Published by Pike & Fischer, a BNA Company

Digital Discovery & e-Evidence

BEST PRACTICES & EVOLVING LAW

http://ddee.pf.com

Reprinted from Vol. 7, No. 1 | January 2007

VENDORS & TECHNOLOGY

Searching in all the Wrong Places: The Effectiveness of Search Tools in E-Discovery

By Bruce Hedin, Ph.D.

Introduction

As technology accelerates the rate at which documents can be created and propagated, corporations and law firms are finding that meeting their document review and management requirements is both increasingly important and increasingly challenging.

For a corporation, failure to meet document retention or production requirements can put the company at serious legal and financial risk. For a law firm, failure to retrieve the documentary evidence needed to advance a line of questioning in a deposition or to support or contest claims at trial can jeopardize the firm's ability to litigate a case effectively. Failure of both kinds is made more likely by the enormous increases in the sizes of the document populations with which companies and law firms must contend.

Growing awareness of these challenges adds to the impetus for industry standards in the management of electronic documents. The Sedona Conference and the Electronic Discovery Reference Model (EDRM) Project represent collaborative attempts at developing accepted standards and best practices for e-discovery.[1]

Such guidelines are a welcome step in bringing order to this new domain. But they don't solve the crucial problem at the heart of current document management efforts: identifying relatively small numbers of relevant records in very large populations of documents.

For help solving this problem, corporations and law firms have turned to various technology-supported information retrieval (IR) systems.[2] Most of the solutions currently on the market rely on user-constructed queries of a document collection, queries which may be entirely manual (e.g., standard keyword or Boolean searches) or assisted by some kind of artificial intelligence (e.g., one of the varieties of "concept search").

However, more than 20 years worth of scientific studies show that query-based approaches to information retrieval (a) perform poorly and (b) fail to provide the user with information on just how poorly the system is performing. This article will provide a brief overview of these studies and identify the key challenges that any successful solution to document review and management problems will have to overcome.

Information Retrieval 101: A Primer on Precision and Recall

The legal community remains for the most part woefully uninformed about search and retrieval concepts that are considered elementary in the field of information retrieval. A basic understanding of these concepts is essential to understanding electronic discovery processes. The more informed members of the legal community are about various search methods and technologies, the better they will be able to apply them in evaluating the performance of different approaches to document retrieval and in assessing the likelihood that a given approach will meet their needs.

First, various terms must be defined. The two most important measures used in the evaluation of IR systems are *recall* and *precision*. These terms are information retrieval terms of art, typically expressed as percentages.

In the context of document review, these two measures can be used to gauge the overall effectiveness of review systems. A measure of both recall and precision is necessary in order to determine the accuracy of review results. Moreover, recall and precision can be used to compute downstream review costs, as well as time to completion.

Recall measures how much of a target set has been found — i.e., how many relevant documents have actually been identified as such. For example, an IR system that achieves 80 per cent recall means that 80 per cent of all

relevant documents were actually found, and 20 per cent of all relevant documents were not found during the review.

Recall is critical in litigation because (1) parties have an obligation to produce documents responsive to discovery requests; and (2) a more complete set of relevant documents enables a more thorough review of evidence to develop an effective case strategy — and reduces the risk of overlooking information that could be critical to the case.

Precision measures how much of a result set is on target — *i.e.,* how many of the returned documents are actually relevant. For example, an IR system that achieves 75 per cent precision means that 75 per cent of all documents that were assessed to be relevant were actually relevant, and 25 per cent of documents were erroneously identified as such.

Precision is critical in litigation because a low precision rate means that a large number of irrelevant documents will need to be reviewed. This translates into considerable costs, as well as significant delays in time to completion.

Studies of the Effectiveness of Manually-Built Boolean Queries

Studies have found that manually created Boolean queries perform poorly on both of the standard performance metrics used in assessing IR systems — recall (how much of the target set is found) and precision (how much of the result set is on-target). Findings of some of the landmark studies in this field are summarized below.

Blair & Maron 1985. This important early study investigated the ability of a manual Boolean full-text retrieval system to locate documents of interest to attorneys. In this study, the lead defense attorneys in a lawsuit formulated a number of information requests pertinent to the case. Paralegals were then charged with turning these requests into Boolean queries and running them over a collection of roughly 40,000 documents. The attorneys evaluated the results and provided the paralegals with feedback to refine their queries.

This feedback loop was repeated until the attorneys felt satisfied that each query had succeeded in retrieving 75 per cent of the material relevant to the request (i.e., until each query had achieved 75 per cent recall). Once the process was completed, both retrieved and unretrieved documents were analyzed in order to obtain recall and precision numbers for the queries.

The analysis found that, while the Boolean queries had reasonably good precision (on average, 79 per cent), the recall was quite low (on average, 20 per cent). Although the attorneys believed, from their evaluation of the returned documents, that they had found three out of every four relevant documents, they had in fact succeeded in retrieving just one out of every five. (Blair & Maron 1985; Dabney 1986)

Turtle 1994. Researchers at Westlaw compared the retrieval performance of manually created Boolean queries to that of automated systems using natural language issue statements. They devised a set of 44 target issues, modeled on the sorts of problems attorneys might research, and compared each system's ability to retrieve relevant material from collections of case law. Expert searchers were assigned the task of devising optimal Boolean queries for each issue. For the natural language system, they created issue statements (typically a single sentence for each issue).

The study found that, both on average and on nearly every target issue, the natural language system outperformed the Boolean system with regard to both recall and precision. In fact, on the test collection with the most reliable relevance judgments, the Boolean approach reached just 24.4 per cent recall, even at the low precision level of 42.3 per cent.

However, the natural language system's results were not particularly impressive either. It achieved recall of 32.9 per cent at a precision level of 57.0 per cent (a little over half of the result set was actually on target). (Turtle 1994)

TREC. The Text REtrieval Conference, or TREC, is co-sponsored by the National Institute of Standards and Technology and the U.S. Department of Defense. Initiated in 1992, its goal is "to support research within the IR community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies." Every year, it creates a particular IR challenge, with a large, static document population, on which experts can test their systems. The results are identified by the challenge number. (1992 was TREC-1, 1993 was TREC-2, and so forth.) For more information, see http://trec.nist.gov.

Ad Hoc Searches (TREC-6). The TREC-6 ad hoc task assigned participating systems the objective of finding material relevant to 50 topics in a population of news and government documents. All kinds of systems were tested on the task, both natural language systems and Boolean, both automated approaches and manual.

The sole contestant based on the manual construction of Boolean queries did poorly, even compared to five other manual approaches. In order to achieve even 50 per cent recall, the Boolean system could get no higher than 20 per cent precision (one out of five documents in the result set being on-target) (Voorhees & Harman 1997; Leong 1997).

Due to this poor showing, manual Boolean systems weren't tested in following TREC studies.

"Concept Searches": Automated Extensions to Boolean Approaches

The studies noted above focus on the performance of systems relying on manually constructed Boolean queries. An obvious limitation of such approaches is that casting a search net is limited entirely by user input. If a user remembers to include a particular variant of a target expression in her query, the query will return documents containing the variant. But if she doesn't think to include a particular variant, documents containing that variant won't be returned.

Recognizing this limitation, researchers have sought to build "concept search" systems that will automatically expand a user's initial input to the full range of variations for expressing the concept it represents. The user need not identify in advance all the ways a particular concept can surface linguistically. He or she must only provide sufficient indication of the target concept, and the system will attempt to expand the query to include conceptually related patterns.

Underlying all varieties of concept search is an information-retrieval model (relying on approaches such as vector space modeling, neural networks, or Bayesian networks) that governs the method by which the user's input is expanded.

With such sophisticated approaches, one would expect superior performance. Indeed, studies of these models' performance have found that they deliver better results than standard manual Boolean queries — but only marginally better.

In Turtle's 1994 study, for example, the natural language system studied (which relied on a variant of a Bayesian model) outperformed the Boolean system with regard to both recall and precision. However, its results were not good enough to be very useful in daily practice. The system found only about one-third of the relevant material, and irrelevant documents made up nearly half its result set.

The two key limitations that make manual Boolean systems so unattractive also hold for the natural language systems (and the "concept searches" built on top of them):

- They require negotiation of a very steep trade-off between recall and precision.
- They leave the user in ignorance as to the actual performance of the system.

This means we are hardly better off than before in terms of solving the document review problem.

Challenges to be Addressed

If query-based approaches are not the answer to the document retrieval challenges corporations and law firms face, what would an adequate answer look like? A complete description of a solution would be too long for purposes of this paper. In short, there are three key challenges that any document management and retrieval system will need to address, if it is to be successful:

- Knowledge transfer between the expert — e.g., a senior litigator — and the reviewers (a teaching challenge);
- Consistency in how each individual reviewer assesses documents (a mental discipline challenge); and
- Quality assurance mechanisms, i.e., providing valid information on the system's performance, in terms of precision and recall (a quality assurance and process design challenge).

These challenges, by their very nature, are not addressed by an end-user search tool. Instead, what is required is an expertly designed combination of technology, efficient human work processes, and rigorous performance measurement techniques.

Conclusion

The studies cited above show that the query-based methods discussed in this document don't solve the document management and review challenges corporations and law firms currently face.

The methods can be expected to identify no more than half, and in all probability much less than half, of the documents they were employed to find. And they will do so at the cost of including a very large amount of off-target material in the result set. But these studies are seldom publicized outside the academic discipline of information retrieval.

Therefore, even when the system is performing poorly, users of the system will be dangerously unaware that it's wasting their time and resources with irrelevant results – and even more dangerously, failing to find the relevant documents.

Bruce Hedin, Ph.D. *is a Senior Consultant in the Professional Services Group of* ***H5*** *(www.h5technologies.com), an automated document analysis and information risk management services firm. Dr. Hedin works with corporate counsel of Fortune 500 companies and senior litigators of AmLaw 200 law firms to help effectively manage the risks involved in missing vital information in large-scale electronic discovery, records retention, and regulatory compliance reviews. Dr. Hedin's areas of expertise include statistics, linguistics, and domain analysis. He is based in San Francisco, California.*

(Endnotes)

1. See http://www.thesedonaconference.org and http://www.sochaconsulting.com/referencemodel.htm.

2. "In the academic world, 'information retrieval' is defined as '[the] actions, methods and procedures for recovering stored data to provide information on a given subject." Baron, "Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery," *The Sedona Conference Journal,* Volume VI at 240 (2005) (citing ISO 2382/1).

