contained in a "stop list" or list of "stop words" or "noise words." The stop words are not indexed because they are too common to be of use in distinguishing between documents. In essence, the stop words have a weight of 0. They are ignored.

Still other weighting schemes exist. Later we will consider systems that use different kinds of vectors to represent texts. Instead of each word having a cell in the vector, various groups of words are assigned to the same cell.

IV. QUERY EXPANSION

Human language use is diverse. People use a wide variety of words to refer to the same concepts (synonymy) and use the same words to refer to different concepts (polysemy). Relying on individual words to indicate the contents of documents produces very poor search results because, for example, the system has no way to know whether a user who enters the word "pitch" is interested in baseball, golf, sales, or aerospace controls. Similarly, a person might use the words "youth," "child," or "boy" in a text and it is difficult to predict which one was actually used without looking at the text. Any given text, therefore, is only one sample of the words that could have been used to express the same idea.

Several techniques have evolved to deal with these problems that derive from people's flexible language use. One of these, restricting the search vocabulary, has already been discussed. Whatever specific words were used in the document, the cataloger coerces the document into one or a few categories from the controlled vocabulary. Similarly, the terms a searcher has in mind have to be coerced, perhaps with the use of a thesaurus, into the same controlled vocabulary terms used by the cataloger.

Controlled vocabulary systems may work well with professionally trained search intermediaries, but they are unacceptable to a the broader search audience. Generally, free text indexing has been found to be no worse and often better than controlled vocabulary approaches. In any case, the process of classifying documents into a small number of categories is very time consuming and highly dependent on the skill of the cataloger.

Other approaches have involved augmenting the original search with additional search terms (called query expansion) or translation schemes that convert natural language queries into formal search queries or into preconstructed and edited search queries (e.g., Ask Jeeves, now Ask.com).

Query expansion converts the term or terms that the searcher enters into a more elaborate query that usually involves the original terms and some additional terms. Query expansion can be based on morphological characteristics of the words, word association patterns, thesauri, taxonomies or ontologies.

The morphology of a word is its semantic structure. Words consist of one or more morphemes. A morpheme is the smallest unit of meaning in the language. The word "unfriendly," for example, consists of three morphemes, "un" meaning "not," "fiend," meaning "companion," and "ly," which indicates that it is a an adverb. Morphological expansion is often useful, in that a system that "knows about" government should also know about "govern" and "governing." In English, stemming is a simple form of morphological analysis. In stemming, the main stem of a word is separated from its inflections. For example, "swimming" consists of a stem, "swim" and an inflection or suffix, "ing." When stemming is used, a search for "swim" or "swimming" will yield the same results, all documents containing either of these word forms. Stemming does sometimes lead to inappropriate expansions (e.g., expanding a query for "Miller" to include "Mills," "Mill," "Milled" and "Milling"), but there are techniques for mitigating such difficulties.

Thesaurus-based expansion depends on the existence of an appropriate and pertinent thesaurus. Some systems employ, for example, semantically encoded dictionaries such as the Longman Dictionary of Contemporary English or George Miller's WordNet[6], a manually constructed thesaurus based on psycholinguistic assessments of lexical relations. Thesauri such as these are difficult to use, because they include multiple senses and multiple meanings for each word. It can be difficult to select the intended meaning from a brief query and it may actually decrease the accuracy of some queries. In some specialized areas with their own distinct dialects, a general thesaurus can greatly interfere with correct selection. For example, in some collections a search for "cleric" might be expanded via the thesaurus to "minister." This expansion would probably be useful in the US, but would give erroneous results in documents about the Middle East, where a minister is a member of the government, and cleric is more tightly related to "Ayatollah." Solving these problems with thesauri can require a great deal of human editing, and they are very brittle-they cannot adapt easily to changes in the language community. For example, it would take considerable expert human intervention to be able to respond to new terms such the emergence of the MP3 file standard, or podcasting.

Similarly, the use of simple co-occurrence patterns, even in text of known relevance, is less than completely satisfactory because these patterns do not capture the overall regularities in the language usage. Equally associated terms are not always equally pertinent. For example, two terms that co-occur frequently in a document may have a different value than two terms that occur only rarely in the document, but always together. These simple mechanisms fail to take redundancy into account.

6 Fellbaum, C. (1998). *Wordnet: An electronic lexical database.* Cambridge, MA: MIT Press.

Some mechanism (either human intervention or some learning/inference system, such as a neural network) is necessary to properly weight the co-occurring terms.

Weighting terms or even pairs of terms indiscriminately, has been found to be inadequate for useful searches. Some form of differential weighting is used by practically every modern search system. For example, most information retrieval systems employ some form of term-frequency weighting-terms that occur more often in a document are considered to be more important (all other things equal) to describing the document's content than terms that occur fewer times. Similarly, terms that occur in fewer documents are considered to be more important (again, all other things equal) than words that occur in many different documents. Although term weighting usually helps to improve performance, no system is yet perfect in retrieving only and all relevant documents. Neural networks can introduce more flexible weighting schemes employing nonlinearities in the weights that vastly improve retrieval precision.

V. SYNTACTIC TECHNIQUES

Another approach to identifying what is important in a document is to do a computerized analysis of its syntactic structure. The vector space model discards word order. Syntactic approaches, in contrast depend strongly on word order and other cues to parse each sentence into a tree structure. The topics of a text are usually contained in its nouns, so identifying the nouns in a text will go a long way toward identifying its most meaningful and characteristic components.

Names, of people, places, organizations, and so on, can also be important carriers of information. Although less than perfect, systems exist that attempt to identify that the word "Bush" is a name in the sentence "President Bush declared Louisiana a disaster area," but not a name in the sentence "He sat in the place that, before the hurricane, was occupied by a blueberry bush."

Some systems are capable of recognizing that "Bush" and "President" are not just names, they refer to the same person or entity. These systems can be very useful for retrieving information where a person could be referred to by his name, his title, the name of his house, and so on. They create a kind of thesaurus that describes the named people, places, and organizations they have detected in the different ways found in the document collection. Like a conventional thesaurus, this entity thesaurus can also be used to expand queries.

Another related source of document information exploits its overall structure. For example, the title of a document is often more valuable in identifying what it is about than any other paragraph in it. In emails it is usually easy to identify who the sender and who the recipient was, but in memos, it may not be so easy for a computer to identify this information.

VI. USER INTERFACE

Systems differ substantially in how they interact with the user. Traditional systems tended to use simple queries that consist of one or a few words. Later systems allowed users to enter complex Boolean expressions as queries (e.g., "(air or water) w/5 (pollution and controls)"). Boolean systems allow users to specify the logical relations between query terms (such as OR, AND, and NOT). Many allow nested input.

These Boolean expressions can become quite complex and creating them is, for many, a formidable task. Many modern systems allow users to enter natural language queries such as:

> What factors are important in determining what constitutes a vessel for purposes of determining liability of a vessel owner for injuries to a seaman under the Jones Act (46 USC 688)? (Turtle, 1994, p. 213)[7]

7 Turtle, H. (1994). "Natural language vs. Boolean query evaluation: a comparison of retrieval performance". Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Pp. 212-220. New York: Springer Verlag.

These queries are not constrained to any particular format. Turtle (1994) found that using natural language queries like this were more effective at retrieving case law documents from Westlaw than were more traditional Boolean expressions, like

> (741 +3 824) FACTOR ELEMENT STATUS FACT /P VESSEL SHIP BOAT /P (46 +3 688) "Jones ACT" /P INJUR! /S SEAMAN CREWNAN WORKER (Turtle, 1994, p. 213)

Other systems do not rely on queries from the user to organize the documents in their collection. They may use various kinds of diagrams or maps to display groups of documents and allow the user to select groups of documents for further investigation.

VII. INFORMATION RETRIEVAL IN E-DISCOVERY

Tools like those described above have been widely applied in electronic discovery, but there are very few published studies that discuss their effectiveness.

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.[8] The attorneys worked with experienced paralegal search specialists to find all of the documents that were relevant to the 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.

VIII. 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. Recall is the proportion of responsive documents that have been retrieved.

$$Precision = \frac{n_{Responsive\_Retrieved}}{n_{Retrieved}} \qquad Recall = \frac{n_{Responsive\_Retrieved}}{n_{Retrieved}}$$

8 Blair D. C. & Maron, M. E. (1985). "An evaluation of retrieval effectiveness for a full-text document-retrieval system", Communications of the ACM, 28, 289-299.

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.

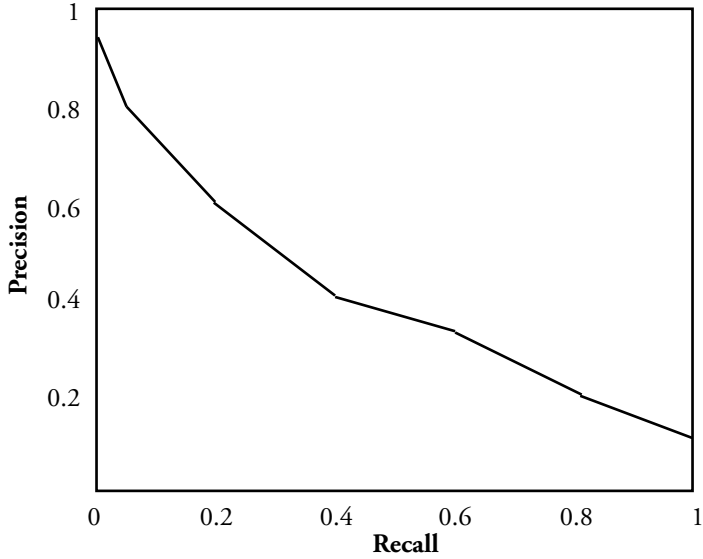

Figure 3. An example of the trade off between precision and recall.

An example of this trade off is shown in Figure 3. 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 the most 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. Because the accuracy of the system is expressed as two numbers, it is often difficult to tell whether one system is better than another if, for example, it yields slightly higher recall and slightly lower precision than another system. Information retrieval investigators have developed a number of measures that combine these two into a single number. One of these is called F1 or van Rijsbergen's F. It is the harmonic mean of precision and recall. Other measures average precision obtained at different levels of recall.

In addition to the system's accuracy, there are other factors to consider when choosing among information retrieval systems. van Rijsbergen[9] suggests the following factors be considered when comparing systems:

- Coverage-how much relevant material can the system access?
- Speed-how quickly does the system return results?
- Effectiveness of the output-does the system present the results in a usable manner?
- Effort-how much trouble is it to set up and use the system?

In addition, evaluations might want to consider these factors:

- Usefulness-does the document provide information that is of value to the user?
- Cost-how much does it cost to get the information?
- System boundaries-to what extent does the system rely on the skill of the user to achieve high levels of performance?

In practical terms, the performance of a system is not just a matter of how well it responds to specific queries. Systems can differ in their coverage. Some systems are better able to extract usable information from a wider variety of document types than others are. Especially in large collections, the speed with which the system operates can be an important factor. The time it takes to respond to a particular query is only one of the speed measures that might be considered. This time might affect the productivity of the system.

Even if a system properly distinguishes between responsive and nonresponsive documents, how it presents this information to the user is also important. Systems differ in the effort it takes to act on the information retrieved. Some systems are just easier to use than others. Some fit with your workflow patterns better than others. Some system have more panache than others. They may look great, but actually be difficult to use.

Documents with the same relevance may differ substantially in their usefulness. Finding the 23rd copy of a significant email, for example, is much less useful than finding the first copy. Cost is almost always a consideration in electronic discovery. Can the cost of finding a few more documents be justified by the value of the documents that are retrieved? Two systems may differ only marginally in their effectiveness. Can you justify the extra expense of using the better system by the quality of the results you will receive?

System performance does not depend only on the mechanical or computational characteristics of the system, but also on the skill and knowledge of the user. The quality of the retrieved document set may depend heavily on the skill of the person formulating the queries. The role of these skills in determining the results should be evaluated along with other measures of system performance.

IX. 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.[10]

9 Rijsbergen, C. van (1979). *Information Retrieval*, London: Butterworths.

Figure 4 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.[10]

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, few real systems 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.

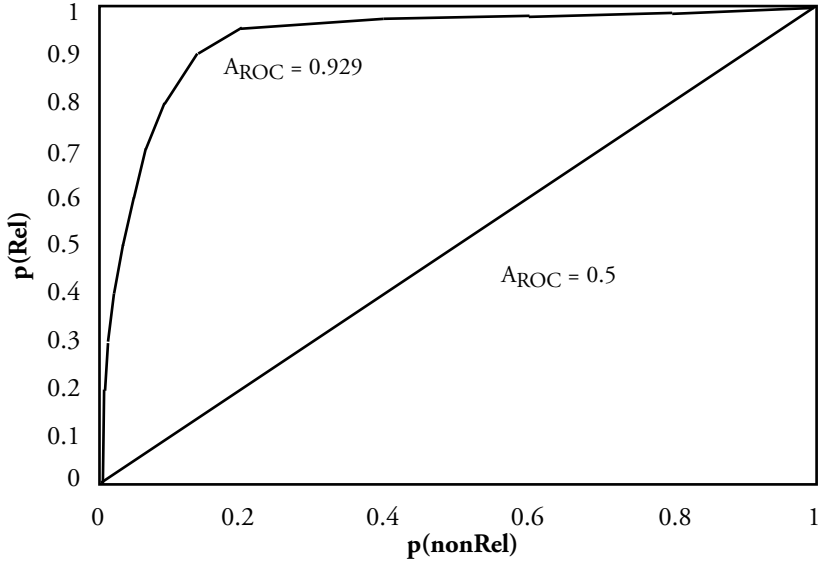

Figure 4. The Retrieval Operating Characteristic (ROC) curve also shows the trade off between precision and recall, but in an easy to summarize way. The accuracy of the system is characterized by the area under the curve (A_{ROC})."

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 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.

10 Swets, J.A., (1963) "Information retrieval systems", Science, 141, 245-250. *See* discussion in Rijsbergen, C. van (1979). *Information Retrieval*, London: Butterworths.

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.[11] 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 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 relevant from irrelevant 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 AROC 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.

X. ESTIMATING RECALL

Measuring system performance by either of these approaches typically requires a rather substantial effort. 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

11 *See* John A. Swets and Ronald M. Pickett, *Evaluation of diagnostic systems: methods from signal detection theory*, Academic Press, New York, 1982.

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 the whole collection in order to know the proportion of those actually 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.

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 documents 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 (α = 1.0 - 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.

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

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. In any case, however, it is far fewer than all of the documents in the collection.

XI. ELUSION AND ELUSION SAMPLING

Another, new, 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. Recall is the proportion of responsive documents that have retrieved and elusition is the proportion of nonretrieved documents that are responsive and should have been retrieved.

One can estimate elusion, 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 sampling procedures to estimate the actual elusion rate, but it is usually simpler 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.

The optimal sample size for this process depends on the confidence level desired and on the desired maximum probability of nonresponsive documents among the nonretrieved set—the quality standard. What is a reasonable effort expended to find responsive documents? How do we know that what we have done is reasonable?

Unlike the use of recall, we will use elusion to determine that there are no substantial numbers of responsive documents that weren't retrieved. 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. Again, we will select a confidence level of 0.98.

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

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

In this 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 (P_S). This percentage can actually be set to any desired value. The lower the percentage, the more items have to be sampled. Reducing the maximum prevalence to 1%, for example, requires that almost 400 documents need to be sampled. The maximum prevalence you set will depend on what you think is reasonable performance of the review process. The number of documents that must be reviewed for various confidence levels and various maximum prevalence is shown in Table 1.

Table 1. The estimated number of documents to review to achieve specified levels of confidence and maximum acceptable error rates (ps).

	Confidence			
*P*s	0.999	0.995	0.99	0.98
0.0001	69075	52981	46050	39119
0.0005	13813	10594	9209	7823
0.001	6905	5296	4603	3911
0.005	1379	1058	919	781
0.01	688	528	459	390
0.02	342	263	228	194

Based on these assumptions, 200 documents (rounded up from 194) 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.[12] This process is repeated until the sample comes up with 0

12 Technically, this repeated sampling introduces a small additional chance of error because of repeatedly sampling from the same population. This error has been ignored here, but it can be accounted for by raising the desired level of confidence.

responsive documents. Rather than merely estimating our level of success, as we would do with recall, this measure 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.