Information Retrieval for E-Discovery
Information Retrieval for E-Discovery

Douglas W. Oard
University of Maryland
College Park, MD 20742
USA
oard@umd.edu

William Webber
University of Maryland
College Park, MD 20742
USA
wew@umd.edu

Full text available at: http://dx.doi.org/10.1561/1500000025
Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2013, Volume 7, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
Information Retrieval for E-Discovery

Douglas W. Oard\textsuperscript{1} and William Webber\textsuperscript{2}

\textsuperscript{1} College of Information Studies and UMIACS, University of Maryland, College Park, MD 20742, USA, oard@umd.edu
\textsuperscript{2} College of Information Studies, University of Maryland, College Park, MD 20742, USA, wew@umd.edu

Abstract

E-discovery refers generally to the process by which one party (for example, the plaintiff) is entitled to “discover” evidence in the form of “electronically stored information” that is held by another party (for example, the defendant), and that is relevant to some matter that is the subject of civil litigation (that is, what is commonly called a “lawsuit”). This survey describes the emergence of the field, identifies the information retrieval issues that arise, reviews the work to date on this topic, and summarizes major open issues.
# Contents

1 Introduction 1

2 The E-Discovery Process 5
   2.1 Civil Discovery 5
   2.2 The Rise of E-Discovery 10
   2.3 The EDRM Reference Model 14
   2.4 An IR-Centric E-Discovery Process Model 18
   2.5 For Further Reading 25

3 Information Retrieval for E-Discovery 27
   3.1 Defining the Unit of Retrieval 28
   3.2 Extraction of Embedded Content 31
   3.3 Representation 31
   3.4 Specification 37
   3.5 Classification 39
   3.6 Clustering 41
   3.7 Other E-Discovery Tasks 45
   3.8 For Further Reading 50

4 Evaluating E-Discovery 53
   4.1 Evaluation Methods and Metrics 54
   4.2 Sampling and Estimation 62
4.3 Measurement Error
4.4 For Further Reading

5 Experimental Evaluation
5.1 Test Collection Design
5.2 The TREC Legal Track
5.3 Other Evaluation Venues
5.4 Results of Research on Test Collection Design
5.5 Research on System and Process Design
5.6 For Further Reading

6 Looking to the Future
6.1 Some Important Things We Don’t Yet Know
6.2 Some Prognostications
6.3 For Further Reading

7 Conclusion

A Interpreting Legal Citations
A.1 Case Law
A.2 Statutes and Rules
A.3 Other Court Documents

Acknowledgments

Notations and Acronyms

References
Regular viewers of the mid-twentieth century courtroom drama *Perry Mason* might be surprised to learn that the Fifth Amendment right against self-incrimination enshrined in the U.S. Constitution applies only to criminal law. In civil law, it is the obligation of parties to a lawsuit to provide documents to the other side that are responsive to proper requests and that are not subject to a claim of privilege (e.g., attorney-client privilege) \[121\]. In the law, this process is called “civil discovery,” and the resulting transfer of documents is called “production.” Amendments to the Federal Rules of Civil Procedure in 2006 made it clear that the scope of civil discovery encompasses all “Electronically Stored Information” (ESI), and thus was born the rapidly growing field that has come to be called “e-discovery” (the discovery of ESI, or Electronic Discovery) \[24\].

A confluence of interest between those working on e-discovery and those working on information retrieval was evident from the outset, although it has taken some time for the key issues to come into sharp focus. E-discovery applications of information retrieval technology are marked by five key challenges. First, e-discovery emphasizes fixed result sets rather than ranked retrieval. Second, e-discovery focuses on high
Introduction

recall, even in large collections, in contrast to the high-precision focus of many end-user applications, such as Web search. Third, e-discovery evaluation must measure not just relative, but also absolute effectiveness. Fourth, e-discovery connects information retrieval with techniques and concerns from other fields (for instance, computer forensics and document management). And fifth, the adversarial nature of civil litigation, and the information asymmetry between requesting party (who makes the request) and responding party (who has the documents), makes e-discovery a substantially arms-length transaction.

While these challenges are not unique to e-discovery, the demands of the e-discovery marketplace has focused research upon them. The market for vendors of e-discovery systems has been estimated at US$ 1 billion in 2010 [87]; several times that figure are spent on the staffing and processing costs to use those systems effectively [108]. In view of these large costs, information retrieval research can help to achieve two important societal goals: (1) improving the return on this investment by enhancing the effectiveness of the process for some given level of human effort (which has important implications for the fairness of the legal system), and (2) reducing future costs (which has important implications for broad access to the legal system by potential litigants).

Furthermore, fundamental technologies developed for e-discovery may have applications in other fields as well. For example, the preparation of systematic reviews of recent research on specific topics in medicine might benefit from advances in high-recall search [67], and personal information management might benefit from advances in search technology that focus specifically on e-mail (which at present is of particular interest in operational e-discovery settings).

With that background in mind, the remainder of this survey is organized as follows. Section 2 on The E-Discovery Process begins with an introduction to the structure of the process of e-discovery, focusing principally on U.S. federal law, but with a brief survey of discovery practice in other jurisdictions. The part of the e-discovery process known as “document review” has been the focus of the greatest investment [108] and is therefore our central focus in this review. The section also introduces the three canonical information seeking processes (linear review, keyword search, and technology-assisted review) that shape current
practice in document review. Section 3 on *Information Retrieval for E-Discovery* examines specific techniques that have been (or could be) applied in e-discovery settings. Section 4 on *Evaluating E-Discovery* discusses evaluation issues that arise in e-discovery, focusing in detail on set-based evaluation, estimation of effectiveness metrics, computation of confidence intervals, and challenges associated with developing absolute as well as relative measures. Section 5 on *Experimental Evaluation* reviews the principal venues in which e-discovery technology has been examined, both those well known in academic research (such as the Legal Track of the Text Retrieval Conference (TREC)), and those more familiar to industry (e.g., the Data Set project of the Electronic Discovery Reference Model (EDRM) organization). Section 6 on *Looking to the Future* draws on our description of the present state of the art to identify important and as yet unresolved issues that could benefit from future information retrieval research. Finally, Section 7, the *Conclusion*, draws together some broader implications of work on e-discovery.
References


References


References


References


References


References


140 References


Full text available at: http://dx.doi.org/10.1561/1500000025


References


References


