IMPORTANT NOTICE:
This Publication Has Been Superseded

See the Most Current Publication at

https://thesedonaconference.org/publication/Commentary_on_Search_and_Retrieval_Methods

The Sedona Conference Journal®

Volume 8 ❖ Fall 2007

Articles

ESI Symposium

The Sedona Conference®

Antitrust Law, Complex Litigation and Intellectual Property Rights

THE SEDONA CONFERENCE® BEST PRACTICES COMMENTARY ON THE USE OF SEARCH AND INFORMATION RETRIEVAL METHODS IN E-DISCOVERY

*A Project of The Sedona Conference® Working Group on Best Practices for Document Retention and Production (WG1), Search & Retrieval Sciences Special Project Team**

August 2007 Public Comment Version

Editor-in-Chief:
Jason R. Baron

Executive Editors:
Richard G. Braman
Kenneth J. Withers

Senior Editors:
Thomas Y. Allman
M. James Daley
George L. Paul

* With valuable input from many other WG1 members and the RFP+ Vendor Panel. This document is for educational purposes only and is not a substitute for legal advice. The opinions expressed herein are consensus views of the editors and authors, and do not necessarily represent the views of any individual participants or authors or any of the organizations to which they belong or clients they represent, nor do they necessarily represent official views of The Sedona Conference®.

Table of Contents

Preface and Acknowledgements

Welcome to another major publication in The Sedona Conference Working Group Series (the "WGS"), *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*. This effort is an outgrowth of our Working Group on Electronic Document Retention and Production (WG1) and represents the work of its Search and Retrieval Sciences Special Project Team, consisting of a diverse group of lawyers and representatives of firms providing consulting and legal services to the legal tech community.

The mission of the Search and Retrieval Sciences Special Project Team has been to explore the nature of the search and retrieval process in the context of civil litigation and regulatory compliance in the digital age. The goal of this Best Practices Commentary is to provide the bench and bar with an educational guide to an area of e-discovery law that we believe will only become more important over time, given the need to accurately and efficiently search for relevant evidence contained within the exponentially increasing volumes of electronically stored information (ESI) that are stored and made subject to litigation, investigations, and regulatory activities. We also understand that the subject of what constitutes best practices in this area will necessarily be subject to change, given the accelerating pace of technological developments that the law is struggling to keep up with. We hope that our efforts will assist the legal profession in this area, and we welcome all feedback at tsc@sedona.net.

This Commentary was originally conceived at the Fourth Annual Meeting of WG1 in Vancouver, B.C., in the fall of 2005. Through the efforts of many individual contributors and editors, several successive drafts were prepared for comment by the full WG1 membership in successive midyear and annual meetings. I especially want to acknowledge the contributions to the overall success of this project made by Jason R Baron, who took the lead role in editing the Commentary, along with all of the special contributions of his fellow Co-chairs of the Search and Retrieval Sciences Team, M. James Daley and Ariana J. Tadler. I also wish to acknowledge the invaluable editorial assistance provided on one or more successive drafts by senior contributing editors Thomas Y. Allman, M. James Daley, and George L. Paul, as well as the drafting contributions provided along the way by Macyl Burke, Christopher Cotton, Matthew Cohen, Conor Crowley, Sherry Harris, William Herr, Joe Looby, Stephanie Mendelsohn, Dan Regard, Herbert Roitblat, Sonya Sigler, and Stephen Whetstone. Lastly, I wish to acknowledge that many other individuals in WG1, including on the Search and Retrieval Sciences Special Project Team and the RFP+ Vendor Panel, spent time in collaborating on earlier proposals for material to be included in the Commentary. On behalf of Richard Braman, Executive Director of The Sedona Conference, I wish to thank everyone involved in devoting their time and attention during the drafting and editing process.

Kenneth J. Withers
Director, Judicial Education and Content
The Sedona Conference
June 2007

Traditional Approaches To Searching For Relevant Evidence Are No Longer Practical Or Financially Feasible

Discovery of the relevant information gathered about a topic in dispute is at the core of the litigation process.[1] However, the advent of "e-discovery" is causing a rapid transformation in how that information is gathered. While discovery disputes are not new, the huge volume of available electronically stored information poses unique challenges. Just a few years ago, a party seeking to review information for production to the other side in a "large" document review case might have been concerned with hundreds of "banker's" boxes of documents.

Today, that same amount of data might be found on a single computer hard drive.[2] Moreover, as the ability to create and store massive volumes of electronic information mushrooms, the cost to store that information inversely plummets. In 1990, a typical gigabyte of storage cost about $20,000; today it costs less than $1 dollar. As a result, more individuals and companies are generating, receiving and storing more data, which means more information must be gathered, considered, reviewed and produced in litigation. But, with billable rates for junior associates at many law firms now starting at over $200 per hour, the cost to review just one gigabyte of data can easily exceed $30,000.[3] These economic realities – *i.e.*, the huge cost differential between the $1 to store a gigabyte of data and the $30,000 to review it – act as a driver in changing the traditional attitudes and approaches of lawyers, clients, courts and litigation support providers about how to search for relevant evidence during discovery and investigations. Escalating data volumes into the billions of ESI objects, review costs, and shrinking discovery timetables, all add up to equaling the need for profound change.

As discussed below in this Commentary, just as technology has given rise to these new litigation challenges, technology can help solve them, too. The emergence of new discovery strategies, best practices and processes, as well as new search and retrieval technologies, are transforming the way lawyers litigate and, collectively, offer real promise that huge volumes of information can be reviewed faster, more accurately, and more affordably than ever before. The good news is that search and retrieval systems are improving and expanding, buoyed by a huge economic wave of activity aimed at improving the "search" experience for users generally.[4] For example, advanced forms of search techniques, including various forms of fuzzy logic, text mining and machine learning all automatically organize electronically stored information in new ways not achieved by past more familiar methods, including the simple use of "keywords" as the only automated aid to conducting manual searches. Although we are at the dawn of a new era, these new techniques hold the potential to increase both accuracy and efficiency. Through statistical sampling and validation techniques we can then confirm the accuracy of the results of either traditional or alternative forms of search, retrieval, and review.

New challenges require new solutions. This Commentary aspires to serve as a guide to enable both the bench and the bar to become more familiar with the new challenges presented by needing to search and retrieve electronically stored information. The Commentary seeks to identify ways to address those challenges, and select the best solution to maximize the just, speedy, and inexpensive determination of every action, consistent with Federal Rule of Civil Procedure 1.

1 *Hickman v. Taylor*, 329 U. S. 495, 507 (1947)("Mutual knowledge of all the relevant facts gathered by both parties is essential to proper litigation").

2 Here's why: One gigabyte of electronic information can generate approximately 70,000-80,000 of text pages, or 35 to 40 banker's boxes of documents (at 2,000 pages per box). Thus, a 100-gigabtye storage device (*e.g.*, a personal computer hard drive), theoretically, could hold as much as the equivalent of 3,500 to 4,000 banker's boxes of documents. By contrast, in 1990, a typical personal computer held just 200 megabytes of data - 1/500 the capacity of a typical hard drive today. Even if only 10% of a computer's available capacity today contains useful or "useable" information (as distinguished from application programs, operating systems, utilities, etc.), attorneys still would need to consider and potentially review 700,000 to 800,000 pages per each device.

3 See Commentary, *infra*, n.13.

4 One indication of the amount of ongoing effort and investment generally to improve search and retrieval capabilities is evidenced by the research and development spending of internet giants Google, Yahoo!, and eBay. According to published reports, Google spent $ 1.23 billion, Yahoo! spent $883 million, and eBay spent $495 million on core research and development activities in 2006. *See* Robert Hertzberg, "I.T.'s Top 84 R&D Spenders," *Baseline* (April 17, 2007), www.baselinemag.com/article2/0,1540,2114821,00.asp.

Executive Summary

Discovery has changed. In just a few years, the review process needed to identify and produce information has evolved from one largely involving the manual review of paper documents to one involving vastly greater volumes of electronically stored information.

A perfect review of the resulting volume of information is not possible. Nor is it economic. The governing legal principles and best practices do not require perfection in making disclosures or in responding to discovery requests.

The Sedona Conference® has helped establish the benchmarks governing the evolution and refinement of reasonable, good faith practices for searching intimidating amounts of data. Principle 6 of The Sedona Principles, Second Edition (2007) notes that "[r]esponding parties are best situated to evaluate the procedures, methodologies and technologies appropriate for preserving and producing their own electronically stored information," and Principle 11 amplifies the point by stating that "[a] responding party may satisfy its good faith obligation to preserve and produce relevant electronically stored information by using electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information."

This Commentary discusses the existing and evolutionary methods by which a party may choose to search unprecedented volumes of information. As the practice of using these "search and retrieval" technologies – the generic term we will utilize in this Commentary – continues to advance, a new understanding will evolve about what is "reasonable" under the particular circumstances of those technologies. Thus, the challenges addressed by this Commentary go beyond litigation and encompass all aspects of the search and retrieval of information from large volumes of data.

The Revolution in Discovery

Just a few years ago all information was stored on physical records such as paper. There was typically only one original document, and the number of duplicative copies and their location was generally limited. Administrative assistants, file clerks, records managers and archivists developed expertise in managing the storage, generally pursuant to pre-existing file systems. It was reasonable, and indeed relatively easy in all but the exceptional case, for the legal profession to gather and then manually review all the individual items collected as part of the discovery process prior to their production.

But with the digital revolution there has also been a paradigm shift in the review process which is feasible. The shift of information storage to a digital realm has, for a variety of reasons, caused an explosion in the amount of information that resides in any enterprise-profoundly affecting litigation. This massive amount of electronically stored information is distributed broadly among different storage devices, from large mainframe computers, to tiny machines capable of storing information equivalent to several warehouses of documents each, all of which are or can be integrated into other systems. These systems are complex, interdependent, and evolve spontaneously, like ecosystems. It is often impossible to find one person, or even one discrete group of people, who completely understand the workings of this new form of "information ecosystem."

Finally, added to the search and retrieval challenge is the fact that a large percentage of the records being searched are expressed in *human language*, not just numbers. Human language is an inherently elastic, ambiguous "living" tool of enormous power. Its elasticity allows for private codes and vocabularies to exist in different subcultures in any enterprise, thus making the identification of the "words" to be searched much more challenging.

Essential Conclusions of this Commentary

This Sedona Conference® "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery" strives to set forth state-of-the-art knowledge about

meeting the challenge of searching enormous databases for relevant information, and then retrieving that information with a minimum of wasted effort.

By way of summary, we set forth our conclusions about the Problems and their Solutions, and summarize our Practical Advice which the balance of the paper articulates.

Problems

- Exponential growth in informational records is a critical challenge to the justice system.
- Electronically stored information contains human language, which challenges computer search tools. These challenges lie in the ambiguity inherent in human language and tendency of people within organizations or networks to invent their own words or communicate in code.
- The comparative efficacy of the results of manual review versus the results of alternative forms of automated methods of review remains very much an open matter of debate. Moreover, simple keyword searching, while itself a valuable tool, has certain known deficiencies.

Solutions

- Much that is useful in selecting information for production in discovery can be learned from other disciplines, including: information retrieval science; the study of linguistics; and implementation of effective management processes, to name just a few.
- Alternative search tools are available to supplement simple keyword searching and Boolean search techniques. These include using fuzzy logic to capture variations on words; using conceptual searching, which makes use of taxonomies and ontologies assembled by linguists; and using other machine learning and text mining tools that employ mathematical probabilities.
- It may be useful and appropriate to seek agreement on ways to measure and evaluate the effectiveness of the search and retrieval process. The metrics currently used in information science, such as "precision" and "recall," as well as more involved concepts are worth studying.

Practical Advice

Practice Point 1. *In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.*

Practice Point 2. *Success in using any automated search method or technology will be enhanced by a well-thought out process with substantial human input on the front end.*

Practice Point 3. *The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed.*

Practice Point 4. *Parties should perform due diligence in choosing a particular information retrieval product or service from a vendor.*

Practice Point 5. *The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.*

Practice Point 6. *Parties should make a good faith attempt to collaborate on the use of particular search and information retrieval methods, tools and protocols (including as to keywords, concepts, and other types of search parameters).*

Practice Point 7. *Parties should expect that their choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials).*

Practice Point 8. *Parties and the courts should be alert to new and evolving search and information retrieval methods.*

How The Legal Community Can Contribute to The Growth of Knowledge

A consensus is forming in the legal community that human review of documents in discovery is expensive, time consuming, and error-prone. There is growing consensus that the application of linguistic and mathematic-based content analysis, embodied in new forms of search and retrieval technologies, tools, techniques and process in support of the review function can effectively reduce litigation cost, time, and error rates.

Recommendations

1. *The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.*

2. *The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.*

Members of The Sedona Conference® community have and will continue to participate in collaborative workshops and other fora focused on issues involving information retrieval. The Sedona Conference® intends to remain in the forefront of the efforts of the legal community in seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships aimed at focused research.

I. INTRODUCTION

The exponential growth in the volume of electronically stored information or "ESI" found in modern enterprises poses a substantial challenge to the justice system. Today, even routine discovery requests can require searches of the storage devices found on mainframes, servers, networked workstations, desktops and laptops, home computers, removable media (such as CDs, DVDs and USB flash drives), and handheld devices (such as PDAs, cell phones and iPods). Complicating things, such information is now almost always flowing robustly throughout a "network," in which it has likely been replicated, distributed, modified, linked, attached, accessed, backed-up, overwritten, deleted, undeleted, fragmented, de-fragmented, morphed and multiplied. Discovery requests for e-mail, as one common example of ESI, often require searching and retrieving information from thousands to millions or even tens of millions of individual messages, with attachments in various file formats.

The volume and complexity of this electronically stored information highlights several issues: First, whether automated search and information retrieval methods are reliable and accurate? Second, whether the legal profession has developed the skills, know-how and processes to use such automated search and retrieval methods intelligently, when applied to huge data sets, in ways that are defensible under the rules governing discovery? Yet another issue is what impact, if any, the changes to the Federal Rules governing e-discovery will have on the search and retrieval process?

The Sedona Principles, Second Edition (2007) issued by The Sedona Conference® have endorsed several highly pragmatic and relevant consensus best practices relevant to this discussion.[5]

First, Principle 6 provides that responding parties are in the best position "to evaluate the procedures, methodologies, and technologies appropriate or preserving and producing their own electronically stored information." Principle 11 expands this concept to include the use of "electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonablylikely to contain relevant information."

Second, the Commentary to Principle 11 provides that the "selective use of keyword searches can be a reasonable approach when dealing with large amounts of electronic data," and goes on to state that it "is also possible to use technology to search for 'concepts,' which can be based on ontologies, taxonomies, or data clustering approaches, for example."[6] This exploits a unique feature of electronic information – the ability to conduct fast, iterative searches for the presence of patterns of words and concepts in large document populations. The Commentary to Principle 11 also states that "[c]ourts should encourage and promote the use of search and retrieval techniques in appropriate circumstances," and suggests that "[i]deally, the parties should agree on the search methods, including search terms or concepts, to be used as early as practicable. Such agreement should take account of the iterative nature of the discovery process and allow for refinement as the parties' understanding of the relevant issues develops."[7]

Third, the Sedona Conference® has recognized that "there are now hundreds of companies offering electronic discovery services."[8] This is also true of search and information retrieval products and services for use in legal contexts – which form a subset of a burgeoning sector of the economy devoted to improving users' "search" experience. However, there remains substantial confusion as to the strengths and weaknesses of such tools. Legal practitioners have a need for guidance as to the appropriate use of search and information retrieval technologies. Such guidance can help practitioners judge the relative costs and benefits of such tools in specific cases.

This Commentary is designed to help educate the justice system – attorneys, judges and litigants alike – about "state of the art" search and retrieval tools, techniques, and methodologies, and

---

5 *The Sedona Principles, Second Edition: Best Practices Recommendations & Principles for Addressing Electronic Document Production (The Sedona Conference® Working Group Series, 2007) ("The Sedona Principles, Second Edition, 2007"), available at* www.thesedonaconference.org.
6 *Id.*, Comment 11.a.
7 *Id.*
8 *The Sedona Conference® Best Practices for the Selection of Electronic Discovery Vendors: Navigating the Vendor Proposal Process* (2007), *available at* http://www.thesedonaconference.org/content/miscFiles/RFP_Paper.pdf.

how they can best be used as part of an overall process to more efficiently manage discovery. This discussion includes the critically important concept of an integrated process of search and retrieval; the ability to differentiate among different search methods; how to evaluate such differences; and what questions to ask before using any particular method or product in a specific legal setting.

The legal community is familiar with keyword and natural language searches on Westlaw® and Lexis® in the context of legal research, and to a lesser extent the use of "Boolean" logic to combine keywords and "operators" (such as "AND," "OR" and "AND NOT" or "BUT NOT") that produce broader or narrower searches. However, the use of keyword, Boolean, and other search and retrieval tools to narrow information to be reviewed for production in discovery is relatively recent.[9] Moreover, to date, the relative efficacy of competing search and retrieval tools used to accomplish production review simply have not been measured. The field is wide open for the development of search and information retrieval best practices that take into account various alternative search and retrieval methods. These methods extend from improvements in basic keyword searching, to more sophisticated systems that use mathematical algorithms and various forms of linguistic techniques to help find, group and present related content.

What follows is an in-depth analysis of the problems lawyers confront in managing massive amounts of data in discovery, including how search and retrieval techniques are used in everyday practice and the key element of "process." This Commentary also provides background on the field of information retrieval and describes the world of search tools, techniques and methodologies that are currently commercially available. It also includes a "practice pointers" guide on the factors to consider in making an overall legal evaluation among different search methods, both on a conceptual and practical level. In a concluding section, the future of search and retrieval efforts is discussed. A more technical discussion of various search methodologies is included in an Appendix. Where appropriate, reference will be made to technical definitions found in the updated Sedona Glossary.

II. THE SEARCH AND INFORMATION RETRIEVAL PROBLEM CONFRONTING LAWYERS

The discovery process of today is drowning in potential sources of information. The exponential increase in volume, especially since the mid-1990s, is principally due to the impact of the PC revolution, the widespread use of email and the growth of networks. Indeed, the implication of this growth in volume is that it places at severe risk the justice system's ability to achieve the "just, speedy and inexpensive" resolution of disputes, as contemplated by Rule 1 of the Federal Rules of Civil Procedure.

The Rise of a Crushing Volume of Information in the Digital Realm

A history of the computer and information technology advances occurring since the mid-1970s is beyond the scope of this Commentary. Suffice it to say that over the last 30 years, there has been a fast-paced and widespread shift from civilization's original physical information storage technologies to new, digital information storage technologies. This "digital realm" was created by an accretion of technological advances, each built on preceding advances, which together have resulted in as fundamental a shift in the way information is shared as that which occurred in 1450 when Johannes Guttenberg invented the printing press. Included among the advances contributing to the new "digital realm" are the invention of the microchip, the development and diffusion of the

9 There may be a role for use of some type of search and retrieval technology in discharging obligations to preserve ESI, as well as during the initial pre-review data culling or "collection" phase, in anticipation of complying with specific ESI and document requests. During the collection phase, for example, the goal is to maximize the amount of potentially relevant evidence in a subset of the greater universe of available ESI, without necessarily selecting only the more relevant information that might be the focus of a production phase review. Accordingly, parties may well end up using (and agreeing to use) differing search methods in the initial collection and later review phases of litigation. While we acknowledge that use of advanced search tools during earlier phases of litigation is truly cutting edge and worthy of future discussion, the primary focus of this Commentary will be on search tools as they are used in the review process. *See generally* Mia Mazza, Emmalena K. Quesada, and Ashley L. Sternberg, "In Pursuit of FRCP 1: Creative Approaches To Cutting and Shifting the Costs of Discovery of Electronically Stored Information," 13 RICHMOND J. LAW & TECHNOLOGY 11 (2007), at Paragraphs 53, 60, http://law.richmond.edu/jolt/v13i3/article11.pdf (discussing the use of concept searching in regard to preservation); *The Sedona Principles, Second Edition*, 2007, Comment 11.a ("Organizations should internally address search terms and other filtering criteria as soon as possible so that they can begin a dialogue on search methods as early as the initial discovery conference.").

personal computer, the spread of various types of networks linking together both computers and other networks, the rise of e-mail and its dominant use in the business world, the plunging cost of computing power and storage, and of course, the spread of the Internet and with it, the World Wide Web.[10]

By the mid-1990s, networked computers and their storage devices had created a true information-based society, with a constant flow of messages in all forms happening on a 24/7 basis. For example, studies reflect that the average U.S. worker sends and receives 100 e-mails per day. The size and nature of the attachments to these emails is also growing, with increased integration of image, audio and video files. Most recently, there has been a similar explosion in the use of instant messaging throughout business enterprises. In many organizations, the average worker maintains several gigabytes of stored data.[11] At the same time, the costs of storage have plummeted from $20,000 per gigabyte in 1990 to less than $ 1 per gigabyte today.[12] Existing technologies are only beginning to grapple with providing a viable automated means for applying records retention requirements, including the ability to implement legal holds, in the new ESI world.

Companies have continued to aggressively leverage technology to increase productivity. No one really controls how, where, how many times, and in how many forms information is stored. For example, the same Word documents can be found on e-mail attachments, local hard drives, network drives, document management systems, websites, and on all manner of removable media, such as USB flash drives, CDs, DVDs, and so on.

Discovery During the Recent Past: Manageable Amounts of Physically Stored Information

Historically, outside counsel played a key role in the discovery process, and the process worked simply. Litigants, assisted by their counsel, identified and collected information that was relevant to pending or foreseeable litigation. Counsel reviewed the information and produced any information that was relevant and not otherwise protected from disclosure by the attorney-client privilege, the attorney work product or by trade secret protections.

This worked fine in the days where most of the potentially relevant information had been created in or was stored in printed, physical form, and in reasonable volumes so that it required only "eyes" to review and interpret it. However, with increasingly complex computer networks, and the exponential increase in the volume of information existing in the digital realm, the venerated process of "eyes only" review has become neither workable nor economically feasible.

The cost of manual review of such volumes is prohibitive, often exceeding the damages at stake. Anecdotal reports indicate that the cost of reviewing information can easily exceed thousands of dollars per custodian, *per event*, for collection and attorney review. Litigants often cannot afford to review all available electronically stored information in the time permitted for discovery.[13] Moreover, efforts to reduce time and cost by use of "claw back"[14] provisions are problematic because of the risk of disclosure of sensitive proprietary and privileged information, as well as the risk of privilege waiver that can be imposed by substantive law, irrespective of new changes in procedural rules.

Accordingly, the conventional discovery review process is poorly adapted to much of today's litigation.[15] Lawyers of all stripes therefore have a vital interest in utilizing automated search and

---

10 *See* George L. Paul and Jason R. Baron, "Information Inflation: Can the Legal System Adapt?," 13 RICHMOND J. LAW & TECHNOLOGY 10 (2007), at Paragraph 1, n.2, http://law.richmond.edu/jolt/v13i3/article10.pdf ("Organizations have thousands if not tens of thousands of times as much information within their boundaries as they did 20 years ago."); Peter Lyman and Hal R. Varian, "How Much Information," 2003, http://www.sims.berkeley.edu/how-much-info-2003.

11 As noted *supra*, n.2, one gigabyte is equivalent in volume to between 70,000 and 80,000 pages of material. At 2000 pages per box, one gigabyte is therefore equivalent to 35 to 40 boxes of documents.

12 Michelle Kessler, "Days of officially drowning in data almost upon us," *USA Today*, Mar. 5, 2007, *available at* www.usatoday.com/tech/news/2007-03-05-data_N.htm.

13 Compare $1 to store a gigabyte of data with $32,000 to review it (*i.e.*, assuming one gigabyte equals 80,000 pages, and assuming that an associate billing $200 per hour can review 50 documents per hour at 10 pages in length, such a review would take 160 hours at $200/hr, or approximately $32,000).

14 "Clawback" and "quick peek" provisions in case management agreements seek to permit large productions of electronic data little or no review, and without waiver of any claim of privilege, work product, etc. *See The Sedona Principles, Second Edition, 2007*, Comment 10.d. *See also* amended Fed. R. Civ. P. 26(f)(4), effective December 1, 2006, and accompanying Committee Note.

15 Not all cases are equally heavy in involving electronic discovery and, from time to time, counsel may forgo production of electronically stored information and rely solely on hard copy documents.

retrieval tools where appropriate. The plaintiff's bar has a particular interest in being able to efficiently extract key information received in mammoth "document" productions, and in automated tools that facilitate the process. The defense bar has an obvious interest in reducing attendant costs, increasing efficiency, and in better risk-management of litigation (including reducing surprises). All lawyers, clients, and judges have an interest in maximizing the quality of discovery, by means of using automated tools that produce a reliable, reproducible and consistent product.

Ideally, then, judges and litigants should strive to increase their awareness of search and retrieval sciences generally, and of their appropriate application in discovery. Some technologies have been used for years to produce documents from large litigant document databases, but often without much critical analysis. The legal system may benefit from the rich body of research available through the information retrieval and library science disciplines. The discussion that follows is designed to provide a common framework and vocabulary for proper application of search and retrieval technologies in this new "age of information complexity" in the legal environment.

The Reigning Myth of "Perfect" Retrieval Using Traditional Means

It is not possible to discuss this issue without noting that there appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible – perhaps even perfect – and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual review of massive sets of electronic data sets (which it does not), the relative efficacy of that approach versus utilizing newly developed automated methods of review remains very much open to debate. Moreover, past research demonstrates the gap between lawyers' expectations and the true efficacy of certain types of searches. The Blair and Maron study (discussed below) reflects that human beings are less than 20% to 25% accurate and complete in searching and retrieving information from a heterogeneous set of documents (*i.e.,* in many data types and formats). The importance of this point cannot be overstated, as it provides a critical frame of reference in evaluating how new and enhanced forms of automated search methods and tools may yet be of benefit in litigation.

The Intelligent Use of Tools

Although the continued use of manual search and review methods may be indefensible in discovery involving significant amounts of electronically stored information, merely adopting sophisticated automated search tools, alone, will not necessarily lead to successful results. Lawyers must recognize that, just as important as utilizing the automated tools, is tuning the *process* in and by which a legal team uses such tools, including a close involvement of lead counsel. This may require an iterative process which importantly utilizes feedback and learning as tools, and allows for measurement of results. The time and effort spent on the front end designing a sophisticated discovery process that targets the real needs of the client must be viewed as a condition precedent to deploying automated methods of search and retrieval.

III. LAWYERS' CURRENT USE OF SEARCH AND RETRIEVAL METHODOLOGIES

Attorneys across all disciplines are generally familiar with search and retrieval methodologies based on their exposure over the past thirty years to using the automated means of searching provided by LexisNexis® and Westlaw® databases. More recently, lawyers have begun to use Google® and other Web-based search engines to hunt down information relevant to their practice. Additionally, law firms and corporate legal departments use search methods for administrative matters, such as searching data on available personnel, to support billing functions, to manage conflicts of interest, and for purposes of contact management. Many products employing search methods of various kinds exist in the legal marketplace to assist lawyers in these functions.

Current Database Tools in The Practice of Law

Litigators use automated search and retrieval tools at many stages of the litigation process. PACER and other automated means are used to uncover data on their opposing counsels' pleadings,

motions, and pretrial filings in similar litigation, as well as showing how a judge has ruled in similar issues even if unreported in legal reporting services. Lawyers also use a variety of search methods with online and CD-ROM databases to dig up facts on opposing parties, witnesses, and even jury pools. At later stages of litigation, lawyers use various litigation management software applications to search through potential exhibits in connection with proceedings held in "electronic courtrooms." But until recently, litigators seldom used automated search and retrieval methods with their clients' or their opponents' growing collections of unstructured ESI.

"De-duplication" in the Processing of ESI

With the exponential increase in the amount of data subject to e-discovery, lawyers have begun to take steps towards employing automated search tools to manage the discovery process. One example of this is "de-duplication" software used to find duplicate electronic files, since ESI often consists of a massively redundant universe. For example, the same email can be copied tens or even hundreds of times in different file locations on a network or on backup media. Such de-duplication software reduces the time attorneys must spend reviewing a large document set and helps to ensure consistent classification of documents for responsiveness or privilege.[16] Increasingly, "near de-duplication" tools also are being used to assist in organizing and expediting overall document reviews, even if the technique is not used to reduce the actual number of unique documents subject to review.[17]

The Use of "Keywords"

By far the most commonly used search methodology today is the use of "keyword searches" of full text and metadata as a means of filtering data for producing responsive documents in civil discovery. For the purpose of this commentary, the use of the term "keyword searches" refers to set-based searching using simple words or word combinations, with or without Boolean and related operators (see below and Appendix for definitions). The ability to perform keyword searches against large quantities of evidence has represented a significant advance in using automated technologies, as increasingly recognized by the courts. As one United States Magistrate Judge stated, "the glory of electronic information is not merely that it saves space but that it permits the computer to search for words or 'strings' of text in seconds."[18]

Courts have not only accepted, but in some cases have ordered, the use of keyword searching to define discovery parameters and resolve discovery disputes. One court has also suggested that a party might satisfy its duty to preserve documents in anticipation of litigation by conducting system-wide keyword searching and preserving a copy of each "'hit."[19]

Because of the costs and burdens (if not impossibility) of reviewing increasingly vast volumes of electronic data, it makes sense for producing parties to negotiate with requesting parties in advance to define the parameters of discoverable information. For example, parties could agree on

---

16 "De-duplication" services work to tag identical documents as duplicates by means of a "binary hash function" (which simply is a mathematical way of comparing the text of two documents -- represented in the underlying digital 1's and O's actually stored on the computer, to see if the documents are in fact perfectly alike). De-duplication by binary hash has been widely used without much notice in court opinions to date. *See Wiginton v. CB Richard Ellis, Inc.*, 229 F.R.D. 568, 571 (N.D. Ill. 2004) (referring to de-duplication process); *Medtronic Sofamor Danek Inc. v. Michelson*, 229 F.R.D. 550, 561 (W.D.Tenn., 2003) (same).

17 "Near de-duplication" involves files that "are not hash value duplicates but are materially similar." *See* http://www.law.com/jsp/legaltechnology/roadmapArticle.jsp?id=1158014995345&hubpage=Processing.

18 *In re Lorazepam & Clorazepate*, 300 F. Supp. 2d 43, 46 (D.D.C. 2004). *See also In re CV Therapeutics, Inc.*, 2006 WL 2458720 (N.D. Cal. Aug. 22, 2006) (court endorses employment of search terms as reasonable means of narrowing production); *J.C. Associates v. Fidelity & Guaranty Ins. Co.*, 2006 WL 1445173 (D.D.C. 2006) (requiring search of files using four specified keywords); *FTC v. Ameridebt, Inc.*, 2006 WL 6188563 (N.D. Cal. Mar. 13, 2006) ("e-mail could likely be screened efficiently through the use of electronic search terms that the parties agree upon"); *Windy City Innovations, LLC v. American Online, Inc.*, 2006 WL 2224057 (N.D. Ill. July 31, 2006) ("[k]eyword searching permits a party to search a document for a specific word more efficiently"); *Reino de Espana v. Am. Bureau of Shipping*, 2006 WL 3208579 (S.D.N.Y. Nov. 3, 2006) (court approves of e-keyword search for names and email addresses as a "targeted and focused discovery search"); *U.S. ex rel. Tyson v. Amerigroup Ill., Inc.*, 2005 WL 3111972 (N.D. Ill. Oct. 21, 2005) (referencing agreement by parties to search terms); *Medtronic Sofamor Danek, Inc., v. Michelson*, 229 F.R.D. 550 (W.D. Tenn. 2003) (court orders defendant to conduct searches using the keyword search terms provided by plaintiff); *Alexander v. FBI*, 194 F.R.D. 316 (D.D.C. 2000) (court places limitations on the scope of plaintiffs' proposed keywords to be used to search White House email).

19 *Zubulake v.UBS Warburg, LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004); *cf. Cache La Poudre Feeds, LLC v. Land O'Lakes, Inc.*, 2007 WL 684001 (D. Colo. Mar. 2, 2007) (where court denied motion for sanctions based on an allegation that the opposing party failed to properly monitor compliance with its discovery obligations by not conducting keyword searches, court also stated that *The Sedona Principles, 2004 Edition* and *Zubulake* were not to the contrary). *See also Zakre v. Norddeutsche Landesbank Girozentrale*, 2004 WL 764895 (S.D.N.Y. Apr. 9, 2004) (court denies plaintiff's request for additional indexing of e-records, holding that defendant's production of CD-ROMS in a text searchable form was sufficient, citing to Guideline 11 of *The Sedona Principles, 2004 Edition*).

conducting a search of only files maintained by relevant or key witnesses, and/or for certain date ranges. They often can also agree to a set of key words relevant to the case. Both sides can often see the advantage to using such protocols or filters to reduce the volume of extraneous information, such as spam, routine listserv notifications, and personal correspondence, which comes with the territory of searching through electronic realms.[20]

In *Treppel v. Biovail Corp*.,[21] the defendant refused to produce documents because the plaintiff would not agree to keyword search terms. Citing to Principle 11 of the *Sedona Principles for Electronic Document Production*, the court held that the defendant was justified in using keyword search terms to find responsive documents and should have proceeded unilaterally to use its list of terms when the plaintiff refused to endorse the list. The Court held that plaintiff's "recalcitrance" did not excuse defendant's failure to produce any records and ordered the company immediately to conduct the automated search, produce the results, and explain its search protocol. Another recent case emphasized the need to confer after plaintiff was successful in obtaining a "mirror image" of data on all of defendant's computers.[22]

Issues With Keywords

Keyword searches work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, although basic keyword searching techniques have been widely accepted both by courts and parties as sufficient to define the scope of their obligation to perform a search for responsive documents, the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[23]

Keyword searches identify all documents containing a specified term regardless of context, and so they can possibly capture many documents irrelevant to the user's query. For example, the term "strike" could be found in documents relating to a labor union tactic, a military action, options trading, or baseball, to name just a few (illustrating "polysemy," or *ambiguity* in the use of language). The problem of the relative percentage of "false positive" hits or noise in the data is potentially huge, amounting in some cases to huge numbers of files which must be searched to find responsive documents.[24]

On the other hand, keyword searches have the potential to miss documents that contain a word that has the same meaning as the term used in the query, but is not specified. For example, a user making queries about labor actions might miss an email referring to a "boycott" if that particular word was not included as a keyword, and a lawyer investigating tax fraud via options trading might miss an email referring to "exercise price" if that term was not specifically searched (illustrating

---

20 *See generally* Kenneth J. Withers, *Computer-Based Discovery in Federal Court Litigation*, 2000 FEDERAL COURTS L. REV. 2, http://www.fclr.org/articles/2000fedctslrev2.pdf (suggesting parties adopt collaborative strategies on search protocols); *see also* R. Brownstone, *Collaborative Navigation of the Stormy e-discovery Seas*, 10 RICHMOND J. LAW & TECHNOLOGY. 53 (2004), http://law.richmond.edu/jolt/v10i5/article53.pdf (arguing that parties must agree to search terms and other selection criteria to narrow the scope to manageable data sets); *see also The Sedona Principles, Second Edition, 2007*, Comment 11.a ("For example, use of search terms could reveal that a very low percentage of files (such as emails and attachments) on a data tape contain terms that are responsive to 'key' terms. This may weigh heavily against a need to further search that source, or it may be a factor in a cost-shifting analysis. Such techniques may also reveal substantial redundancy between sources (*i.e.*, duplicate data is found in both locations) such that it is reasonable for the organization to preserve and produce data from only one of the sources.").

21 233 F.R.D. 363 (S.D.N.Y. 2006).

22 *Balboa Threadworks v. Stucky*, 2006 WL 763668 (D. Kan. Mar. 24, 2006) (court orders parties to meet and confer on the use of a search protocol, including key word searching).

23 Some case law has held that keyword searches were either incomplete or overinclusive, *see Alexander v FBI, supra*, n.18; *Quinby v. WestLB*, AG, 2006 WL 2597900 (S.D.N.Y. Sept. 5, 2006) (court narrows party's demand for 170 proposed search terms in part due to the inclusion of commonly used words).

24 *See, e.g.*, G. Paul and J. Baron, "Information Inflation," *supra*, n.10, at Paragraph 20 (discussing potential time and cost of searching through 1 billion emails); Craig Ball, "Unlocking Keywords: How you frame your search words will shape your success," 14 No. 1 *Law Technology News* 56 (January 2007) (discussing how to improve keyword search results by use of various techniques, including eliminating "noise words" such as "law" and "legal"). *See also* Steven C. Bennett, "E-Discovery by Keyword Search," 15 No. 3 *Prac. Litigator 7* (2004).

"synonymy" or *variation* in the use of language). And of course, if authors of records are inventing words "on the fly," as they have done through history, and now are doing with increasing frequency in electronic communications, such problems are compounded.[25]

Keyword searches can also exclude common or inadvertently misspelled instances of the term (*e.g.*, "Phillip" for "Philip," or "strik" for "strike") or variations on "stems" of words (*e.g.* "striking"). So too, it is well known that even the best of optical character recognition (OCR) scanning processes introduce a certain rate of random error into document texts, potentially transforming would-be keywords into something else. Finally, using keywords alone results in a return set of potentially responsive documents that are not weighted and ranked based upon their potential importance or relevance. In other words, each document is considered to have an equal probability of being responsive upon further manual review.

More advanced keyword searches using "Boolean" operators and techniques borrowed from "fuzzy logic" may increase the number of relevant documents and decrease the number of irrelevant documents retrieved. These searches attempt to emulate the way humans use language to describe concepts. In essence, however, they simply translate ordinary words and phrases into a Boolean search argument. Thus, a natural language search for "all birds that live in Africa" is translated to something like ("bird* + liv* + Africa").

At the present time, it would appear that the majority of automated litigation support providers and software continue to rely on keyword searching. Such methods are limited by their dependence on matching a specific, sometimes arbitrary choice of language to describe the targeted topic of interest.[26] The issue of whether there is room for improvement in the rate of "recall" (as defined in the next section) of relevant documents in a given collection is something lawyers must consider when relying on simple and traditional input of keywords alone.

Use of Alternative Search Tools and Methods

Lawyers are beginning to feel more comfortable using alternative search tools to identify potentially relevant electronically stored information. These more advanced text mining tools include "conceptual search methods" which rely on semantic relations between words, and/or which use "thesauri" to capture documents that would be missed in keyword searching. Specific types of alternate search methods are set out in detail in the Appendix.

"Concept" search and retrieval technologies attempt to locate information that relates to a desired concept, without the presence of a particular word or phrase. The classic example is the concept search that will recognize that documents about Eskimos and igloos are related to Alaska, even if they do not specifically mention the word "Alaska." At least one reported case has referenced the possible use of "concept searching" as an alternative to strict reliance on keyword searching.[27]

Other automated tools rely on "taxonomies" and "ontologies" to help find documents conceptually related to the topic being searched, based on commercially available data or on specifically compiled information. This information is provided by attorneys or developed for the business function or specific industry (*e.g.*, the concept of "strike" in labor law *vs.* "strike" in options trading). These tools rely on the information that linguists collect from the lawyers and witnesses about the key factual issues in the case – the people, organization, and key concepts relating to the business as well as the idiosyncratic communications that might be lurking in documents, files, and emails. For example, a linguist would want to know how union organizers or company officials might

---

25 Philosophers use colorful imagery to describe the dynamism and complexity of human language. *See, e.g.* Ludwig Wittgenstein, THE PHILOSOPHICAL INVESTIGATIONS, Section 18 (G.E.M. Anscombe, trans., The Macmillan Co., 1953 ("[T]o imagine a language is to imagine a form of life…. [L]anguage can be seen as an ancient city; a maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new boroughs with straight regular streets and uniform houses").

26 See Part IV, *infra*; see *generally*, S.I. Hayakawa, LANGUAGE IN THOUGHT AND ACTION (Harcourt 1990) (5th ed.) (such methods are inherently limited by their specific choice of language to describe a specific object or reality).

27 *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, 2007 WL 1585452 (D.D.C. June 1, 2007) (citing to G. Paul and J. Baron, *supra*, n.10); *see generally* M. Mazza, E. Quesada, and A. Sternberg, "In Pursuit of FRCP 1: Creative Approaches To Cutting and Shifting the Costs of Discovery of Electronically Stored Information," *supra* n.9, at Paragraph 54 (discussing concept searching).

communicate plans, any special code words used in the industry, the relationships of collective bargaining units, company management structure, and other issues and concepts.

Another type of search tool relies on mathematical probabilities that a certain text is associated with a particular conceptual category. These types of machine learning tools, which include "clustering" and "latent semantic indexing," are arguably helpful in addressing cultural biases of taxonomies because they do not depend on linguistic analysis, but on mathematical probabilities. They can also help to find communications in code language and neologisms. For example, if the labor lawyer were searching for evidence that management was targeting neophytes in the union, she might miss the term "n00b" (a neologism for "newbie"). This technology, used in government intelligence, is particularly apt in helping lawyers find information when they don't know exactly what to look for. For example, when a lawyer is looking for evidence that key players conspired to violate the labor union laws, she will usually not know the "code words" or expressions the players may have used to disguise their communications.

Anecdotal information suggests that a small number of companies and law firms – particularly those that have gained significant experience in e-discovery – are using alternative search methods to either identify responsive documents (reducing expensive attorney review time) or to winnow collections to the key documents for depositions, pretrial pleadings, and trial.

The document databases that can assist lawyers in developing advanced ontologies and mathematical models are not limited to "discovery" documents. Search tools can be used in overall case management to search across pleadings, legal research, discovery responses, expert reports, and attorney work product. For example, in addition to searching discovery documents, a legal team in a labor dispute might want to search the interrogatory responses, pleadings, and depositions for all references to the concept of "strike." This is a potential growth area for vendors specializing in case management software.

Apart from the authorities listed in this section, there is still little by way of published reports or cases discussing or challenging the use of these various tools. It is only a matter of time, however, before more widespread deployment will lead to the development of a fuller body of case law.

Resistance by the Legal Profession

Some litigators continue to primarily rely upon manual review of information as part of their review process.[28] Principal rationales are: (1) concerns that computers cannot be programmed to replace the human intelligence required to make complex determinations on relevance and privilege; (2) the perception that there is a lack of scientific validity of search technologies necessary to defend against a court challenge; and (3) widespread lack of knowledge (and confusion) about the capabilities of automated search tools.

Other parties and litigators may accept simple keyword searching, yet be reluctant to use alternative search techniques. They may not be convinced that the chosen method would withstand a court challenge. They may perceive a risk that problem documents will not be found despite the additional effort; and an opposite risk that documents might be missed which would otherwise be picked up in a straight keyword search. Moreover, acknowledging that there is no one solution for all situations, they may opt for a tried-and-true lowest common denominator. Finally, litigators lack the time and resources to sort out these highly complex technical issues on a case-by-case basis.[29]

---

28 *But see In re Instinet Group, Inc.*, 2005 WL 3501708 (Del. Ch. Dec. 14, 2005). The court reduced plaintiffs' attorneys' fee claim by $1 million (75% of the total claim) for "obvious" inefficiencies in plaintiffs' counsel's review of paper printouts ("blowbacks") from digital files. The court stated that plaintiffs' counsel's decision to "blow back" the digital documents to paper "both added unnecessary expense and greatly increased the number of hours required to search and review the document production."

29 *See, e.g.,* R. Friedmann, http://prismlegal.com/wordpress/?cat=9 (Feb. 4, 2005)(suggesting that not one solution fits all cases); *see also id.* (July 30, 2003) (questioning the incremental value of sophisticated searching over simple searching because of the costs of implementation and need to build taxonomies and to test methodologies).

Challenging the Choice of Search Method

The challenge to a choice of search methodology used in a review prior to production can arise in one of two contexts: (1) a requesting party's objection to the unilateral use of a search method by a responding party; or (2) a court's *sua sponte* review of the use of a method or technology. Accordingly, the preferable method to reduce challenges – advocated by the proponents of the 2006 Federal Rules Amendments and experienced practitioners – is for a full and transparent discussion among counsel of the search terminology. Where the parties are in agreement on the method and a reasonable explanation can be provided, it is unlikely that a court will second-guess the process.

Absence agreement, a party has the presumption, under Sedona Principle 6, that it is in the best position to choose an appropriate method of searching and culling data. However, a unilateral choice of a search methodology may be challenged due to lack of a scientific showing that the results are accurate, complete and reliable. Since all automated search tools rely on some level of science, the challenging party may argue that the process used by the responding party is essentially an expert technology which has not been validated by subjecting it to peer review, and unbiased empirical testing or analysis.

The probability of such a challenge is greater if the technology is patented or proprietary to a developer or vendor (*i.e.*, in a so-called "Black Box"). In such circumstances, e-discovery and litigation support vendors that use these technologies may be several degrees of separation from the original developers. A requesting party may demand the responding party to "prove up" the use of such search technology. This could set the stage for a difficult and expensive battle of experts.

As a practical matter, however, those who might object to a particular search and retrieval technology face several challenges. First, the legal system has, for decades blessed the use of keyword search tools and databases for discovery review. Second, even if such a challenge were permitted to proceed, the lack of a formally acknowledged baseline by which to measure the comparative accuracy and reliability of any search method precludes a comparison of the "new" method to traditional methods. And third, if human review or even keyword searching is the benchmark for accuracy and reliability, it arguably should not be difficult to compare the new technology favorably with either keyword searching or human review, especially when guided by a reasonable process. The discovery standard is, after all, reasonableness, not perfection.

Given the continued exponential growth in information, we would expect that a body of precedent will develop over time which references, if not critically analyzes, new and alternative search methods in use in particular legal contexts.

IV. SOME KEY TERMS, CONCEPTS AND HISTORY IN INFORMATION RETRIEVAL TECHNOLOGY

The evaluation of information retrieval ("IR") systems has, until now, largely been of greatest interest to computer scientists and graduate students in information and library science. Unlike performance benchmarking for computer hardware, there are no agreed-upon objective criteria for evaluating the performance of information retrieval systems. That is, for IR systems, the notion of effectiveness is subjective. Human judgment is ultimately the criteria for evaluating whether an IR system returns the relevant information in the correct manner. Two users may have differing needs when using an IR system. For example, one may want to find all potentially relevant documents. Another may want to correctly sort information by priority. Additionally, the subject matter and information type impact a user's information retrieval requirements.

Over the past 50 years, a large body of research has emerged concerning the evaluation of IR systems. The study of IR metrics helps quantify and compare the benefits of various search and information retrieval systems. In 1966, C.W. Cleverdon listed various "metrics" which have become

the standard for evaluating IR systems within what has become known as the "Cranfield tradition."[30] Two of the metrics, *precision* and *recall*, are based on binary relationships. That is, either a document is relevant or it is not, and either a document is retrieved or it is not. Several modifications and additional metrics have been added in the IR literature since then, as the scientific field continues to add and refine techniques for measuring the efficiency of IR systems – both in terms of retrieval and also in user access to relevant information.

Measuring the effectiveness of information retrieval methods

Recall, by definition, is "an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved."[31] Another way to think about it is: out of the total number of relevant documents in the document collection, how many were retrieved correctly?

Precision is defined as "an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant."[32] Put another way, how much of the returned result set is on target?

Recall and precision can be expressed by simple ratios:

$$\textbf{Recall} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of responsive documents overall}}$$

$$\textbf{Precision} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of documents retrieved}}$$

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.

Importantly for the practitioner, 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.

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.[33]

---

30 *See* Cyril W. Cleverdon et al., ASLIB CRANFIELD RESEARCH PROJECT: FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS (1966) (Vol. I , Design), *available at* http://www-nlpir.nist.gov/projects/irlib/pubs/cranv1p1/cranv1p1_index/cranv1p1_toc.html; Cyril W. Cleverdon et al., ASLIB CRANFIELD RESEARCH PROJECT: REPORT OF CRANFIELD II (1966) (Vol. II, Test Results), *available at* http://www-nlpir.nist.gov/projects/irlib/pubs/cranv2/cranv2_index/cranv2_toc.html; *see generally*, C.J. VAN RIJISBERGEN, INFORMATION RETRIEVAL (2d ed. 1979), *available at* http://www.dcs.gla.ac.uk/Keith/Preface.html.

31 *See* Ricardo Baeza-Yates & Berthier Ribeiro-Neto, MODERN INFORMATION RETRIEVAL 437-455 (1999) (glossary), *available at* http://www.sims.berkeley.edu/~hearst/irbook/glossary.html.

32 *Ibid*.

33 There are many other common metrics that are considered in IR literature, including f-measure, mean average precision and average search length. F-measure is an approximation of the cross-over point between precision and recall, which allows one to see where the compromise is between the two. Mean average precision determines the existing precision level for each retrieved relevant item. Average search length is the average position of a relevant retrieved item. Still other terms include "fallout," the ratio of the number of non-relevant items retrieved to the total number of items retrieved," and "elusion," the proportion of responsive documents that have been missed.

Measuring the Efficiency of Information Retrieval Methods

Efficiency is important to the usability of an IR system, but it does not affect the quality of the results. Efficiency is measured in two ways. The first measurement is the mean time for returning search results. This can be measured by average time to return the results or the computational complexity of the search. The second measurement is the mean time it takes a user to complete a search. This measurement is more subjective and is a function of the usability of the IR system.

The Blair and Maron Study

The leading study testing recall and precision in a legal setting was conducted by David Blair and M.E. Maron in 1985.[34] It is a classic in showing the problem caused by the rich use of human language among the many people that can be involved in a dispute, and how difficult it is to take such richness into account in a search for informational records.

Indeed, Blair and Maron found that attorneys were only about 20% effective at thinking up all of the different ways that document authors could refer to words, ideas, or issues in their case.

The case involved a San Francisco Bay Area Rapid Transit (BART) accident in which a computerized BART train failed to stop at the end of the line. There were about 40,000 documents totaling about 350,000 pages in the discovery database. 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 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 a "disaster." Other documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were often not mentioned.

As Roitblat notes, *supra*, n.34, Blair and Maron even 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."

The Impact of Ambiguity and Variation on Precision and Recall

Since the Blair and Maron study, some further efforts have been made to study the precision/recall issues in a legal discovery context, some of which have been performed by members of The Sedona Conference®.[35] This field requires further study.

The limitation on search and retrieval methodology exposed in the Blair and Maron study was not the ability of the computer to find documents that met the attorneys' search criteria, but rather the inability of the attorneys and paralegals to anticipate all of the possible ways that people could refer to the issues in the case. The richness of human language causes a severe challenge in identifying informational records.

Ambiguity refers to the tendency of words and expressions to have different meanings when used in different contexts. These contexts are "referential variants" or *variation*. If one and only one word or expression is found in only one and only one context, it would present no ambiguity and no

---

34 David C. Blair & M.E. Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM* 289 (1985). The discussion that follows of the Blair and Maron study is drawn directly from Herbert L. Roitblat, "Search and Information Retrieval Science," 8 *Sedona Conf. J. at* 225 (2007).

35 *See, e.g.*, Anne Kershaw, "Automated Document Review Proves Its Reliability," DIGITAL DISCOVERY & E-EVIDENCE, Nov. 2005, at 10, 10-12 (client-sponsored private study); Howard Turtle, "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," 1994 PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 212-220 (using structured caselaw in Westlaw databases); *see also* Text REtrieval Conference, http://trec.nist.gov/, discussed *infra* Part VII.C.

variation. A search for that term would retrieve all of the documents in which the term appeared, and all of the documents would be relevant. While there may not be an exact mathematical comparison, generally speaking, the lower the variation in the contexts, the lower the likely overall recall, and the lower the ambiguity of the search term, the better the precision of the result.

But as the Blair and Maron study demonstrates, human language is highly ambiguous and full of variation. In the years since Blair and Maron, the IR community has been engaged in research and development of methods, tools, and techniques that compensate for endemic ambiguity and variation in human language, and thus maximize the recall and precision of searches.

V. BOOLEAN AND BEYOND: A WORLD OF SEARCH METHODS, TOOLS AND TECHNIQUES

In the twenty years since the Blair and Maron study, a variety of new search tools and techniques have been introduced to help find relevant information and to help weed out irrelevant information. Understanding these various tools and methods is critical. All automated methods are not created equal, and do not perform the same function and task. It is important to know what each methodology does when it is used alone or in conjunction with other methodologies.

Clearly, different search methods have different functions and values in different circumstances. There is no one best system for all situations, a key fact for practitioners learning the technique of search and retrieval technology.

A more detailed description of search methods and techniques is set out in the Appendix. These methods can be grouped into three broad categories, but there are hybrid and cross-cutting approaches that defy easy placement in any particular "box."

Keywords and Boolean Operators

First, there are *keyword based methods*, ranging from the simple use of keywords alone, to the use of strings of keywords with what are known as "Boolean operators" (including AND, OR, "AND NOT" or "BUT NOT").

Statistical Techniques

Second, there are a variety of *statistical techniques*, which analyze word counts (how many times the same keyword will appear in a document, or will appear near other keywords). One such approach is called "Bayesian," derived from a famous mathematical theorem. Querying the data set using combinations of one or more of these types of Bayesian methods may well result in returning a broader slice of the data than merely using a simple keyword search, or a keyword search with Boolean operators.

Categorizations of Data Sets

Third, there are other techniques depending on *categorizations of the entire data* set with various methodologies heavily reliant on setting up (*i.e.*, coming to a consensus on) a *thesaurus, taxonomy* or "*ontology*" of related words or terms. These techniques can be used to categorize the entire data set into specified categories all at once – or continually, as more data is added to the data set.

However, data sets generally need to be indexed to use any of the latter alternative methodologies – where the indexing will take more time depending on what one indexes (*e.g.*, indexing all of the data will take substantially longer than indexing selected coded fields).

There are a variety of indexing tools, some of which are available as open source tools. Indexing structured data may take less time than indexing data in an unstructured form. Indexing a set number of structured fields (*i.e.* coded data) will be much faster because only those designated

fields are indexed. Indexing an unstructured data set is time consuming because of the need to index all the *words* (except for and, a, the, or other common words). Knowing what is being indexed will be important to set expectations in terms of timing and making the data useful for querying or review.

Alternative search methods to keywords can, in some instances, free the user from having to guess, for every document, what word the author might have used. For example, there are more than 120 words that could be used in place of the word "think" (*e.g.*, guess, surmise, anticipate). As the Blair and Maron study shows, people coming in after the fact are actually very poor at guessing the right words to use in a search – words that find the documents a person is looking for without overwhelming the retrieval with irrelevant documents. In light of this fact, alternative search methods may serve to help to organize large collections of documents in ways that people have trouble doing.

Using a thesaurus, taxonomy, or ontology generally gives the results one would expect, because these systems explicitly incorporate one's expectations about what is related to what. They are most useful when one has (or can buy) a good idea of the conceptual relations to be found in one's documents – or one has the time and resources needed to develop them. Clustering, Bayesian classifiers, and other types of systems have the power to discover relationships in the text that might not have been anticipated. This means that one gets unexpected results from time to time, which can be of great value, but can also be somewhat disconcerting (or even wrong). An example: after training on a collection of medical documents, one of these systems learned that Elavil and Klonopin were related (they are both anti-anxiety drugs). A search for Elavil turned up all the documents that contained that word, along with documents containing the word "Klonopin" even without the word "Elavil."

Such systems can discover the meaning of at least some acronyms, jargon, and code words appropriate to the context of the specific document collection. No one has to anticipate their usage in all possible relational contexts; the systems, however, can go help to derive them directly from the documents processed.

Finally, none of these systems is magical. Language is sometimes shared just between two people, who have invented a shorthand or code. All tools require common sense, based on a thought-out approach. Some techniques may be difficult to understand to those without technical backgrounds, but they need not be mysterious. If a vendor will not explain how a system works, it is most likely because of ignorance. Ask for someone who can provide an explanation.

There is no magic to the science of search and retrieval: only mathematics, linguistics, and hard work. If lawyers do not become conversant in this area, they risk surrendering the intellectual jurisdiction to other fields.

VI. PRACTICAL GUIDANCE IN EVALUATING THE USE OF AUTOMATED SEARCH AND RETRIEVAL METHODS

Practice Point 1. In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.

For the reasons articulated in prior sections, the demands placed on practitioners and parties in litigation and elsewhere increasingly dictate that serious consideration be given to the use of automated search and retrieval methods in a wide variety of cases and contexts. Particularly, but not exclusively, in large and complex litigation, where discovery is expected to encompass hundreds of thousands to hundreds of millions of potentially responsive electronic records, there is no reasonable possibility of marshalling the human labor involved in undertaking a document by document, manual review of the potential universe of discoverable materials. This is increasingly true both for parties responding to a discovery request, and for parties who propound discovery only to

receive a massive amount of material in response. Where the infeasibility of undertaking manual review is acknowledged, utilizing automated search methods may not only be reasonable and valuable, but necessary.

Even in less complex settings, sole reliance on manual review may nevertheless be an inefficient use of scarce resources. This is especially the case where automated search tools used on the front end of discovery may prove to be useful in a variety of ways, including for sampling, categorizing or grouping documents in order to facilitate later manual review.

Of course, the use of automated search methods is not intended to be mutually exclusive with manual review; indeed, in many cases, both automated and manual searches will be conducted: with initial searches by automated means to cull down a large universe of material to more manageable size, followed by a secondary manual review process. So too, while automated search methods may be used to find privileged documents out of a larger set, it remains the case that the majority of practitioners still will rely on manual review processes to identify the bases for privilege to be asserted for each document.

Practice Point 2. Success in using any automated search method or technology will be enhanced by a well-thought out process with substantial human input on the front end.

As discussed above, the decision to employ an automated search method or technology cannot be made in a vacuum, on the assumption that the latest "tool" will solve a discovery obligation. Rather, to maximize the chances of success in terms of finding responsive documents, a well-thought out strategy capitalizing on "human knowledge" available to a party should be put into action at the earliest opportunity. This knowledge can take many forms.

First, an evaluation of the legal setting a party finds itself in is of paramount importance, since the nature of the lawsuit or investigation, the field of law involved, and the specific causes of action under which a discovery obligation arises must all be taken into account. For example, keyword searches alone in highly technical patent cases may prove highly efficacious. In other types of cases, including those with broad causes of action and involving subjective states of intent, a practitioner should consider alternative search methods.

Second, in any legal setting involving consideration of automated methods for conducting searches, counsel and client should perform a "relevance needs analysis," to first define the target universe of documents that is central to the relevant causes of action. This would include not only assessing relevant subject areas, and "drilling down" with as much specificity as possible, but also analyzing the parties who would be the "owners" of relevant data. Time and cost considerations must also be factored in, including budgeting for human review time. These practice points apply whether your client is a defendant and holds a universe of potentially discoverable data, or your client is a plaintiff party who is expecting to receive similarly massive data in response to requests for documents.

Practice Point 3. The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed.

The choice of a search and retrieval method for a given situation depends upon a number of factors.

For example, a search method that eliminates false positive "noise" (achieving high levels of precision) may not yield the highest number of relevant documents. In other cases, such as sampling, a search method will be graded on its ability to measure statistical significance of the occurrence of a particular word or concept. There are a number of overarching factors that lawyers should consider in evaluating the use of particular search and retrieval methods in particular settings.

First, the "heterogeneity" of the overall relevant universe of electronically stored information is a significant factor. Electronically stored information that is potentially relevant may be found in

multiple locations and in a variety of forms, including structured and unstructured active computer environments, removable media, backup tapes, and the variety of email applications and file formats. In some cases, information that provides historical, contextual, tracking or managerial insight (such as metadata) may be relevant to a specific matter and demand specialized data mining search tools. Yet in other cases, it will be irrelevant.

Next, the volume and condition of the electronically stored information, and the extent to which electronically stored information is contained within static or dynamic electronic applications is relevant to the decisions made by the advocate or investigator.

Third, the time it will take to use a particular search and information retrieval method and its cost, as compared to other automated methods or human review, must be considered.

Fourth, the goals of the search are a factor (*e.g.* capturing or finding as many responsive documents as possible regardless of time and cost vs. finding responsive documents as efficiently as possible, *i.e.*, with the least number of nonresponsive documents). In other words, one must consider the desired trade off between recall and precision. Given the particular setting, the party seeking to employ one or more search methods should assess the relative importance in that setting of finding responsive electronically stored information versus the importance of eliminating non-responsive data. Depending on this assessment, one or more alternative search methodologies may prove to be a better match in the context of a particular task.

Fifth, one must consider the skills, experience, financial and practical logistical constraints of the representatives of the party making the selection (attorneys, litigation support staff, vendors).

Sixth, there is the status of electronic discovery in the matter, including the extent to which activities including preservation and collection are occurring in addition to processing and/or attorney review.

Seventh, one must investigate published papers supporting the reliability of the search and information retrieval method for particular types of data, or in particular settings.

Practice Point 4. Parties should perform due diligence in choosing a particular information retrieval product or service from a vendor.

The prudent practitioner should ask questions regarding search and retrieval features and the specific processing and searching rules that are applied to such features. Some tools are fully integrated into a vendor's search and review system, whereas others are "stand alone" tools that may be used separately from the review platform. It is essential not only to understand how the various tools function, but also to understand how the tools fit within the overall workflow planned for discovery. A practitioner should inquire as to what category or categories the specific tool fits into, how it functions, and what third party technology lies behind the tool.

It is also essential that specific methods or tools be made understandable to the court, opposing parties, and your own client. How data is captured and indexed (and how long it takes to build an index) also may affect a decision on use: it is therefore important to understand how a particular system deals with rolling input and output over time, in terms of its flexibility. The ability to perform searches across metadata, to search across multiple indices or stores of data, to search embedded data, to refine search results (nested searches), to save queries, to capture duplicates and perform de-duplication, to trace email threads, and to provide listings of related terms or synonyms, are all examples of the kind of specific functional requirements that should be inquired about.

Other types of due diligence inquiries may involve administrative matters (*e.g.*, understanding maintenance and upkeep, additional charges, system upgrades, availability of technicians, system performance), quality control issues (*e.g.*, prior testing of the method or tool in question; how databases and dictionaries supporting concept searching were populated; how strong is

the application development group of the provider), and, finally, any relevant licensing issues, involving proprietary software or escrow agreements with third parties.

Practice Point 5. The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.

Just as with past practice involving manual searches through traditional paper document collections, there is no requirement that "perfect" searches will occur – only that lawyers and parties act reasonably in the good faith performance of their discovery and legal obligations. From decades of information retrieval research, we know that a 100% rate of recall, *i.e.*, the ability to retrieve *all* responsive documents from a given universe of electronic data, is an unachievable goal. As discussed in prior sections, the richness of human language, with its attendant elasticity, results in all present day automated search methods falling short.

It is also important to recognize that there will be a measure of statistical variation associated with alternative search methods, *i.e.*, some responsive documents will be found by one search method while being missed by others. Even the same search method (such as one based on statistical properties of how words appear in the data set), may return different results if new documents are added to the searched universe.

Particularly in the context of a large data set, a search method should be judged by its overall results (such as using average measures of recall and precision), rather than being judged by whether it produces the identical document set as compared with a different technique. One possible benchmark to employ when considering use of an alternative search method is to compare the results of such a search against a similar search utilizing keywords and Boolean operators alone.

However, it is important not to compare "apples with oranges." Given the present state of information science, it would be a mistake to assume that one search method will work optimally across all types of possible inquiries or data sets (*e.g.*, what works well in finding word processing documents in a given proprietary format may not be as optimal for finding information in structured databases, or in a collection of scanned images). This is another area where, consistent with the above principles, a good deal of thought should be given at the outset to the precise problem, in terms of its scope and relevancy considerations, before committing to a particular search method.

Practice Point 6. Parties should make a good faith attempt to collaborate on the use of particular search and information retrieval methods, tools and protocols (including as to keywords, concepts, and other types of search parameters)

The *Treppel* decision and other recent case law indicates that courts are becoming more comfortable with addressing search and retrieval issues, particularly in the context of blessing or ordering parties to share information that would lead to the development of more refined search protocols. The fact that some courts have waded into these issues demonstrates how rapidly the law has been evolving even in advance of the 2006 amendments to the Federal Rules of Civil Procedure.[36]

Under newly modified Rule 26(f), the parties' initial planning is expected to address "[a]ny issues relating to disclosure or discovery of electronically stored information," as well as "[a]ny issues relating to preserving discoverable information." These initial discussions on preservation and production easily should encompass a specific discussion on search methods and protocols to be employed by one or both parties. While disclosure of these methods and protocols is not mandated or legally required under this rule, the advantages of collaborating should strongly be considered.

---

36 *See* Kenneth J. Withers, "The December 2006 Amendments to the Federal Rules of Civil Procedure," 4 NW. J. OF TECH. & INTELL. PROP. 171 (2006), *available at* http://www.law.northwestern.edu/journals/njtip/v4/n2/3 (what "probably strikes the reader [of *Treppel*] as matter-of-fact, sensible, and routine, would have been extraordinary a scant six years ago, when the last major revision of the discovery rules went into effect [in 2000])."

Reaching an early consensus on the scope of searches has the potential to minimize the overall time, cost, and resources spent on such efforts, as well as minimizing the risk of collateral litigation challenging the reasonableness of the search method employed.[37]

Practice Point 7. Parties should expect that their choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials).

Counsel should be prepared to explain what keywords, search protocols, and alternative search methods were used to generate a set of documents, including ones made subject to subsequent manual searches for responsiveness and privilege. This explanation may best come from a technical "IT" expert, a statistician, or an expert in search and retrieval technology. Counsel must be prepared to answer questions, and indeed, to prove the reasonableness and good faith of their methods.

Practice Point 8. Parties and the courts should be alert to new and evolving search and information retrieval methods.

What constitutes a reasonable search and information retrieval method is subject to change, given the rapid evolution of technology. The legal community needs to be vigilant in examining new and emerging techniques and methods which claim to yield better search results. In particular settings, lawyers should endeavor to incorporate evolving technological progress at the earliest opportunity in the planning stages of discovery or other legal setting involving search and retrieval issues.

VII. FUTURE DIRECTIONS IN SEARCH AND RETRIEVAL SCIENCE

What prospects exist for improving present day search and retrieval methodologies? And how can lawyers play a greater role in working with the information retrieval research community based on a shared interest in how to improve the accuracy and efficiency of information retrieval?

A. Harnessing the Power of Artificial Intelligence (AI)

A statement from page 36 of The Sedona Conference®, Navigating The Vendor Proposal Process (2007 ed.), under the general heading "Advanced Search and Retrieval Technology," bears repetition here: "Technology is developing that will allow for electronic relevancy assessments and subject matter, or issue coding. These technologies have the potential to dramatically change the way electronic discovery is handled in litigation, and could save litigants millions of dollars in document review costs. Hand-in-hand with electronic relevancy assessment and issue coding, it is anticipated that advanced searching and retrieval technologies may allow for targeted collections and productions, thus reducing the volume of information involved in the discovery process."

The growing enormity of data stores, the inherent elasticity of human language, and the unfulfilled goal of computational thinking to approximate the ability and subtlety of human language behavior all present steep challenges to the AI community in developing optimal search and retrieval techniques.

But the future continues to hold promise. Not only is there the possibility of applying sophisticated artificial intelligence means to data mining of traditional texts, but looming immediately on the horizon are new and better approaches to image and voice pattern recognition. Clearly, all forms of data stored in corporations and institutions will be fair game in terms of being within the scope of future information demands in legal settings.

Finding information on the Web sometimes is easier than finding documents on one's own hard drive. The post-Google burst of interest in building better search engines for the Web can only

37 *See* G. Paul and J. Baron, *Information Inflation, supra* n.10, at Paragraphs 50-55 (discussing an iterative collaboration process that includes adoption of multiple "meet and confers" to discuss and refine preliminary search results).

help lead to new and better search techniques applied in more well-defined contexts, such as corporate and institutional intranets and data stores.

A recent "2020 Science" report issued by Microsoft anticipates the near-term development of "novel data mining technologies and novel analysis techniques," including "active learning" in the form of "autonomous experimentation" and "artificial scientists," in replacement of "'traditional' machine learning techniques [that] have failed to bring back the knowledge out of the data."[38] Beyond the short-term horizon, scientists are expected to embrace emergent technologies including the use of genetic algorithms, nanotechnology, quantum computing, and a host of other advanced means of information processing. The field of future AI research in the specific domain of search and retrieval is wide open.

B. The Role of Process in the Search and Retrieval Challenge

Every search and retrieval technology has its own methodology to ensure the technology works properly – a set of instructions outlining the workflow for the tool. How well these methods are applied significantly impacts the performance, and therefore, the results generated by the technology. This is where process comes in. Process functions to provide order and structure by setting guidelines and procedures designed to ensure that a technology performs as intended. Effectively applied, process will then drive the consistent and predictable application of the search and retrieval technology. The results derived from the consistent and predictable application of search and retrieval tools will then establish the technology's credibility and value.

The Important Nature of Process

A process is a considered series of events, acts or operations leading to a result or an effect. A process, like a technology, is a "tool" that can be used to assist in completing a task. The use of a well-defined and controlled process promotes consistency, reliability and predictability of the results and ensures the efficient use of the resources required to produce them. As such, a process does not find the answer to, or attain the objective of a task on its own. Process, no matter how well designed and executed can not replace the exercise of judgment, however, process promotes the exercise of judgment by ensuring that the most accurate and reliable information is available when making decisions. In the search and retrieval context, this means the availability of consistent and reliable information to assist parties in making informed decisions.

The use of process promotes consistency by establishing a defined approach to a task. The resulting consistency promotes reliability and predictability. Reliability and predictability allow for better planning, performance and cost management. All together, risk is reduced and confidence is promoted.

Search and retrieval should be visualized as a process which enables a party to distinguish potentially discoverable information from among a broader set of electronic data for purposes of production. It consists of several process steps that take place in the context of a particular search and retrieval technology. Because the application of process is flexible, it can be used to address unique conditions that might be associated with a technology, such as where the use of a search and retrieval technology itself creates issues. For example, the use of search and retrieval technologies to address significant volumes of information may not address all problems: as review volumes increase, even with carefully crafted and tested search criteria, the likelihood of being swamped by false positives increases greatly. Additionally, greater volume increases the likelihood of the omission of some relevant documents. By developing and implementing process steps that consistently address these issues, their impact can be diminished and the reasonableness and good faith of the technology can be established.

38 *See* http://research.microsoft.com/towards2020science/downloads.htm, p 15.

"Process" as a Measure of Reasonableness and Good Faith

Search and retrieval in this new era requires the establishment and recognition of a new standard. A standard of absolute perfection is and always has been unrealistic, but now, with quantitative data available, we know perfection is not only unrealistic, but also quite simply unachievable.

Rather than perfection, which expects that every relevant, non-privileged document will be found and produced, the standard against which we measure these new technologies and processes must be based upon the same principles that have traditionally governed discovery – reasonableness and good faith. Although these terms conjure thoughts of ambiguity and uncertainty, they can actually represent a well-defined set of expectations when placed within the context of process.

A process that emphasizes reasonableness and good faith is fully consistent with what is required under the discovery process. Discovery of information relevant to a dispute gathered by an opponent is often central to a fair and efficient resolution.[39] A party need only identify and produce that which is relevant, as defined by the rules, with the degree of diligence expected and available by experienced practitioners acting reasonably.[40] As noted in Sedona Principles 6 and 11, a party may choose to implement this approach in a reasonable manner, which is left to the good judgment of the party.

Sound process applied to the use of search and retrieval technology can readily establish a measurable means for conducting discovery that satisfies the rules. Reasonableness and good faith can be defined and measured by identifying performance criteria based on their attributes. Accordingly, the unreasonable and unattainable goal of "perfection" should not be allowed to be an enemy of the attainable and measurable goal of reasonableness.

As search and retrieval technologies and associated processes are developed, parties will no doubt want to use them in order to achieve defensible and credible results. If a party fails to adhere to appropriate performance guidelines it will be subject to scrutiny and criticism. Therefore, established process in conjunction with sound technology can serve as a benchmark for conducting discovery in the future. Furthermore, defensibility in court will very likely depend on the implementation of, and adherence to, processes developed for use with a search and retrieval technology.

Implementing Process

Using a search and retrieval technology in conjunction with an implementing process in the complex context of electronic discovery will involve multiple phases of activity, with iterative feedback opportunities at appropriate decision points to allow integration of what a case team learns after each exercise of the process in order to calibrate and maximize the technology's capability to identify relevant information. It is through this feedback that case teams will acquire sound information to use in making both strategic and tactical decisions.

The initial search and retrieval process should be designed with the intent that it serve as a pilot process that can be evaluated and modified as the team learns more about the corpus of information to be reviewed. One useful approach is to initiate the process by focusing on the information collected from a few of the custodians who were at the center of the facts at issue in the litigation or investigation. Focusing on information collected from the core custodians, which has a higher likelihood of being relevant, will help the team efficiently develop its understanding of the issues and language used by the custodians, thus allowing them to more efficiently develop and implement an appropriate search and retrieval process.

---

39 *Hickman v. Taylor*, 329 U.S. at 507; *see supra*, n.1.

40 Under Rule 26(g)(1), an attorney of record is expected to certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that disclosures are "complete and correct" as of the time they were made. Similarly, under Rule 26(g)(2), an attorney must certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that discovery requests, responses, and objections" are made "consistent with these rules."

The initial selection and refinement of search terms can also benefit from the application of sampling techniques that can help the review team to rank the precision and recall of various terms or concepts. Reviewing samples of information that include selected search terms or concepts and ranking their relative value based on their efficacy in retrieving relevant information (recall) and their efficiency in excluding non-relevant information (precision) can help the review team to focus the selection of terms.[41]

The development of process control logs and second-level review techniques can also help the review team to ensure that the designed process is consistently applied to all of the information to be reviewed. Additionally, a second-level review process based on statistical sampling techniques can ensure the achievement of acceptable levels of quality. While these techniques are relatively unknown in the typical review processes in use today, their widespread adoption in businesses of all types should drive their implementation in large document review projects in the near future.

C. How The Legal Community Can Contribute to The Growth of Knowledge

A consensus is forming in the legal community that human review of documents in discovery is expensive, time consuming, and error-prone. There is growing consensus that the application of linguistic and mathematic-based content analysis, search and retrieval technologies, and tools, techniques and process in support of the review function can effectively reduce the cost, time, and error rates.

Recommendations

1. ***The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.***

2. ***The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.***

As stated, in the 20 years since the Blair and Maron study, there has been little in the way of peer-reviewable research establishing the efficacy of various methods of automated content analysis, search, and retrieval as applied to a legal discovery context. A program of research into the relative efficacy of search and retrieval methods would acknowledge that each alternative should be viewed in the context of its suitability to specific document review tasks. Different technologies, tools and techniques obviously have different strengths. Moreover, the outcome of the application of advanced content analysis, search and retrieval methods can have significant differences based on expertise of the operator. Ideally, a research program would advance the goals of setting minimum or baseline standards for what constitutes adequate information retrieval, as well as reaching agreement on how to benchmark competing methods against agreed-upon objective evaluation measures.

In this regard, The Sedona Conference® supported the introduction of a new "Legal Track" in 2006 for the TREC research program run by the National Institute of Standards and Technology. NIST is a federal technology agency that works with industry to develop and apply technology, measurements and standards. TREC is designed "to encourage research in information retrieval from large text collections."[42] The TREC legal track involves an evaluation of a set of search methodologies

41 *See* text at Part IV, *supra*.

42 The Text Retrieval Conference (TREC) was started in 1992. *See* http://trec.nist.gov. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC is overseen by a program committee consisting of representatives from government, industry, and academia. Each TREC track involves a test database of documents and topics Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST generally pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The TREC cycle ends with a workshop that is a forum for participants to share their experiences. The TREC test collections and evaluation software are available to the retrieval research community at large, so organizations can evaluate their own retrieval systems at any time. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and of facilitating technology transfer, and many of today's commercial search engines include technology first developed in TREC.

based on lawyer relevancy assessments on topics drawn from a large public database of OCR-ed documents. The results coming out of the 2006 legal track represent the type of objective research study into the relative efficacy of Boolean and other search methods that the legal community should further encourage.[43]

However, a need exists to scale up the TREC research to accommodate the potential retrieval of millions or tens or hundreds of millions of arguably relevant documents among a greater universe of terabytes, petabytes, exabytes, and beyond.

Members of The Sedona Conference® community have and will continue to participate in collaborative workshops and other fora focused on issues involving information retrieval.[44] How best to leverage the work of the IR community to date is an enterprise beyond the scope of this paper. The Sedona Conference® intends to remain in the forefront of the efforts of the legal community in seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships aimed at focused research.

43 *See* Jason R. Baron, David D. Lewis & Douglas W. Oard, "TREC 2006 Legal Track Overview," 2006 FIFTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2006) PROCEEDINGS, *available at* http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf; *see also* TREC 2007 Legal Track, http://trec-legal.umiacs.umd.edu/ (additional documentation relating to TREC 2006 Legal Track).

44 *See, e.g., Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings* ("DESI Workshop"), held at the Eleventh International Conference on Artificial Intelligence and Law (ICAIL 2007), June 4, 2007, Palo Alto, CA, papers *available at* http://www.umiacs.umd.edu/~oard/desi-ws/.

APPENDIX: Types of Search Methods

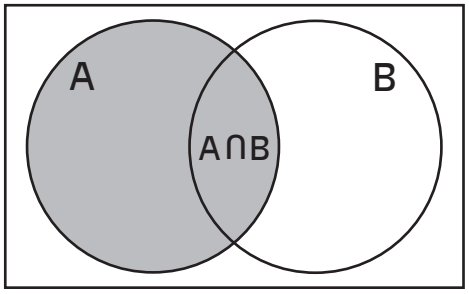
This appendix is a "survey" of different forms of search methods found in the information science literature, and which form the basis of offerings by vendors in the legal marketplace. The list is not definitive. Indeed, as the main body of the Commentary makes clear, rapid technological progress will inevitably affect how methods are described, perfected, and then replaced with new ways of performing search and retrieval.

A second caveat: the following search methods are not intended to be mutually exclusive. Indeed, many products tout the benefits of hybrid, combined, or cumulative approaches to performing searches.

A. Boolean Search Models

A "Boolean search" utilizes the principles of Boolean logic named for George Boole, a British born mathematician. Boolean logic is a method for describing a "set" of objects or ideas. Boolean logic was applied to information retrieval as computers became more widely accepted. Boolean search statements can easily be applied to large sets of unstructured data and the results exactly match the search terms and logical constraints applied by the operators.

As used in set theory, a Boolean notation demonstrates the relationship between the sets or groups, indicating what is in each set alone (set union), what is jointly contained in both (set intersection), and what is contained in neither (set differences). The operators of AND (intersection or ∩), OR (union or ∪) and AND NOT or BUT NOT (difference) are the primary operations of Boolean logic. These relationships can easily be seen within a Venn diagram (see below).



OR is a Boolean operator that states the set may contain any, some or all of the keywords searched. The purpose of this command is to encompass alternative vocabulary terms. OR is represented by the union of the sets A ∪ B (the entire shaded areas above). The use of OR expands the resulting Boolean set.

AND is a Boolean operator used to identify the intersection of two sets or two keywords. The purpose of this command is to help construct more complex concepts from more simple vocabulary words. AND is represented by the middle intersecting area above (A ∩ B). The use of AND restricts the resulting Boolean set.

NOT is a Boolean operator used to eliminate unwanted terms. The purpose of this command (preceded by either AND or BUT) is to help suppress multiple meanings of the same term; in other words, eliminating ambiguity.

Different search engines or search tools may provide additional Boolean-type operators or connectors to create more complex search statements. These may include:

- **Parenthesis:** A Boolean search may include the use of parentheses to force a logical order to the execution of the search, as well as to create more refined and flexible criteria. Any number of logical ANDs (or any number of logical ORs) may be chained

together without ambiguity; however, the combination of ANDs and ORs and AND NOTs or BUT NOTs sometimes can lead to ambiguous cases. In such cases, parentheses may be used to clarify the order of operations. The operations within the innermost pair of parentheses are performed first, followed by the next pair out, etc., until all operations are completed.

- **Proximity or NEAR/WITHIN operator:** Another extension to Boolean searching, this technique checks the position of terms and only matches those within the specified distance. This is a useful method for establishing relevancy between search criteria, as well as for paring down irrelevant matches and getting better results (improving precision). Some search engines let you define the order, in addition to the distance. For example: *budget w/10 deficit* might mean "deficit within the 10 words following the word budget".

- **Phrase Searching:** Some search engines provide an option to search a set of words as a phrase, either by typing in quote marks (" ") or by using a command. When they receive this kind of search, the engines will generally locate all words that match the search terms, and then discard those which are not next to each other in the correct order. To perform this task efficiently, the index typically will store the position of the word in the document, so the search engine can tell where the words are located.

- **Wildcard operators** (also sometimes referred to as truncation and stemming). This search capability allows the user to widen the search by searching a word stem or incomplete term. It is typically a symbol such as a question mark (?), asterisk (*), or exclamation point (!). The search system may also allow the user to restrict the truncation to a certain number of letters by adding additional truncation symbols. For example: Teach?? would find teaches and teacher but would not find teaching. In addition, some systems will allow for internal truncation such as wom?n would find women or woman. The * and ! terms have broader application: for example, hous* would find house, housemate, Houston, household or other similar words with the stem "hous."

B. Probabilistic Search Models: Bayesian Classifiers

Probability theories are used in information retrieval to make decisions regarding relevant documents. The most prominent of these are so-called "Bayesian" systems or methods, based on Bayes' Theorem. The theorem was developed by Thomas Bayes, an eighteenth century British mathematician. A Bayesian system sets up a formula that places a value on words, their interrelationships, proximity and frequency. By computing these values, a relevancy ranking can be determined for each document in a search result. This weighting may be based on a variety of factors:

- Frequency of terms within a document- the more times it appears, the more weight it carries.

- Closer to the top – documents with the term in the title are more heavily weighted

- Adjacency or proximity – the closer the terms are to each other, the higher the weighting

- Explicit or implicit feedback on relevance

(Note: other types of search models apply these types of concepts or ideas as well.)

Bayesian systems frequently utilize a "training set" of highly relevant documents to increase understanding, and therefore the probability measures of the system. During training, the system examines each word in the training documents and computes the probability with which that word occurs in each category.

For example, the word "potato" may occur in 5 documents in the category "kitchen tools" (*e.g.*, "potato peeler"), in 7 documents in the category "farm products," and in one document in the category "garden tools." When a new document is then found to contain the word "potato," the Bayesian classifier will interpret this new document as most likely to be a member of the category "farm products" than either of the other two. The same process is repeated for all of the words in the document. Each word in the document provides evidence for which of the categories the document belongs to. The Bayesian classifier combines all of this evidence, using Bayes' rule, and determines the most likely category.

Bayesian classifiers provide powerful tools for comparing documents and organizing documents into useful categories with a moderate amount of effort.

C. Fuzzy Search Models

Boolean and probabilistic search models rely on exact word matches to form the results to a query. Exact matching is very strict: either a word matches or it doesn't. Fuzzy search is an attempt to improve search recall by matching more than the exact word: fuzzy matching techniques try to reduce words to their core and then match all forms of the word. The method is related to stemming in Boolean classifiers, discussed above.

Some algorithms for fuzzy matching use the understanding that the beginning and end of English words are more likely to change than the center, so they count matching letters and give more weight to words with the matching letters in the center than at the edges. Unfortunately, this can sometimes bring up results that make little sense (a search for tivoli might bring up ravioli).

Many systems allow one to assign a degree of "fuzziness" based on the percentage of characters that are different. Fuzzy searching, or matching, has at least two different variations: finding one or more matching strings of a text, and finding similar strings within a fixed string set often referred to as a dictionary. Fuzzy searching has many applications in legal information retrieval including: spellchecking, email addresses and OCR clean-up.

D. Statistical Methods: Clustering

Systems may use statistics or other machine-learning tools to recognize what category certain information belongs to. The simplest of these is the use of "statistical clustering." Clustering is the process of grouping together documents with similar content. There are a variety of ways to define similarity, but one way is to count the number of words that overlap between each pair of documents. The more words they have in common, the more likely they are to be about the same thing.

Many clustering tools build hierarchical clusters of documents. Some organize the documents into a fixed number of clusters. The quality or "purity" of clustering (*i.e.*, the degree to which the cluster contains only what it should) is rarely as high as that obtained using custom built taxonomies or ontologies, but since they require no human intervention to construct, clustering is often an economical and effective first pass at organizing the documents in a collection.

Some systems improve the quality of clusters that are produced by starting with a selected number of clusters, each containing selected related documents. These selected documents then function as "seeds" for the clusters. Other related documents are then joined to them to form clusters that correspond to their designer's interests. Then, additional documents are added to these clusters if they are sufficiently similar.

E. Machine Learning Approaches to Semantic Representation

Bayesian classifiers are often considered "naïve" because they assume that every word in a document is independent of every other word in the document. In contrast, there is a class of concept learning technologies that embrace the notion that words are often correlated with one another, and that there is value in that correlation. These methods are also referred to as "dimensionality reduction techniques" or "dimension reduction systems."

These systems recognize there is redundancy among word usage and take advantage of that redundancy to find "simpler" representations of the text. For example, a document that mentions "lawsuits" is also likely to mention "lawyers," "judges," "attorneys," etc. These words are not synonyms, but they do share certain meaning characteristics. The presence of any one of these words would be suggestive of their common theme. Documents that mentioned any of these terms would likely be about law. Conversely, in searching for one of these words, one might be almost as satisfied to find a document that did not contain that exact word, but did contain one of these related words.

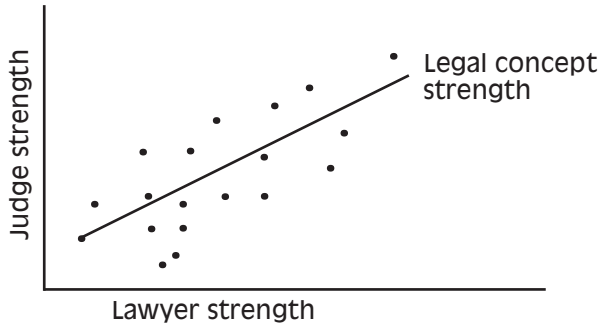


Figure 1. Dimension reduction – the original dimensions of "lawyer" and "judge" are combined into a single dimension. Each point in the graph represents a document. Its location in the graph shows how much the document is related to each dimension.

The figure above illustrates the kind of relationships such systems find. The word "lawyer" tends to occur in the same context as the word "judge." Each document has a certain strength along the "lawyer" dimension, related, for example, to how many times the word "lawyer" appears. Similarly, documents have strength along the "judge" dimension, related, for example, to how many times the word "judge" appears. These systems find a new dimension that summarizes the relationship between "lawyer" and "judge." In this example, we are reducing the dimensions from two to one.

Mathematically, we can then describe documents by how much strength they have along this dimension and not concern ourselves with its strength along the original "lawyer" or "judge" dimensions. The new dimension is a summary of the original dimensions, and the same thing can be done for all words in all the documents. We can locate documents along these new, reduced, dimensions or we can represent words along these dimensions in a similar way.

Similarly, multiple words can be represented along dimensions. And, instead of having just one summary dimension, we can have many of them. Instead of describing a document by how it relates to each of the words it contains, as is done with Vector Space Models,[45] we can describe the document by how it relates to each of these reduced dimensions. Latent Semantic Indexing (LSI, also called Latent Semantic Analysis) is probably the best known of these dimension-reducing techniques, but there are others, including neural networks and other kinds of statistical language modeling.

Such techniques are similar to one another in that they learn the representations of the words in the documents from the documents themselves. Their power comes from reducing the dimensionality of the documents. They simplify representation, and make recognizing meaning easier.

For example, a collection of a million documents might contain 70,000 or more unique words. Each document in this collection can be represented as a list of 70,000 numbers, where each number stands for each word (say the frequency with which that word occurs in that document). Using these techniques, one can represent each document by its strength along each of the reduced dimensions.

45 *See* H. Roitblat, *supra*, n.34.

One can think of these strengths as a *meaning signature*, where similar words will have similar meaning signatures. Documents with similar meanings will have similar meaning signatures. As a result, the system can recognize documents that are related, even if they have different words, because they have similar meaning signatures.

F. Concept and Categorization Tools: Thesauri, Taxonomies and Ontologies

To deal with the problem of synonymy, some systems rely on a thesaurus, which lists alternative ways of expressing the same or similar ideas. When a term is used in a query, the system uses a thesaurus to automatically search for all similar terms. The combination of query term and the additional terms identified by the thesaurus can be said to constitute a "concept."

The quality of the results obtained with a thesaurus depends on the quality of the thesaurus, which, in turn, depends on the effort expended to match the vocabulary and usage of the organization using it. Generic thesauri, which may attempt to represent the English language or are specialized for particular industries, are sometimes available to provide a starting point, but each group or organization has its own jargon and own way of talking that require adjustment for effective categorization. In America, for example, the noun "jumper" is a child's one-piece garment. In Australia, the noun "jumper" is a sweater. In America, a 3.5 inch removable disk device was called a "floppy" during its heyday. But in Australia, it was called a "stiffy."

Taxonomies and ontologies are also used to provide conceptual categorization. Taxonomy is a hierarchical scheme for representing classes and subclasses of concepts. The figure below shows a part of a taxonomy for legal personnel. Attorneys, lawyers, etc. are all kinds of law personnel. The only relations typically included in a taxonomy are inclusion relations. Items lower in the taxonomy are subclasses of items higher in the taxonomy. For example, the NAICS (North American Industry Classification System) is one generally available taxonomy that is used to categorize businesses. In this taxonomy, the category "Information" has subclasses of "Publishing" and "Motion Picture and Sound Recording Industries" and "Broadcasting."

One can use this kind of taxonomy to recognize the conceptual relationship among these different types of personnel. If your category includes law personnel, then any document that mentions attorney, lawyer, paralegal, etc. should be included in that category. Like thesauri, there are a number of commercially available taxonomies for various industries.

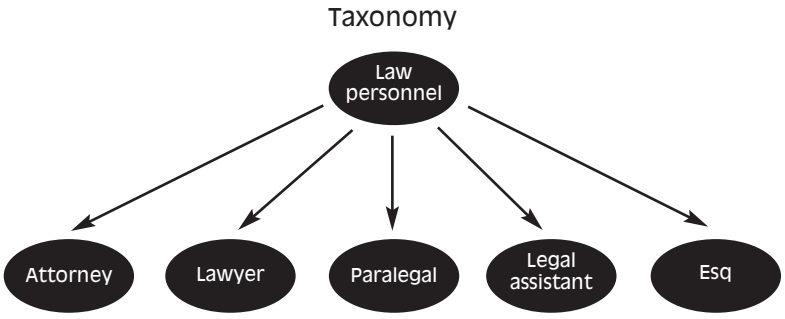


Figure 2. A simple taxonomy for law personnel.

Predefined taxonomies exist for major business functions and specific industries. It may be necessary to adapt these taxonomies to one's particular organization or matter.

An ontology is a more generic species of taxonomy, often including a wider variety of relationship types than are found in the typical taxonomy. An ontology specifies the relevant set of conceptual categories and how they are related to one another. The figure below shows part of an ontology covering subject matter similar to that described in the preceding taxonomy. For clarity, only a subset of the connections between categories is shown. According to this ontology, if your category includes attorneys, you may also be interested in documents that use words such as "lawyer," "paralegal," or "Esq." Like taxonomies, ontologies are most useful when they are adapted to the specific information characteristics of the organization.

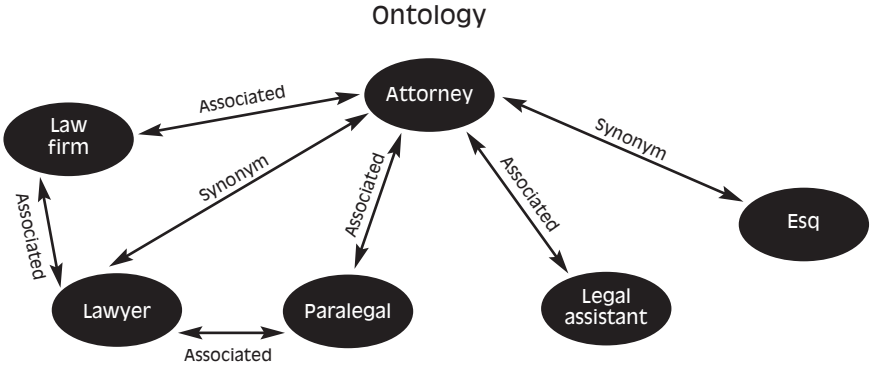


Figure 3. A section of an ontology of legal personnel.

Taxonomies, ontologies, and thesauri are all knowledge structures. They represent explicit knowledge about some subject. An expert writes down the specific relations she knows about. Although there are tools that help the expert create these structures, they still tend to represent only the information the expert can explicitly describe as important.

The structure of the thesaurus, taxonomy, or ontology can be used as the organizing principle for a collection of documents. Rules are derived that specify how documents with specific words in them are related to each of these categories, and the computer can then be used to organize the documents into the corresponding categories.

These rules can be created explicitly, or they can be created using machine-learning techniques. Explicit rules are created by knowledge engineers. For example, one rule might include a Boolean statement like this: (acquir* or acquisition or divest* or joint venture or alliance or merg*) and (compet* or content or program*) that specifies the critical words that must appear for a document to be assigned to the "merger" category. The effectiveness of rules like these depends critically on the ability of the knowledge engineers to guess the specific words that document authors actually used. Syntactic rules may also be employed by some systems. For example, a system may only look for specific words when they are part of the noun phrase of a sentence.

G. Presentation/Visualization Tools

Presentation and visualization software technologies may incorporate search and retrieval functionality that may be found to have useful applications. These technologies can organize information (*e.g.*, emails) so that a researcher can more efficiently study the research topic (including finding relevant emails). They also are good at highlighting patterns of "social networks" within an organization that would not necessarily be apparent by more traditional searches. Subject to some exceptions, the results of any search and retrieval query can be presented in a variety of forms, including as a:

1. List – items in sequence, for example messages ordered by sent date
2. Table – items aggregated into rows by columns, for example messages by sender
3. Group – items categorized or totaled, for example count of messages by sender

4. Cluster – items in groups organized by spatial proximity, for example relevant groups spiraling out to less relevant groups
5. Tree – items in parent/child hierarchy, for example, folder and subfolder(s)
6. Timeline – items arrayed by a time element, for example a list/group of items arrayed by sent date
7. Thread – items grouped by conversation
8. Network – items arrayed by person, for example a diagram of message traffic between sender(s) and recipient(s)
9. Map – items plotted by geography, for example items plotted by city and state of origin
10. Cube – items in a multi-dimensional pivot table; includes, table, group, timeline and tree functionality

In practice, a researcher can load search results into a presentation technology for an organized view, and then drill-down to access discrete items of significant interest or concern. This often iterative process may help a researcher to learn more about, act on, and manage search results.

ADVANCED LEARNING IN A PANORAMIC SETTING℠

Visit www.thesedonaconference.org