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

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is D. W. Oard and W. Webber, Information Retrieval for E-Discovery, Foundation and Trends® in Information Retrieval, vol 7, nos 2–3, pp 99–237, 2013

ISBN: 978-1-60198-678-8

Foundations and Trends® in Information Retrieval
Volume 7 Issues 2–3, 2013

Editorial Board

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.

Foundations and Trends® in
Information Retrieval
Vol. 7, Nos. 2–3 (2013) 99–237

DOI: 10.1561/1500000025

now
the essence of knowledge

Information Retrieval for E-Discovery

Douglas W. Oard[1] and William Webber[2]

[1] *College of Information Studies and UMIACS, University of Maryland, College Park, MD 20742, USA, oard@umd.edu*
[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 75
4.4 For Further Reading 79

5 Experimental Evaluation 81

5.1 Test Collection Design 82
5.2 The TREC Legal Track 84
5.3 Other Evaluation Venues 93
5.4 Results of Research on Test Collection Design 95
5.5 Research on System and Process Design 101
5.6 For Further Reading 110

6 Looking to the Future 113

6.1 Some Important Things We Don't Yet Know 113
6.2 Some Prognostications 117
6.3 For Further Reading 118

7 Conclusion 119

A Interpreting Legal Citations 123

A.1 Case Law 124
A.2 Statutes and Rules 125
A.3 Other Court Documents 126

Acknowledgments 129

Notations and Acronyms 131

References 133

1

Introduction

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

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

[1] R. Agrawal, H. Mannila, R. Srikant, H. Tolvonen, and A. I. Verkamo, "Fast discovery of association rules," *Advances in Knowledge Discovery and Data Mining*, vol. 12, pp. 307–328, 1996.

[2] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, pp. 119–126, 1998.

[3] J. Aslam and V. Pavlu, "A practical sampling strategy for efficient retrieval evaluation," Technical report, Northeastern University, 2008.

[4] J. Aslam, V. Pavlu, and E. Yilmaz, "A statistical method for system evaluation using incomplete judgments," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, eds.), Washington, USA: Seattle, pp. 541–548, 2006.

[5] S. Attfield and A. Blandford, "Discovery-led refinement in e-discovery investigations: Sensemaking, cognitive ergonomics and system design," *Artificial Intelligence and Law*, vol. 18, pp. 387–412, 2010.

[6] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz, "Relevance assessment: Are judges exchangeable and does it matter?," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds.), Singapore, pp. 667–674, 2008.

[7] P. Baldi, S. Brunak, Y. Chavin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, pp. 412–424, 2000.

[8] S. Bales and P. Wang, "Consolidating user relevance criteria: A meta-ethnography of empirical studies," in *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, 2006.

[9] K. Balog, "People search in the enterprise," PhD thesis, University of Amsterdam, 2008.

[10] T. Barnett, S. Godjevac, J.-M. Renders, C. Privault, J. Schneider, and R. Wickstrom, "Machine learning classification for document review," in *DESI III: The ICAIL Workshop on Globaal E-Discovery/E-Disclosure*, 2009.

[11] J. R. Baron, "The TREC Legal Track: Origins and reflections on the first year," *The Sedona Conference Journal*, vol. 8, 2007.

[12] J. R. Baron, "Towards a new jurisprudence of information retrieval: What constitutes a 'reasonable' search for digital evidence when using keywords?," *Digital Evidence and Electronic Signature Law Review*, vol. 5, pp. 173–178, 2008.

[13] J. R. Baron, "E-discovery and the problem of asymmetric knowledge," *Mercer Law Review*, vol. 60, p. 863, 2009.

[14] J. R. Baron, "Law in the age of exabytes: Some further thoughts on 'information inflation' and current issues in e-discovery search," *Richmond Journal of Law and Technology*, vol. 17, 2011.

[15] J. R. Baron, D. D. Lewis, and D. W. Oard, "TREC-2006 legal track overview," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-272, pp. 79–98, 2006.

[16] J. R. Baron, D. W. Oard, T. Elsayed, and L. Wang, "No, not that pmi: Creating search technology for e-discovery," Powerpoint slides, 2008.

[17] L. J. Barris, *Understanding and Mastering The Bluebook*. Carolina Academic Press, 2nd ed., 2010.

[18] R. S. Bauer, D. Brassil, C. Hogan, G. Taranto, and J. S. Brown, "Impedance matching of humans and machines in high-q information retrieval systems," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 97–101, 2009.

[19] S. Bennett and S. Millar, "Multinationals face e-discovery challenges," *International Financial Law Review*, vol. 25, pp. 37–39, 2006.

[20] M. D. Berman, C. I. Barton, and P. W. Grimm, eds., *Managing E-Discovery and ESI: From Pre-Litigation Through Trial*. American Bar Association, 2011.

[21] W. E. Bijker, T. P. Hughes, and T. J. Pinch, eds., *The Social Construction of Technological Systems: New Directions in the Sociology of History and Technology*. MIT Press, 1987.

[22] D. 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, 1985.

[23] B. B. Borden, *E-discovery Alert: The Demise of Linear Review*. 2010.

[24] B. B. Borden, M. McCarroll, B. C. Vick, and L. M. Wheeling, "Four years later: How the 2006 amendments to the federal rules have reshaped the e-discovery landscape and are revitalizing the civil justice system," *Richmond Journal of Law and Technology*, vol. 17, 2011.

[25] D. Brassil, C. Hogan, and S. Attfield, "The centrality of user modeling to high recall with high precision search," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 91–96, 2009.

[26] K. R. W. Brewer and M. Ḥanīf, *Sampling with Unequal Probabilities*. Springer, 1983.

[27] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Annual Symposium on Combinatorial Pattern Matching*, pp. 1–10, 2000.

[28] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical Science*, vol. 18, pp. 101–133, 2001.

[29] C. Buckley and E. Voorhees, "Retrieval system evaluation," in *Voorhees and Harman, Chapter 3*, 2005.

[30] J. P. Buonaccorsi, *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC, 2010.

[31] J. L. Carroll, "Proportionality in discovery: A cautionary tale," *Campbell Law Review*, vol. 32, pp. 455–466, 2010.

[32] B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds.), Singapore, pp. 651–658, 2008.

[33] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania, "Efficient techniques for document sanitization," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pp. 843–852, 2008.

[34] D. T. Chaplin, "Conceptual search — ESI, litigation and the issue of language," in *DESI II: Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*, London, UK, 2008.

[35] H. Chu, "Factors affecting relevance judgment: A report from TREC legal track," *Journal of Documentation*, vol. 67, pp. 264–278, 2011.

[36] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds.), Singapore, pp. 659–666, 2008.

[37] C. W. Cleverdon, "The Cranfield tests on index language devices," *Aslib Proceedings*, vol. 19, pp. 173–192, 1967.

[38] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404–413, 1934.

[39] W. G. Cochran, *Sampling Techniques*. John Wiley & Sons, 3rd ed., 1977.

[40] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[41] J. G. Conrad, "E-discovery revisited: The need for artificial intelligence beyond information retrieval," *Artificial Intelligence and Law*, vol. 18, pp. 321–345, 2010.

[42] G. V. Cormack, M. R. Grossman, B. Hedin, and D. W. Oard, "Overview of the TREC 2010 Legal Track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, pp. 1:2:1–45, 2010.

[43] T. Curtis, "Declassification overview," in *Proceedings of the 1997 Symposium on Document Image Understanding Technology*, (D. Doermann, ed.), p. 39, 1997.

[44] B. Dervin and L. Foreman-Wernet, eds., *Sense-Making Methodology Reader: Selected writings of Brenda Dervin*. Hampton Press, 2003.

[45] D. Doermann, "The indexing and retrieval of document imgaes: A survey," *Computer Vision and Image Understanding*, vol. 70, pp. 287–298, 1998.

[46] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of ACM International Conference on Information and Knowledge Management*, pp. 148–155, 1998.

[47] S. T. Dumais, T. Joachims, K. Bharat, and A. S. Weigend, "Sigir 2003 workshop report: Implicit measures of user interests and preferences," *SIGIR Forum*, vol. 37, pp. 50–54, 2003.

[48] E. N. Efthimiadis and M. A. Hotchkiss, "Legal discovery: Does domain expertise matter?," *Proceedings of American Society on Information Science Technology*, vol. 45, pp. 1–2, 2008.

[49] D. Eichmann and S.-C. Chin, "Concepts, semantics and syntax in e-discovery," in *DESI I: The ICAIL Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*, Palo Alto, CA, USA, 2007.

[50] T. Elsayed, D. W. Oard, and G. Namata, "Resolving personal names in email using context expansion," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 941–949, 2008.

[51] J. M. Facciola and J. M. Redgrave, "Asserting and challenging privilege claims in modern litigation: The Facciola-Redgrave framework," *The Federal Courts Law Review*, vol. 4, pp. 19–54, 2009.

[52] S. Fischer, R. E. Davis, and M. D. Berman, "Gathering, reviewing, and producing esi: An eight-stage process," in *Berman et al. eds., Chapter 14*, 2011.

[53] D. C. Force, "From Peruvian Guano to electronic records: Canadian e-discovery and records professionals," *Archivaria*, vol. 69, pp. 1–27, 2010.

[54] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, "INEX: INitiative for the Evaluation of XML Retrieval," in *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, pp. 1–9, 2002.

[55] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database Systems — The Complete Book*. Pearson Education, 2nd ed., 2009.

[56] C. Görg and J. Stasko, "Jigsaw: Investigative analysis on text document collections through visualization," in *DESI II: Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*, London, UK, 2008.

[57] P. W. Grimm, L. Yurmit, and M. P. Kraeuter, "Federal rule of evidence 502: Has it lived up to its potential?," *Richmond Journal of Law and Technology*, vol. 17, 2011.

[58] M. R. Grossman and G. V. Cormack, "Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review," *Richmond Journal of Law and Technology*, vol. 17, pp. 11:1–48, 2011.

[59] M. R. Grossman and G. V. Cormack, "Inconsistent assessment of responsiveness in e-discovery: Difference of opinion or human error?," in *DESI IV: The ICAIL Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, Pittsburgh, PA, USA, pp. 1–11, 2011.

[60] M. R. Grossman, G. V. Cormack, B. Hedin, and D. W. Oard, "Overview of the TREC 2011 Legal Track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, p. 20, 2011.

[61] S. Harter, *Online Information Retrieval: Concepts, principles, and Techniques*. Academic Press, 1986.

[62] Harvard Law Review, *The Bluebook: A Uniform System of Citation*. The Harvard Law Review Association, 19th ed., 2010.

[63] B. Hedin and D. W. Oard, "Replication and automation of expert judgments: Information engineering in legal e-discovery," in *SMC'09: Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, pp. 102–107, 2009.

[64] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard, "Overview of the TREC 2009 Legal Track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-278, pp. 1:4:1–40, 2009.

[65] H. Henseler, "Network-based filtering for large email collections in e-discovery," in *DESI III: The ICAIL Workshop on Globaal E-Discovery/ E-Disclosure*, 2009.

[66] R. J. Heuer, *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, 1999.

[67] J. P. T. Higgins and S. Green, *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, 2008.

[68] T. C. Hoad and J. Zobel, "Methods for identifying versioned and plagiarized documents," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 203–215, 2003.

[69] C. Hogan, R. S. Bauer, and D. Brasil, "Automation of legal sensemaking in e-discovery," *Artificial Intelligence and Law*, vol. 18, pp. 431–457, 2010.

[70] C. Hogan, D. Brassil, S. M. Rugani, J. Reinhart, M. Gerber, and T. Jade, "H5 at TREC 2008 legal interactive: User modeling, assessment and measurement," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-277, pp. 2:18:1–9, 2008.

[71] B. J. Jansen, A. Spink, and I. Taksa, eds., *Handbook of Research on Web Log Analysis*. Information Science Reference, 2009.

[72] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (E. Yannakoudis, N. J. Belkin, M.-K. Leong, and P. Ingwersen, eds.), Athens, Greece, pp. 41–48, 2000.

[73] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of European Conference on Machine Learning*, (C. Nédellec and C. Rouveirol, eds.), pp. 137–142, 1998.

[74] S. Joshi, D. Contractor, K. Ng, P. M. Deshpande, and T. Hampp, "Auto-grouping emails for fast e-discovery," *Proceedings of VLDB Endowment*, vol. 4, pp. 1284–1294, 2011.

[75] S. Joty, G. Carenini, G. Murray, and R. T. Ng, "Exploiting conversation structure in unsupervised topic segmentation for emails," in *Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts, USA, pp. 388–398, 2010.

[76] P. Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval*, vol. 1, pp. 233–334, 2006.

[77] L. Katz, "Confidence intervals for the number showing a certain characteristic in a population when sampling is without replacement," *Journal of the American Statistical Association*, vol. 48, pp. 256–261, 1953.

[78] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds., *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010.

[79] J. Kekäläinen and K. Järvelin, "Using graded relevance assessments in IR evaluation," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 1120–1129, 2002.

[80] A. Kershaw and J. Howie, *Report on Kershaw–Howie Survey of E-discovery Providers Pertaining to Deduping Strategies*. 2009.

[81] A. Kershaw and J. Howie, "Exposing context: Email threads reveal the fabric of conversations," *Law Technology News*, 2010.

[82] S. Kiritchenko, S. Matwin, and S. Abu-hakima, "Email classification with temporal features," in *Proceedings of International Intelligent Information Systems*, 2004.

[83] B. Klimt and Y. Yang, "Introducing the Enron corpus," in *Proceedings of Conference on Email and Anti-Spam*, Mountain View, CA, USA, p. 2, 2004.

[84] R. Laplanche, J. Delgado, and M. Turck, "Concept search technology goes beyond keywords," *Information Outlook*, vol. 8, 2004.

[85] V. L. Lemieux and J. R. Baron, "Overcoming the digital tsunami in e-discovery: Is visual analytics the answer?," *Canadian Journal of Law and Technology*, vol. 9, pp. 33–50, 2011.

[86] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (W. B. Croft and C. J. van Rijsbergen, eds.), Dublin, Ireland, pp. 3–12, 1994.

[87] D. Logan and S. Childs, "Magic quadrant for e-discovery software," Gartner Research Report, 2012.

[88] T. R. Lynam and G. V. Cormack, "MultiText Legal experiments at TREC 2008," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-277, pp. 2:56:1–5, 2008.

[89] D. A. Mackenzie and J. Wajeman, eds., *The Social Shaping of Technology*. McGraw Hill, 1985.

[90] W. Magdy and G. J. F. Jones, "PRES: A score metric for evaluating recall-oriented information retrieval applications," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (H.-H. Chen, E. N. Efthimiadis, J. Savoy, F. Crestani, and S. Marchand-Maillet, eds.), Geneva, Switzerland, pp. 611–618, 2010.

[91] R. Manmatha, C. Han, and E. Riseman, "Word spotting: A new approach to indexing handwriting," in *Proceedings of 1996 IEEE Computer Science Conferen on Computer Vision and Pattern Recognition*, pp. 631–637, 1996.

[92] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[93] R. L. Marcus, "E-discovery & beyond: Toward *Brave New World* or *1984*?," *The Review of Litigation*, vol. 25, pp. 633–689, 2006.

[94] K. Marley and P. Cochrane, Online training and practice manual for ERIC database searchers, *ERIC Clearinhouse on Information Resources*. 2nd ed., 1981.

[95] S. Martin, A. Sewani, B. Nelson, K. Chen, and A. D. Joseph, "Analyzing behavioral features for email classification," in *Proceedings of Conference on Email and Anti-Spam*, p. 8, 2005.

[96] J. McGann, "Lesson from the news of the world scandal: Data is forever," *Forbes*, 2010.

[97] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 873–889, 2001.

[98] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transactions on Information Systems*, vol. 27, pp. 1–27, 2008.

[99] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 36, pp. 354–359, 1983.

[100] S. D. Nelson and J. Simek, "Technology tips for cutting e-discovery costs," *Information Management Journal*, vol. 43, 2009.

[101] D. W. Oard, "Multilingual information access," in *Encyclopedia of Library and Information Sciences*, (M. J. Bates and M. N. Maack, eds.), Taylor and Francis, 2009.

[102] D. W. Oard, J. R. Baron, B. Hedin, D. D. Lewis, and S. Tomlinson, "Evaluation of information retrieval for E-discovery," *Artificial Intelligence and Law*, pp. 1–40, 2010. published online.

[103] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron, "Overview of the TREC 2008 legal track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-277, pp. 3:1–45, 2008.

[104] R. T. Oehrle, "Retrospective and prospective statistical sampling in legal discovery," in *DESI IV: The ICAIL Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, Pittsburgh, PA, USA, 2011.

[105] J. S. Olsson and D. W. Oard, "Combining evidence from lvcsr and ranked utterance retrieval for robust domain-specific ranked retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 91–98, 2009.

[106] J. O'Neill, C. Privault, J.-M. Renders, V. Ciriza, and G. Bauduin, "DISCO: Intelligent help for document review," in *DESI III: The ICAIL Workshop on Globaal E-Discovery/E-Disclosure*, 2009.

[107] P. Oot, A. Kershaw, and H. L. Roitblat, "Mandating reasonableness in a reasonable inquiry," *Denver Law Review*, vol. 87, pp. 533–559, 2010.

[108] N. M. Pace and L. Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*. RAND, 2012.

[109] G. L. Paul and J. R. Baron, "Information inflation: Can the legal system adapt?," *Richmond Journal of Law and Technology*, vol. 13, 2007.

[110] A. Perer, B. Shneiderman, and D. W. Oard, "Using rhythms of relationships to understand email archives," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 1936–1948, 2006.

[111] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1998.

[112] S. Robertson, "A new interpretation of average precision," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds.), Singapore, pp. 689–690, 2008.

[113] S. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of Text REtrieval Conference*, (D. Harman, ed.), Gaithersburg, Maryland, USA: NIST Special Publication 500-225, pp. 109–126, 1994.

[114] H. L. Roitblat, A. Kershaw, and P. Oot, "Document categorization in legal electronic discovery: Computer classification vs. manual review," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 70–80, 2010.

[115] R. Russeth and S. Burns, "Why my human document reviewer is better than your algorithm," *ACC Docket: The Journal of the Association of Corporate Counsel*, vol. 28, pp. 18–31, 2010.

[116] G. Salton and R. K. Waldstein, "Term relevance weights in on-line information retrieval," *Information Processing and Management*, vol. 14, pp. 29–35, 1978.

[117] M. Sanderson, "Test collection based evaluation of information retrieval systems," *Foundations and Trends in Information Retrieval*, vol. 4, pp. 247–375, 2010.

[118] M. Sanderson and J. Zobel, "Information retrieval system evaluation: Effort, sensitivity, and reliability," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, eds.), Salvador, Brazil, pp. 162–169, 2005.

[119] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 2126–2144, 2007.

[120] A. Sayeed, S. Sarkar, Y. Deng, R. Hosn, R. Mahindru, and N. Rajamani, "Characteristics of document similarity measures for compliance analysis," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pp. 1207–1216, 2009.

[121] S. Scheindlin, D. J. Capra, and The Sedona Conference, *Electronic Discovery and Digital Evidence: Cases and Materials (American Casebook)*. West Publishers, 2nd ed., 2012.

[122] F. Scholer, A. Turpin, and M. Sanderson, "Quantifying test collection quality based on the consistency of relevance judgements," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua, and W. B. Croft, eds.), Beijing, China, pp. 1063–1072, 2011.

[123] S. Scott and S. Matwin, "Feature engineering for text classification," in *Proceedings of International Conference on Machine Learning*, pp. 379–388, 1999.

[124] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[125] C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," *Internal Journal on Document Analysis and Recognition*, vol. 3, pp. 232–247, 2001.

[126] D. L. Simel, G. P. Samsa, and D. B. Matchar, "Likelihood ratios with confidence: Sample size estimation for diagnostic test studies," *Journal of Clinical Epidemiology*, vol. 44, pp. 763–770, 1991.

[127] I. Soboroff, "A comparison of pooled and sampled relevance judgments," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, eds.), Amsterdam, The Netherlands, pp. 785–786, 2007.

[128] R. D. Solomon and J. R. Baron, "Bake offs, demos & kicking the tires: A practical litigator's brief guide to evaluating early case assessment software & search & review tools," 2009.

[129] K. Spärck Jones and J. R. Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer LNAI 1083, 1995.

[130] K. Spärck Jones and C. J. van Rijsbergen, "Report on the need for and provision of an 'ideal' test collection," Technical Report, University Computer Laboratory, Cambridge, 1975.

[131] B. Stein, M. Koppel, and E. Stamatatos, "Plagiarism analysis, authorship identification, and near-duplicate detection: Pan'07," *SIGIR Forum*, vol. 41, pp. 68–71, 2007.

[132] T. Sterenzy, "Equivio at TREC 2009 Legal Interactive," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-278, pp. 1:17:1–3, 2009.

[133] C. Stevens, "Knowledge-based assistance for handling large, poorly structured information spaces," PhD thesis, University of Colorado, 1993.

[134] S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, pp. 677–680, 1946.
[135] E. Stewart and P. N. Banks, "Preservation of information in nonpaper formats," in *Preservation: Issues and Planning*, (P. N. Banks and R. Pilette, eds.), Chapter 18. ALA Editions, 2000.
[136] A. B. Sunter, "List sequential sampling with equal or unequal probabilities without replacement," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 26, pp. 261–268, 1977.
[137] R. S. Taylor, "Process of asking questions," *American Documentation*, vol. 13, pp. 391–396, 1962.
[138] A. Tenenbein, "A double sampling scheme for estimating from binomial data with misclassifications," *Journal of the American Statistical Association*, vol. 65, pp. 1350–1361, 1970.
[139] 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*, vol. 8, pp. 189–223, 2007.
[140] The Sedona Conference, "The Sedona Guidelines: Best practice guidelines & commentary for managing information & records in the electronic age," 2007.
[141] The Sedona Conference, "The Sedona Principles, Second Edition: Best practice recommendations and principles for addressing electronic document production," 2007.
[142] The Sedona Conference, "The Sedona Canada principles: Addressing electronic discovery," 2008.
[143] The Sedona Conference, "The Sedona Conference commentary on non-party production & rule 45 subpoenas," 2008.
[144] The Sedona Conference, "The Sedona Conference commentary on preservation, management and identification of sources of information that are not reasonably accessible," 2008.
[145] The Sedona Conference, "The Sedona Conference cooperation proclamation," 2008.
[146] The Sedona Conference, "The Sedona Conference commentary on proportionality in electronic discovery," 2010.
[147] The Sedona Conference, "The Sedona Conference glossary: E-discovery & digital information management," 2010.
[148] The Sedona Conference, "The Sedona Canada commentary on practical approaches for cost containment," 2011.
[149] The Sedona Conference, "The Sedona Conference database principles: Addressing the preservation & production of databases & database information in civil litigation," *Public Comment Version*, 2011.
[150] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE Computer Graphics and Applications*, vol. 26, pp. 10–13, 2006.
[151] S. K. Thompson, *Sampling*. New York: John Wiley & Sons, 3rd ed., 2012.
[152] S. Tomlinson, "Learning task experiments in the TREC 2011 Legal Track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, p. 14, 2011.

[153] S. Tomlinson, D. W. Oard, J. R. Baron, and P. Thompson, "Overview of the TREC 2007 legal track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA: NIST Special Publication 500-274, pp. 5:1–34, 2007.

[154] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 2nd ed., 1979.

[155] E. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing & Management*, vol. 36, pp. 697–716, 2000.

[156] E. Voorhees, "The philosophy of information retrieval evaluation," in *Proceedings of Workshop of the Cross-Lingual Evaluation Forum*, vol. 2406 of *Lecture Notes in Computer Science*, (C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, eds.), Darmstadt, Germany: Springer, pp. 355–370, 2002.

[157] E. Voorhees and D. Harman, eds., *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

[158] A. Wang, "The Shazam music recognition service," *Communications of the ACM*, vol. 49, pp. 44–48, 2006.

[159] J. Wang, "Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery," in *Proceedings of SIGIR Information Retrieval for E-Discovery Workshop*, Beijing, China, pp. 1:1–10, 2011.

[160] J. Wang, C. Coles, R. Elliott, and S. Andrianakou, "ZL Technologies at TREC 2009 legal interactive: Comparing exclusionary and investigative approaches for electronic discovery using the TREC Enron corpus," in *The Text REtrieval Conference Proceedings (TREC 2009)*, 2009.

[161] J. Wang and D. Soergel, "A user study of relevance judgments for e-discovery," *Proceedings of American Society on Information Science Technology*, vol. 47, pp. 1–10, 2010.

[162] R. Warren, "University of waterloo at TREC 2011: A social networking approach to the Legal Learning track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, p. 4, 2011.

[163] W. Webber, "Re-examining the effectiveness of manual review," in *Proceedings of SIGIR Information Retrieval for E-Discovery Workshop*, Beijing, China, pp. 2:1–8, 2011.

[164] W. Webber, "Approximate recall confidence intervals," In submission, 2012.

[165] W. Webber, D. W. Oard, F. Scholer, and B. Hedin, "Assessor error in stratified evaluation," in *Proceedings of ACM International Conference on Information and Knowledge Management*, Toronto, Canada, pp. 539–548, 2010.

[166] W. Webber, F. Scholer, M. Wu, X. Zhang, D. W. Oard, P. Farrelly, S. Potter, S. Dick, and P. Bertolus, "The Melbourne team at the TREC 2010 legal track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, pp. 49:1–12, 2010.

[167] W. Webber, B. Toth, and M. Desamito, "Effect of written instructions on assessor agreement," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, eds.), Portland, Oregon, USA, 2012. to appear.

[168] T. D. Wickens, *Elementary Signal Detection Theory*. Oxford University Press, 2002.

[169] T. Wilson, "Models in information behaviour research," *Journal of Documentation*, vol. 55, pp. 249–270, 1999.

[170] E. Yilmaz and J. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of ACM International Conference on Information and Knowledge Management*, (P. S. Yu, V. Tsotras, E. A. Fox, and B. Liu, eds.), Arlington, Virginia, USA, pp. 102–111, 2006.

[171] P. Zeinoun, A. Laliberte, J. Puzicha, H. Sklar, and C. Carpenter, "Recommind at TREC 2011 Legal Track," in *Proceedings of Text REtrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Gaithersburg, Maryland, USA, p. 12, 2011.

[172] J. Zhang, Y. Yong, and M. Lades, "Face recognition: Eigenface, elastic matching, and neural nets," *Proceedings of the IEEE*, vol. 85, pp. 1423–1435, 1997.

[173] F. C. Zhao, D. W. Oard, and J. R. Baron, "Improving search effectiveness in the legal e-discovery process using relevance feedback," in *ICAIL 2009 DESI III Global E-Discovery/E-Disclosure Workshop*, 2009.

[174] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger, "Signature detection and matching for document image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[175] J. Zobel, "How reliable are the results of large-scale information retrieval experiments?," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, eds.), Melbourne, Australia, pp. 307–314, 1998.