Systematic Review is e-Discovery in Doctor's Clothing

Matthew Lease[†], Gordon V. Cormack[△], An T. Nguyen[‡], Thomas A. Trikalinos[▽], Byron C. Wallace[†]

[†]School of Information

[△]School of Computer Science
University of Waterloo

[‡]Dept. of Computer Science
University of Texas at Austin

[▽]Health Services, Policy & Practice
Brown University

ABSTRACT

Systematic review [6] and electronic (e-)discovery [9] present similar high-recall information retrieval tasks arising in distinct domains. Both traditionally involve an initial high-recall Boolean search followed by expensive expert review, ultimately yielding a low percentage of relevant documents. Both must account for human error in document review in their processes and evaluation, both face tremendous scalability challenges with ever-larger document collections being searched, and both are increasingly turning to active learning in response [3, 11]. While parallels between these tasks have been briefly noted before [8], research on each still remains largely disjoint today. We argue that the time is now ripe for cross-pollinating research on these tasks.

1. INTRODUCTION

Systematic reviews are a key tool of *Evidence-based Medicine*, which seeks to inform health decisions using all available relevant evidence. Systematic reviews identify, describe, critique, and synthesize empirical studies relevant to a well-defined clinical question. Conducting a review requires performing several sequential tasks, including: (i) formulating a precise clinical question to be addressed; (ii) designing and executing a high-recall (and possibly low-precision) Boolean query to retrieve potentially relevant citations; (iii) screening query results for eligibility for inclusion in the review; (iv) extracting the information of interest from the relevant (screened in) articles; (v) assessing the trustworthiness of each study and its applicability to the question at hand, and finally, (vi) synthesizing the information, often using quantitative methods (meta-analysis) [10].

Systematic reviews are laborious and expensive to conduct, a problem exacerbated by the exponential expansion of the biomedical literature [1]. *Screening* (step iii), the main information retrieval task in systematic reviews, typically involves experts (usually with an MD or other terminal degree in a health field) reading the entire set of citations retrieved via database search (in the several or many thousands) to identify the small subset (up to a few hundred) that are potentially eligible for the review.

Electronic discovery (e-discovery) arises in U.S. civil law, when a party is obligated to identify "as nearly as practicable" [4] *all* documents in the party's custody and control that are "responsive" to a set of "requests for production" (RFPs) from another party [8]. E-discovery typically involves the following steps: (i) a written set of RFPs, initially formulated by the requesting party, is finalized either through negotiation or a court ruling, considering the relevance, specificity, and burden of the requests; (ii) a typically large and high-recall/low-precision set of potentially relevant electronic "documents" are extracted from repositories under the responding party's control; (iii) the documents are reviewed for responsiveness by counsel for the responding party; (vi) the responsive documents are further reviewed for privilege and to analyze their importance to the case; (v) all responsive non-privileged documents are produced to the requesting party; (vi) both parties use the documents to inform their legal strategies, and possibly as evidence [8]. Responsiveness review (step iii) typically involves the review by lawyers of tens of thousands, if not hundreds of thousands, or millions of documents, driving new research and products for *technology-assisted review* (TAR).

Call to Action. The brief overviews above of systematic review (SR) and e-discovery reveal common structure in which inclusion criteria are determined, a broad set of potentially relevant documents is identified, the set is manually screened, and documents meeting criteria are further scrutinized. While we are not the first to note such parallels (cf. [8] and the NIST TREC 2015 Total Recall Track), research in each area remains largely disjoint and these parallels have not been deeply explored. Moreover, while IR research on e-discovery has flourished in recent years, thanks to the TREC Legal Track (2006-2011) [4] and 2015 Total Recall track[1], SR has received relatively scant attention in the SIGIR community. We argue that the time is now ripe for cross-pollinating research in these two related domains, to tackle shared problems and explore common opportunities.

2. SIMILARITIES & DIFFERENCES

While both e-discovery and SR practitioners target exhaustive retrieval of relevant documents, the former are beginning to acknowledge the infeasibility of this goal [9] while the latter appear to believe that exhaustivity remains within reach. While both communities note the challenge of achieving complete recall in the face of vast and rapidly growing collection sizes, a typical relevance review for e-discovery (e.g., searching a corporate email archive) involves *orders of magnitude* more documents than typical SR screening. In addition, SR practice often massages the initial Boolean

ACM SIGIR 2016 Medical Information Retrieval (MedIR) Workshop. The authors retain copyright. We thank Doug Oard (UMD) for inspiration and suggestions, including for our paper's title.

[1] http://plg.uwaterloo.ca/~gvcormac/total-recall

query to return a manageable corpus that can be exhaustively searched, thereby shifting concern about sufficient recall from document screening to query formulation without actually resolving the underlying problem.

With e-discovery, the initial Boolean query may be negotiated by the two parties involved in the civil suit. Traditionally, such negotiation has sought to similarly massage the query to balance the desire for total recall against the greater cost of having to manually screen more documents (especially a concern when one or both parties have only limited resources). With the rise of TAR, however, parties are now able to cast a wider recall net with the Boolean query and rely on TAR to focus limited human screening effort where it will have the greatest impact. As a related note, the impact of such two-party negotiation on Boolean query quality has never been studied. In contrast, a SR team formulates their query internally; while a pharmaceutical company could hypothetically lobby for inclusion/exclusion of certain studies, this would be highly unusual.

While both e-discovery and systematic review seek exhaustive retrieval, the question of how much is "enough" is at least in part legally-determined for e-discovery by established law or judicial decree in a given litigation [3], hinging on effort required of additional document review vs. estimated responsiveness rate and expected impact on legal proceedings. In contrast, while systematic reviews seek comprehensive coverage of the published literature in a given area to support meta-analysis, there are professional guidelines but no medical liability or law characterizing "good" vs. "bad" systematic reviews or legal consequences.

As noted, both domains must account for fallible humans in document review. However, systematic reviewers tend to achieve comparatively high inter-annotator agreement [6], and errors that are made tend to be in one direction (i.e., false positives, which are later culled upon further assessment). This may stem from having a more narrowly-defined scientific domain and set of inclusion criteria for determining relevance, with less variation in types of documents being reviewed, medical training of the screeners, and common, recurring categories of inclusion criteria for relevance.

We must also consider the context in which document review occurs in defining appropriate evaluation methodology for screening. For example, do we assume the process allows automated methods to screen out documents without any human review, thereby incurring no cost, but possibly resulting in false negatives? Do we allow documents to be similarly screened-in (for free) without any manual review? Both domains require further manual analysis of documents once screened in: reviewing responsive documents for privilege in e-discovery, and information extraction in systematic review for statistically synthesis. How should these subsequent stages impact evaluation of document screening?

3. A JOINT RESEARCH AGENDA

As we have noted, both domains have traditionally followed a simple 3-stage pipeline: (1) Boolean search, (2) first-pass screening, and (3) final review and use. Such staging is inherently lossy: later stages cannot recover documents missed by earlier stages. A joint process combining stages would not be so limited. The process ought to require human reviewing cost $C = aR + b$ proportional to the number of relevant documents R, plus some fixed overhead b [2]. $a < 2$ and $b < 1000$ seem readily achievable, and even combining only stages 1&2 or 2&3 would still be a huge win.

In addition to reducing the number of documents to be manually screened and picking those documents intelligently, we might also investigate new human labor models for manual review. For example, SR research is already exploring use of volunteer screeners (e.g., taskexchange.cochrane.org) and paid laypeople [7]. In fact, crowd screening decisions have been found to correlate reasonably well with expert judgments at far lower cost. While e-discovery's review for responsiveness is often outsourced today, we know of no research on crowdsourcing responsiveness review.

Though crowdsourcing offers new savings and scalability, it also poses new risks. When documents to review contain confidential information (e-discovery), one must supply one's own trusted crowd or somehow automatically redact sensitive information before manual document review. In addition, little research has investigated the potential risk of an adversary organizing worker attacks to manipulate system outcomes [5]. In a truly adversarial setting in which one party seeks to manipulate document review outcomes to miss important documents or include documents supporting a particular position, crowdsourcing offers a new vulnerability to exploit, as is already being done in other areas [12].

REFERENCES

[1] Hilda Bastian, Paul Glasziou, and Iain Chalmers. "Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?" In: *PLoS medicine* 7.9 (2010), e1000326.

[2] Gordon V Cormack and Maura R Grossman. "Engineering Quality and Reliability in Technology-Assisted Review". In: *SIGIR*. 2016.

[3] Gordon V Cormack and Maura R Grossman. "Evaluation of machine-learning protocols for technology-assisted review in electronic discovery". In: *SIGIR*. 2014, pp. 153–162.

[4] Maura R Grossman, Gordon V Cormack, Bruce Hedin, and Douglas W Oard. "Overview of the TREC 2011 Legal Track." In: *20th Text Retrieval Conference (TREC)*. 2012.

[5] Walter S Lasecki, Jaime Teevan, and Ece Kamar. "Information extraction and manipulation threats in crowd-powered systems". In: *Proc. ACM CSCW*. 2014, pp. 248–256.

[6] Farrah J Mateen, Jiwon Oh, Ana I Tergas, Neil H Bhayani, and Biren B Kamdar. "Titles versus titles and abstracts for initial screening of articles for systematic reviews". In: *Clinical epidemiology* 5 (2013), pp. 89–95.

[7] An Thanh Nguyen, Byron C. Wallace, and Matthew Lease. "Combining Crowd and Expert Labels using Decision Theoretic Active Learning". In: *Proc. of the AAAI Conference on Human Computation (HCOMP)*. 2015, pp. 120–129.

[8] Douglas W Oard and William Webber. "Information retrieval for e-discovery". In: *Foundations and Trends in Information Retrieval* 7.2-3 (2013), pp. 99–237.

[9] George L Paul and Jason R Baron. "Information inflation: Can the legal system adapt?" In: *Richmond Journal of Law & Technology* 13 (2007), pp. 10–17.

[10] Byron C Wallace, Issa J Dahabreh, Christopher H Schmid, Joseph Lau, and Thomas A Trikalinos. "Modernizing the systematic review process to inform comparative effectiveness: tools and methods". In: *Journal of comparative effectiveness research* 2.3 (2013), pp. 273–282.

[11] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. "Active learning for biomedical citation screening". In: *Proceedings of ACM KDD*. 2010, pp. 173–182.

[12] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y Zhao. "Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers". In: *23rd USENIX Security Symposium (USENIX Security 14)*. 2014, pp. 239–254.