Concept Search: Perceived Security, Actual Risk

H5 White Paper

H5

Abstract

Large organizations increasingly face document management and review requirements that exceed the capacity of traditional manual approaches. Recognizing this, corporations and law firms are turning to various technology-supported information retrieval systems. Most of the approaches on the market rest on user-constructed queries, which may be entirely manually constructed or assisted by some form of automation. Studies have consistently found that such approaches to information retrieval perform poorly, requiring the negotiation of an expensive trade-off between recall and precision. Furthermore, these approaches fail to provide users with crucial data on how much of the target set has been captured by a query. Without such data users cannot assess the performance of a search or identify the optimal trade-off between recall and precision. The result is often users who dangerously believe they have found more than they have actually found and who think they know more than they actually do know. Such users can put themselves, their organizations, and their clients at great risk. To avoid such risk an approach is needed that both reduces the severity of the recall-precision trade-off and provides users with visibility into performance on recall and precision. Only such an approach will allow businesses and law firms to meet their document management and review objectives in a clear-sighted, safe, and cost-effective manner.

Introduction

As technology accelerates the rate at which documents can be created, corporations and law firms find that meeting their document management and review requirements is both increasingly important and increasingly challenging. For a corporation, a failure to meet document retention or production requirements puts the company at serious legal and financial risk; for a law firm, a 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 jeopordize the firm's ability to litigate a case. Failure is made more likely by the enormous increases in the sizes of the document populations with which companies and law firms must contend.

As awareness of these challenges spreads, the impetus for establishing industry standards in the management of electronic documents grows. The Sedona Conference Guidelines represent a collaborative attempt at developing accepted standards for e-discovery.[1] The Electronic Discovery Reference Model is another attempt at developing a set of best practices around e-discovery.[2] While such guidelines are a welcome step in bringing order to this new field, they do not solve the crucial problem that remains at the heart of current document management efforts: the identification of relatively small numbers of relevant records in very large document populations.

For assistance in solving this problem, corporations and law firms have turned to various technology-supported information retrieval systems. Most of the solutions currently on the market rely on user-constructed queries of a document collection, queries which may be entirely manually constructed (e.g., standard manually constructed Boolean queries) or assisted by some automated system of accuracy improvement (e.g., one of the varieties of "concept search"). The effectiveness of such approaches has been well studied, both in the

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

Text REtrieval Conference (TREC)[3] series and elsewhere. In this paper, we review the academic literature on query-based approaches to information retrieval, summarize the limitations of these approaches, and identify the key challenges that a successful solution to the document management and review problem sketched above will have to overcome.

Studies of the Effectiveness of Manually Built Boolean Queries

Studies have found that methods utilizing manually created Boolean queries perform poorly on both of the standard performance metrics used in assessing information retrieval systems: recall (how much of the target set is found) and precision (how much of the result set is on target). A summary of the findings of a cross-section of these studies follows.

Studies on the TREC Ad Hoc Task. The TREC Ad Hoc Task is designed to test the performance of information retrieval systems in identifying material on a range of different pre-defined topics in a static document population. 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. A wide range of systems were tested on the task, both natural language systems and Boolean, both automated approaches and manual.

Just one of the systems tested in TREC-6 relied on the manual construction of Boolean queries. When compared with the others, the manual Boolean system was found to perform worse than five other manual approaches. In fact, in order to achieve even 50 percent recall (i.e., bringing half of the target documents into the result set), the system required the acceptance of 20 percent precision (i.e., only one out of five documents in the result set being on target). (Voorhees & Harman 1997; Leong 1997)

Perhaps because of the limited promise shown by the approach, no further testing of manual Boolean systems on the Ad Hoc Task was undertaken following TREC-6. All testing on the Ad Hoc Task was discontinued following TREC-8.

Studies on the TREC Interactive Track. The TREC Interactive Track studies the effects of a user's interaction with a text retrieval system. The Interactive Track assigns participating systems the objective of finding material relevant to a small set of pre-defined topics in a finite population of documents. For TREC-6 (1997), participants were tasked with finding material relative to six topics; for TREC-7 (1998), participating systems were tested on their ability to retrieve material relevant to eight topics. The population searched in both TREC-6 and TREC-7 was a collection of news reports from the *Financial Times*. In each submission to the Interactive Track, the performance of an experimental system is compared to that of a control system implemented at all participating sites.

In a pair of related studies, one for TREC-6 and the other for TREC-7, researchers at the Oregon Health Sciences University (OHSU) compared an approach to text retrieval that used manually created Boolean queries to an approach that used an automated natural language system. The TREC-6 study found that the Boolean approach performed worse than the natural language searching system. In fact, while the OHSU team was among the top performers on the natural language system with regard to recall, OHSU's manual Boolean system scored among the lowest on recall of those submitted to the track. (Voorhees & Harman 1997; Hersh & Day 1997)

3 The Text REtrieval Conference 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 information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies." For more information, see http://trec.nist.gov.

OHSU's TREC-6 study relied on a relatively small result set (data was collected from just four searchers) making it difficult to establish the significance of the findings. For this reason, the OHSU researchers followed up their 1997 study with a larger-scale experiment for TREC-7. In the TREC-7 study, data was collected from 24 searchers, all of whom were skilled at and enjoyed information retrieval tasks (all were information professionals with library degrees; on average, subjects had 7.8 years of on-line search experience). This study found that, in the hands of these highly experienced subjects, the Boolean system was able to reach performance levels equal to that attained by the natural language system with which it was compared. However, the study also found that both approaches performed poorly: at approximately 70 percent precision, neither system was able to exceed 35 percent recall (i.e., neither system was able to find more than about one-third of the relevant material). (Voorhees & Harman 1998; Hersh, et al. 1998)

The OHSU studies are evidence that, in the hands of highly skilled searchers, a Boolean system may perform equally as well as a natural language system, but the performance of that natural language system is itself unsatisfactory.

Turtle 1995. In a 1994 study published in 1995, researchers at Westlaw compared the retrieval performance of manually created Boolean queries to that of automated systems using natural language issue statements as input. For the study, researchers devised a set of 44 target issues modeled on the sorts of problems attorneys might research. The ability of each system to retrieve material relevant to the target issues from collections of case law was then compared. In implementing the Boolean system, expert searchers were assigned the task of iteratively devising the optimal Boolean query for each issue. In implementing the natural language system, the issue statements (typically a single sentence characterizing the issue) were fed into each of the tested systems.

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 was able to achieve just 24.4 percent recall (i.e., the expert searchers less than about one-quarter of the target documents) even at the low precision level of 42.3 percent (i.e., less than half of the result set was on target). By comparison, the natural language system achieved recall of 32.9 percent (i.e., the system found about one-third of the target documents) at a precision level of 57.0 percent (i.e., more than half of the result set was actually on target). (Turtle 1995)

Blair & Maron 1985. In a seminal study on information retrieval, the ability of a manual Boolean full-text retrieval system to locate documents of interest to attorneys was investigated. 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 creating and executing Boolean queries on a collection of roughly 40,000 documents in order to retrieve material relevant to the requests. The attorneys evaluated the results returned by the queries and provided the paralegals with feedback so that the queries could be improved. The process was allowed to continue through as many iterations as were needed until the attorneys were satisfied that each query had succeeded in retrieving three-quarters of the material relevant to the request (i.e., until each query had achieved 75 percent recall).

Once the process was completed, both retrieved and unretrieved documents were analyzed in order to obtain recall and precision metrics for the queries. The analysis found that while the Boolean queries performed well with regard to precision (achieving, on average, 79 percent), the recall was quite low (on average, 20 percent). 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)

The findings of the cited studies are consistent with the findings of other studies of the performance of manually created Boolean queries. All point to poor performance with regard to both recall and precision and to the necessity of aquiring even modest levels of recall at the price of very poor (and in terms of necessary review resources, extravagantly wasteful) levels of precision.

"Concept Searches": Automated Extensions to Boolean Approaches

The studies discussed above focus on the performance of systems relying on manually constructed Boolean queries. An obvious limitation of such approaches is that the scope of the search is limited entirely by user input: if the user remembers to include a particular variant of a target expression in his query, the query will return documents containing the variant; if the user does not think to include a particular variant, documents containing the variant will not be returned.

Recognizing this limitation, researchers have sought to put systems in place that will allow the automatic expansion of a user's initial input to the full range of variants that can express that same concept. The result is that the user need not identify in advance all the ways a particular concept can surface linguistically; the user needs only to provide sufficient indication of the target concept, and the system, relying on one approach or another will attempt to expand the query to include conceptually related patterns. Such automatically expanded searches are called "concept searches."

Underlying all varieties of concept search is an information retrieval model (be it a neural networks model, a vector space model, or a Bayesian networks model) that governs the method by which the user's input is expanded. The performance of the models underlying current versions of concept search has been well studied.

Studies of these models' performance have indeed found that such information retrieval approaches can achieve levels of performance superior to those achieved by standard manual Boolean systems. However, the studies have also found that the improvement in performance realized by these approaches is, marginal at best. 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. But this achievement is of dubious practical utility: the natural language system still found only about one-third of the relevant material, and it did so by pulling in a result set of which nearly half was not relevant. 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): (1) they require negotiation of a steep trade-off between recall and precision and (2) they leave the user ignorant of the actual performance of the system.

Challenges to be Addressed

Query-based approaches to document retrieval, whether involving standard manually created Boolean queries or some sort of natural language expansion of a query, are the approaches corporations and law firms most frequently turn to for their document management and

review needs. But, as shown above, these approaches perform poorly and fail to provide the user with information on just how poorly the system is performing.

If query-based approaches to document retrieval are not the answer to the document management challenges corporations and law firms face, what would an adequate answer look like? It is not the purpose of this paper to provide a complete description of a solution; for purposes of this paper, we simply note three key challenges that any document management and retrieval system will have to address if it is to be successful:

- Knowledge transfer between the expert, such as a senior litigator, and the reviewers (a teaching challenge)
- Consistency in reviewers' assessment of documents, both on an individual basis and among different reviewers (a mental discipline and coordination challenge)
- Valid information on the performance, in terms of precision and recall, of the system, whether manual, semi-automated, or automated (a quality assurance and process design challenge)

These challenges, by their very nature, are not addressed by an end-user search tool. Rather expertly designed combination of processes, technologies, and rigorous performance measurement is required.

Conclusion

The data cited above are evidence that the query-based methods discussed in this document will fail to meet the document management and review challenges corporations and law firms currently face. These methods can be expected to succeed in identifying 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. Moreover, the data are evidence that even when the system is performing poorly, users of the system will be dangerously unaware of that fact.

References

Blair, David C., and M. E. Maron. 1985. An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM* 28 (3): 289-299.

Dabney, Daniel P. 1986. The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval. *Law Library Journal* 78 (5): 5-40.

Hersh, William R., and Bikram Day. 1997. A Comparison of Boolean and Natural Language Searching for the TREC-6 Interactive Task. In *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6),* ed. E. M. Voorhees and D. K. Harman, 585-595. Gaithersburg, MD: NIST.

Hersh, William R., Susan Price, Dale Kraemer, Benjamin Chan, Lynetta Sacherek, and Daniel Olson. 1998. A Large-Scale Comparison of Boolean vs. Natural Language Searching for the TREC-7 Interactive Track. In *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7),* ed. E. M. Voorhees and D. K. Harman, 429-438. Gaithersburg, MD: NIST.

Leong, Mun-Kew. 1997. Concrete Queries in Specialized Domains: Known Item as Feedback for Query Formulation. In *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6),* ed. E. M. Voorhees and D. K. Harman, 541-549. Gaithersburg, MD: NIST.

Turtle, Howard. 1995. Text Retrieval in the Legal World. *Artificial Intelligence and the Law* 3 (1-2): 5-54.

Voorhees, Ellen M., and Donna Harman. 1997. Overview of the Sixth Text Retrieval Conference (TREC-6). In *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6),* ed. E. M. Voorhees and D. K. Harman, 1-24. Gaithersburg, MD: NIST.

Voorhees, Ellen M., and Donna Harman. 1998. Overview of the Seventh Text Retrieval Conference (TREC-7). In *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7),* ed. E. M. Voorhees and D. K. Harman, 1-24. Gaithersburg, MD: NIST.

H5 | New York | San Francisco | 415 625 6700 | info@h5.com | www.h5.com