Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Adapting to the Data Explosion:
Ensuring Justice for All

Julia L. Brickell
Lecturer, Executive Master of Science
in Technology Management
Columbia University
New York, NY
jlb130@columbia.edu
&
H5
New York, NY
jbrickell@h5.com

Arthur M. Langer
Senior Director, Center for Technology, Innovation,
and Community Engagement
Fu Foundation School of Engineering
and Applied Science
Columbia University
New York, NY
al261@columbia.edu

***Abstract*—The age of electronic information will continue to engender a technological evolution to which the efforts of human beings will have to adjust. In the legal realm in the United States, the time and costs associated with meeting a party's obligations to search and turn over relevant electronic data for litigation threatens to drive litigants out of the courts. There is now judicial recognition that technological tools, especially for automated search, will have to play a role in case discovery, which used to be handled by humans alone. Improved understanding of human-machine interplay is of growing importance to provide effective, and cost effective, information retrieval in this litigation context. Pairing technology and human expertise in a sophisticated way can provide benefits not possible when either is used on its own.**

***Keywords*—human-machine cooperation and systems, information retrieval, legal factors, search methods**

I. INTRODUCTION

Civil litigation in the United States, from a tenant-landlord dispute to a multi-million dollar corporate anti-trust case, is now challenged by the growing volume of electronic information generated by people who, at home and at work, create information that may need to be searched for litigation. Supreme Court Justice Stephen Breyer has commented that "If it really costs millions to [review and produce electronic information], then you're going to drive out of the litigation system a lot of people who ought to be there." [1] In essence, because the exchange of information between parties in litigation is essential to our legal system, an economic obstacle to that effort may stand in the way of access to justice, and ultimately to justice itself.

The Federal Rules of Civil Procedure [2], under which civil litigation is carried out in the federal courts of the United States, proffers as its first rule the goal of "just, speedy and inexpensive determination of every action." It also provides rules (Rules 26-27) under which parties are allowed to ask for, and must give to each other "discovery" of information that may have a bearing on the case. Given the enormous volume of information that is now created and stored electronically, discovery (now more commonly electronic discovery or "e-discovery") even in small cases can be cost-prohibitive, sometimes exceeding the amount in dispute. According to one estimate, a "midsize" lawsuit is now expected to generate between $2.5 and $3.5 million in e-discovery costs alone, while large corporate litigation may compile costs of $10M or more [3]. Many litigants will settle or handle their disputes outside the court system to avoid the disproportionate expense of discovery.

Further, judges increasingly demonstrate their expectation that people and organizations can find in a timely manner discoverable information among their own data, imposing what may be substantial monetary sanctions and/or adverse inferences if this expectation is not met.[1] For the legal system to continue to be available to those who struggle with the growing volume of data and its attendant risks, new strategies must be employed, including technological tools and methods, that will enable a reasonable, defensible, and cost-effective retrieval of evidentiary information. Improved methods for the search and review of information along with rules that demand focus on the important information for a case could reduce the discovery burden and again permit a just, speedy and inexpensive determination of an action.

II. VOLUME AND COMPLEXITY OF DATA

Since the late 1980s, companies and individuals have increasingly created information electronically. It is now estimated that more than 93% percent of corporate data originates in electronic form [4]. With increased storage capacity at decreased cost, the volume of electronically stored information in the corporate data center and personal computer alike has reached staggering proportions. In 2006, the amount of digital information created, captured, and replicated was roughly 161 billion gigabytes, about 3 million times the information in all books ever written. Between 2006 and 2010, digital information is expected to increase more than six-fold [5].

[1] See *Coleman (Parent) Holdings, Inc. v. Morgan Stanley & Co.*, 2005 WL 679071 (Fla.Cir.Ct. Mar 1, 2005), *rev'd on other grounds*, 955 So.2d 1124 (Fla.4th CA 2007); *Qualcomm Inc. v. Broadcom Corp.*, 2008 WL 66932 (S.D.Cal. Jan. 7, 2008).

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

Although such volumes impose challenges to the discovery of information for litigation, quantity is not the only issue and litigation is not the only demand. Statistics show that the employee, awash in information that flows 24/7, wastes hours searching for, sorting, and assessing information, all of which adds up to significant organizational productivity cost. It has been estimated that an enterprise with 1,000 knowledge workers loses a minimum of $6 million a year in the time workers spend searching for—and not finding—needed information [6].

Information retrieval becomes a formidable task as unstructured data formats spawned by new technologies proliferate, adding to the already complex data stores created by email, voicemail, and myriad mobile devices. Complicating matters even more, software applications and storage methods are subject to ongoing upgrade and evolution presenting technical obstacles that may render associated data unintelligible or inaccessible to search without costly conversion or forensic expertise. Routine data management may permanently alter or destroy existing data without the owner's permission or knowledge. Shared data environments such as e-rooms and wikis where digital information is freely exchanged and modified can obfuscate the owner or creator of digital material with implications for litigation when discovery from specific individuals is needed. Demonstration of the chain of custody may be inadvertently impaired by the movement of data. Metadata, system records, and deleted files may become discoverable in and of themselves, expanding the potential scope of discovery (although such discovery may be conditioned on a showing that it is likely to result in substantive, material evidence not otherwise available) [7].

Although most digital information is created by individuals, organizational entities ultimately become responsible for most of it [5] and they are the ones most often subject to complex and repeated litigation. A *2006 Fulbright & Jaworski Litigation Trends Survey* showed that the average U.S. company faces 305 suits at any one time; that number jumps to 556 for companies with $1 billion or more in revenue [8]. The duty to preserve electronic information that may be pertinent to pending or potential litigation arises in each of these cases; most of these suits will require some form of interrogation of electronically stored information for the purposes of e-discovery, imposing search and review burdens on the parties involved.

III. THE RULES OF THE COURT SYSTEM: DISCOVERY

The Federal Rules of Civil Procedure [2] envision that if the facts bearing on a matter are fully known, a just result will ensue. The rules entitle a party to seek discovery from the opponent in several ways: a party can ask for documents and inspection of things (Rule 34), send written questions (Rule 33), interview witnesses under oath (Rules 30, 31), and examine people (Rule 35). Requests for information are limited to seeking matter relevant to a claim or defense and "reasonably calculated to lead to the discovery of admissible evidence"(Rule 26 (b)). The rules specify the manner in which information may be requested by and produced to opposing parties.

In 2002, some members of the legal community concerned about whether rules and concepts developed largely for paper discovery would be adequate to address issues of electronic discovery had already begun to consider and write about the challenges related to production of electronic information [9]. The Sedona Conference®, a legal think tank, produced the *Sedona Principles Addressing Electronic Document Production (*now in its *Second Edition)* the goal of which was to "stimulate productive discussion and promote the formulation of legal doctrine consistent with principles of fairness, equity and efficiency." [9] Subsequently, amendments to the rules were approved in 2006 to address the differences between paper and electronic discovery, although many problematic issues related to electronically stored information still remain [9]. The rules amendments served to acknowledge that characteristics related to volume, format, mutability, transience, storage, and accessibility, among others, confer a distinction on electronic data that impacts the way in which parties must address their discovery obligations. Rule 34 goes so far as to include the concept of sampling data from an entire dataset to determine if additional discovery is warranted [2], a concept that also appears in Sedona's 11th Principle 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" [9].

Ideally, reasonable discovery allows for disclosure of all relevant facts pertinent to a case while not imposing excessive burdens and costs on litigants and the court [9]. The volume of electronically stored information, however, is routinely posing a burden, often excessive, both in terms of cost and human resources.

IV. E-DISCOVERY AND THE INTERACTION AND EVOLUTION OF LAW, HUMAN BEINGS, AND TECHNOLOGY

The purview of the legal domain is to address the application of justice in all of its aspects. The rule of law is necessarily evolving to accommodate the challenges of electronically stored information as its effects are better understood, exemplified by various rules amendments and recent case rulings. Jurists are increasingly weighing in on the more robust aspects of technology as applied to e-discovery, with one notable reference to the complexities and pitfalls of search techniques as an "interplay, at least, of the sciences of computer technology, statistics and linguistics."[2] In fact, law in the service of human beings and technology in the service of their legal efforts is an intricate and interdependent relationship.

The processes, procedures, and tools that legal practitioners rely upon in their work, especially as related to discovery, are evolving as well. Each step in the e-discovery process—identifying and collecting relevant data, processing it for review, the review itself, and its production to the requesting party—now requires a unique collaboration of human beings and technology. Human review, the costly process carried out by legal professionals, has traditionally been the means to

[2]See *United States v. O'Keefe*, 537 F. Supp. 2d 14 (D.D.C. 2008)

locate information to support a party's own contentions in a lawsuit and to determine what subset of documents must be produced to the other side, identifying in the process that which may be handled differently due to confidentiality or privilege concerns. Although there is wide variation due to circumstance, the cost of document review can account for 50-90% of the total cost of a litigation matter [10]. Unlike efforts pre-dating the advent of electronic information when bankers' boxes of hardcopy documents were collected and reviewed by attorneys, electronic tools are now available for e-discovery activities at every step along the way, from information retrieval to data preparation, hosting and review platforms, forensic analysis, and production to the requesting party. Rule 34(b) [2] was in fact amended to support the production of material to the other side in electronic format, in lieu of the traditional massive print effort.

Not surprisingly, the technology that fosters such staggering data volumes is evolving to help manage it (as are the offerings of e-discovery services providers who now see an aggregate yearly revenue approaching $3 billion [11]). Tools that assist people in organizing or categorizing electronic information before it is stored (commonly known as "knowledge management" systems) are proliferating, although any benefit these systems may present in the e-discovery realm is as yet unmeasured. Optical character recognition (OCR) technology, which translates non-digital material into digital format is improving, making the contents of OCR'd material more accessible to search. Hash algorithms and de-duplication tools are applied to enable single instance document identification, reducing the volume of data for review. In the legal domain, document review software applied to repositories of data relieves burdens on the human reviewer with features such as "mass tagging" of documents with similar content, saving untold hours of review time.

Advances in data management and end-user tools, however, which may reduce some of the onus on the human being, are subordinate to those related to the information retrieval effort when it comes to e-discovery. The legal imperative to locate and produce relevant information related to a particular topic dictates that the discovery of electronically stored information "stands on equal footing with the discovery of paper documents" ([12] See Committee Note for Rule 34(a)). That may be so, but it is clear that information retrieval efforts cannot be on equal footing, since human beings cannot retrieve responsive documents from the corporate data center the way they can from the office file drawer. The Sedona Conference® Practice Point 1 notes that "[m]anual search processes for the purpose of finding responsive documents may be infeasible or unwarranted in many instances and the use of automated search methods should be viewed as reasonable, valuable, and even necessary" [13]. In this context, automated search becomes a review tool in and of itself, albeit one which includes the ambiguities of human input, as its execution results in determinations made about what is potentially responsive.

Just as information on the Internet would be unknowable, not to mention inaccessible, without the requisite advances in technology that allow retrieval of a few pieces of pertinent information, so too with enterprise corpora, from which litigants' requested information cannot be accurately retrieved without even more sophisticated search methods. Further, considering the high cost of document review and the attendant risk of missing relevant material, the size of the dataset subjected to human review is a major factor in the overall cost and risk of e-discovery; that is where effective and defensible information retrieval methods play a vital role.

V. TECHNOLOGY AND IR METHODOLOGIES EVOLVING TO HELP HUMANS TO REDUCE TIME, COSTS, RISKS

The perfect search result for an e-discovery effort would be the retrieval of all relevant documents (100% "recall") and only relevant documents (100% "precision"). Not surprisingly, this goal has never been met by any known search and retrieval method to date, either manual or electronic. "Precision-recall" has been a topic of several published and proprietary studies alike and the results have shown that manual review, like most automated approaches, entails a steep trade-off between recall and precision. The operative concept is that any search, manual or electronic, involves to a greater or lesser degree the participation of human beings: a search engine, such as one executing a keyword or Boolean search, can operate on its own only after human input.

A well-known study by Blair & Maron found that manually built Boolean queries achieved 20% recall at 79% precision (although users had the impression that they had achieved 75% recall) [14]. Other unpublished studies show that straight manual review, long considered the "gold standard," does not perform much better, and surely does not lessen the need to negotiate the recall-precision trade-off: historically, to capture more than 50% of targeted material, generally more than 50% of unwanted material will be contained in the result set. Inversely, to achieve a result set that contains less than 50% unwanted material, less than 50% of targeted material will be included in the result set. In the e-discovery context, the stated demand is generally that "all" responsive information be produced so searches are often crafted to that end. Although the evidence is that those attempts usually fall short, getting close to the goal raises the cost of the review effort since precision will usually drop, subjecting a host of "false positive" documents to human examination. Search methods that emphasize precision over recall may miss responsive documents, either incriminating and/or exculpatory, each of which presents its own element of risk.

Although the legal domain has historically been in the forefront of advanced search technology, pioneering the development of text retrieval when it first became available in the 1970's, [15] it is not advancing as quickly in adopting more sophisticated search and retrieval methods for e-discovery. Research and development of new algorithms for information retrieval applicable to e-discovery have been underway since the early 1990's with government funded programs such as Tipster and Translingual Information Detection Extraction and Summarization (TIDES) [15]. Tools that supplement keyword searching and Boolean techniques, including fuzzy logic to capture variations on words, the use of taxonomies and ontologies assembled by linguists in the service of concept search, and tools that employ mathematical probabilities are slowly finding recognition in the legal community. However,

the use of traditional keyword and Boolean search persists as the tool of choice among lawyers engaged in e-discovery [13].

Recently, several high-profile cases have shone a spotlight on keyword search as well as the agreements lawyers make in using them.[3] These cases contribute to the recent trend in e-discovery case law toward greater discussion by the judiciary of the competence of counsel in handling their clients' electronic discovery. Simultaneously, the judiciary has increased its involvement in making parties negotiate about how they propose to go about conducting automated searches for relevant evidence [16]. One recent case, *In re Fannie Mae Securities Litigation,*[4] provides an eye-opening example to consider: counsel for the Office of Federal Housing Enterprise Oversight, a non-party to the litigation in question, agreed to a keyword list that ended up retrieving 80% of the agency's email. An attempt to enlist the court's help to mitigate the result failed, both at the trial and appellate level, and its predicted spend for the review and production was $6 million, or 9% or the agency's operating budget [17].

According to the Sedona Conference®, the recognition that human review of documents in discovery is expensive, time consuming, and error-prone is leading to a growing consensus that more sophisticated information retrieval methods can effectively reduce litigation cost, time, and error rates lessening the burden on the litigant and making justice more affordable. Such information retrieval methods have only recently been subjected to the comparative assessments and benchmarking standards that would provide legal practitioners with a basis for making informed search method selections. One of these efforts is TREC Legal Track.

VI. The Interplay of Technology and Human Knowledge: TREC Legal Track and the Interactive Task

The TREC Legal Track, part of the Text REtrieval Conference (TREC) sponsored by National Institute of Standards and Technology (NIST), was established in 2006 as a platform to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records [18]. The far-reaching TREC Legal Track objective for bench and bar, articulated in an open letter from The Sedona Conference® to law firms and companies in the legal tech sector, is a "credible, collaborative, and independent process and protocol by which both long-established and emerging search methods used for document review may be evaluated and benchmarked [19]." In 2008, the TREC Legal Track included a revised "Interactive Task," which models more accurately and completely the real-world conditions in which companies and law firms, and the e-discovery firms they engage, must meet their document retrieval objectives and obligations [20]. The Interactive Task focused on the retrieval of documents in a dataset relevant to specific topics in order to benchmark the resulting recall and precision of an approach.

Interestingly (and what should be obviously), emulating "real-world conditions" meant introducing into the overall process a very important element: the knowledgeable human being. In the Interactive Task, a single individual was designated to act as an authority for defining the intent and scope of a topic along with a provision that allowed participants to engage with that authority for purposes of clarifying relevance to a topic as well [21]. Not surprisingly, entrants spent varying time with the topic authority and used the information garnered with different degrees of success. But one can hypothesize that the interaction provided an opportunity for improved results over what technology alone could accomplish. TREC Legal Track 2009 is underway, anticipating a broader range of participants who will be using a different publically available data repository. As information retrieval methods and tools are evaluated and benchmarked, legal practitioners will continue to gain insight in the selection of effective methods that they can harness to bring down the cost of discovery.

VII. Conclusion

There is no doubt that the age of electronic information will continue to engender a technological evolution to which both the commonplace and arcane efforts of human beings will have to adjust. In the legal realm, there is now overt judicial acknowledgement that technological tools, especially for automated search, will have to play a role in the pursuit of just efforts to resolve many disputes.

Improved understanding of human-machine interplay is of growing importance, especially in the area of information retrieval for e-discovery, which will soon be (if it is not already) impossible to undertake without the assistance of automated search tools. Pairing technology and human expertise in a sophisticated way can provide benefits not possible when either is used on its own. Understanding and refining the nature of that interaction for an optimal result is certainly an area worthy of future exploration.

From a legal standpoint, there is a growing need for standards and benchmarks by which automated search tools can be measured to show their reasonableness and defensibility when used in a legal context. Efforts along these lines are only in their infancy, but will likely accelerate as the need for them meets the reality of justice impaired without them. With the growing participation by the legal, academic, and scientific communities to contribute to efforts such as TREC, there will be a body of knowledge that the legal system can tap to ensure that the evidentiary challenge of justice is being reasonably and fairly met.

[3] See, for example: *Victor Stanley v. Creative Pipe* 250 F.R.D. 251 (D. Md. 2008), *United States. v. O'Keefe,* 537 F. Supp. 2d 14, 24 (D.D.C. 2008); *In re Fannie Mae Securities Litigation*, 552 F.3d 814 (D.C. Cir. 2009).

[4] See *In re Fannie Mae Securities Litigation*, 552 F.3d 814 (D.C. Cir. 2009).

References

[1] "And Justice for All: How the Electronic Information Explosion Is Transforming the American Legal System." Comments from the Georgetown University Law Center Summit held March 20, 2007.

[2] U.S.A. The Committee on the Judiciary House of Representatives. *Federal Rules of Civil Procedure*. Washington, DC: U.S. Government Printing Office, 2007. [Online]. Available: http://www.uscourts.gov/rules/civil2007.pdf. [Accessed March 30, 2009].

[3] Institute for the Advancement of the American Legal System at the University of Denver, "The Emerging Challenge of Electronic Discovery: Strategies for American Businesses," *Institute for the*

Advancement of the American Legal System at the University of Denver, 2008. [Online]. Available: http://www.du.edu/legalinstitute/pubs/EDiscovery-Strategies.pdf. [Accessed March 30, 2009].

[4] M. C. S. Lange, "Sarbanes-Oxley has major impact on electronic evidence," *National Law Journal*. Jan. 2, 2003. [Online]. Available: http://www.law.com/jsp/article.jsp?id=1039054510969#. [Accessed March 30, 2009].

[5] "The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010." IDC white paper, March 2007. [Online]. Available: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf. [Accessed March 30, 2009].

[6] R. Rao, "From unstructured data to actionable intelligence," *IT Professional,* vol. 5, no. 6, pp. 29-35, Nov./Dec. 2003. [Online]. Available: http://www.ramanarao.com/papers/rao-itpro-2003-11.pdf. [Accessed March 30, 2009].

[7] K. J. Withers, "Is Digital Different? Electronic Disclosure and Discovery in Civil Litigation," section III.c(4). December 30, 1999. [Online]. Available: http://www.kenwithers.com/articles/bileta/index.htm. [Accessed March 30, 2009].

[8] Fullbright & Jaworski, "Third Annual LitigationTrends Survey Findings," 2006. [Online]. Available: http://www.fulbright.com/mediaroom/files/2006/FulbrightsThirdAnnualLitigationTrendsSurveyFindings.pdf. [Accessed March 30, 2009].

[9] The Sedona Conference®, "The Sedona Principles: Second Edition, Best Practices Recommendations & Principles for Addressing Electronic Document Production," page iv-v, June 2007. [Online] Available: from http://thesedonaconference.org. [Accessed March 30, 2009].

[10] B. Burney, "Subdue the costs of document review," *Special to Law.com*, June 23, 2008. [Online]. Available: http://www.law.com/jsp/legaltechnology/pubArticleLT.jsp?id=1202422450816. [Accessed March 30, 2009].

[11] G. Socha and T. Gelbman, "2008 Socha-Gelbman 6th Annual Electronic Discovery Survey." [Online]. Available: http://www.sochaconsulting.com/2008survey/results_001.php. [Accessed March 30, 2009].

[12] *Amendments to the Federal Rules of Civil Procedure*. [Online]. Available: http://www.uscourts.gov/rules/EDiscovery_w_Notes.pdf. [Accessed March 30, 2009].

[13] The Sedona Conference®, "The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery," *The Sedona Conference Journal,* August 2007, (Public Comment Version) [Online]. Available: http://www.thesedonaconference.org. [Accessed March 30, 2009].

[14] D. C. Blair, and M. E. Maron "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM,* vol. 28, pp. 289-299, March 1985.

[15] J. R. Baron and P. Thompson, "The search problem posed by large heterogeneous data sets in litigation: possible future approaches to research," *ICAIL,* 2007. [Online]. Available: http://d.scribd.com/docs/rkma5r9thm33qs9wcph.pdf. [Accessed March 30, 2009].

[16] J. R. Baron, "E-discovery and the problem of asymmetric knowledge," presented at the Mercer Law School Ethics in the Digital Age Symposium, November 7, 2008. [Online]. Available: http://www.law.mercer.edu/academics/centers/mclep/postremarksfinalx.pdf. [Accessed March 30, 2009].

[17] H. C. Boehning and D. J. Toal, "D.C. circuit delivers high-cost EDD lesson." *New York Law Journal.* February, 2009. [Onine]. Available: http://www.law.com/jsp/legaltechnology/pubArticleLT.jsp?id=1202428590930S. [Accessed March 30, 2009].

[18] J. R. Baron, D. D. Lewis, and D. W. Oard. "TREC -2006 Legal Track Overview." [Online]. Available: http://trec-legal.umiacs.umd.edu/LegalTrackOverview2006Final.pdf

[19] The Sedona Conference®, "An Open Letter to Law Firms and Companies in the Legal Tech Sector, Re: Invitation To Participate In The TREC Legal Track." [Online]. Available: http://www.thesedonaconference.org/content/miscFiles/TREC_open_letter.pdf. [Accessed March 30, 2009].

[20] "TREC 2008 Legal Track Interactive Task - Guidelines." [Online]. Available: http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf. [Accessed March 30, 2009].

[21] D. Oard, B. Hedin, S. Tomlinson, and J. Baron. Overview of the TREC 2008 Legal Track. In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*, 2009.