Test Collection Based Evaluation of Information Retrieval Systems

Test Collection Based Evaluation of Information Retrieval Systems

Mark Sanderson

The Information School
University of Sheffield
Sheffield
UK
m.sanderson@shef.ac.uk

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 M. Sanderson, Test Collection Based Evaluation of Information Retrieval Systems, Foundation and Trends® in Information Retrieval, vol 4, no 4, pp 247–375, 2010

ISBN: 978-1-60198-360-2

Foundations and Trends® in Information Retrieval

Volume 4 Issue 4, 2010

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, 2010, Volume 4, 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. 4, No. 4 (2010) 247–375
© 2010 M. Sanderson
DOI: 10.1561/1500000009

now
the essence of knowledge

Test Collection Based Evaluation of Information Retrieval Systems

Mark Sanderson

The Information School, University of Sheffield, Sheffield, UK
m.sanderson@shef.ac.uk

Abstract

Use of test collections and evaluation measures to assess the effectiveness of information retrieval systems has its origins in work dating back to the early 1950s. Across the nearly 60 years since that work started, use of test collections is a de facto standard of evaluation. This monograph surveys the research conducted and explains the methods and measures devised for evaluation of retrieval systems, including a detailed look at the use of statistical significance testing in retrieval experimentation. This monograph reviews more recent examinations of the validity of the test collection approach and evaluation measures as well as outlining trends in current research exploiting query logs and live labs. At its core, the modern-day test collection is little different from the structures that the pioneering researchers in the 1950s and 1960s conceived of. This tutorial and review shows that despite its age, this long-standing evaluation method is still a highly valued tool for retrieval research.

Contents

1

Introduction

An examination of the opening pages of a number of Information Retrieval (IR) books reveals that each author defines the topic of IR in different ways. Some say that IR is simply a field concerned with organizing information [210]; and others emphasize the range of different materials that need to be searched [286]. While others stress the contrast between the strong structure and typing of a database (DB) system with the lack of structure in the objects typically searched in IR [262, 244]. Across all of these definitions, there is a constant, IR systems have to deal with incomplete or *underspecified* information in the form of the queries issued by users. The IR systems receiving such queries need to fill in the gaps of the users' underspecified query.

For example, a user typing "nuclear waste dumping" into the search engine of an academic repository is probably looking for multiple documents describing this topic in detail, he/she probably prefers to see documents from reputable sources, but all he/she enters into the search engine are three words. Users querying on a web search engine for "BBC" are probably looking for the official home page of the corporation, yet they fully expect the search engine to infer that specific information request from the three letters entered. The fact that the content being searched is typically unstructured and its components (i.e., words) can have multiple senses, and different words can be used to express the same concept, merely adds to the challenge of locating relevant items. In contrast to a DB system, whose search outputs are deterministic, the accuracy of an IR system's output cannot be predicted with any confidence prior to a search being conducted; consequently, empirical evaluation has always been a critical component of Information Retrieval.[1]

The typical interaction between a user and an IR system has the user submitting a query to the system, which returns a ranked list of objects that hopefully have some degree of relevance to the user's request with the most relevant at the top of the list. The success of such an interaction is affected by many factors, the range of which has long been considered. For example, Cleverdon and Keen [61, p. 4] described five.

1. *"The ability of the system to present all relevant documents*
2. *The ability of the system to withhold non-relevant documents*
3. *The interval between the demand being made and the answer being given (i.e., time)*
4. *The physical form of the output (i.e., presentation)*
5. *The effort, intellectual or physical, demanded of the user (i.e., effort)."*

To this list one could add many others, e.g.:

- the ability of the user at specifying their need;
- the interplay of the components of which the search algorithm is composed;
- the type of user information need;
- the number of relevant documents in the collection being searched;
- the types of documents in the collection;

[1] This is not to say that researchers haven't tried to devise non-empirical approaches, such as building theoretical models of IR systems. However, Robertson [197] points out that a theory of IR that would allow one to predict performance without evaluation remains elusive.

- the context in which the user's query was issued; and
- the eventual use for the information being sought.

Evaluation of IR systems is a broad topic covering many areas including information-seeking behavior usability of the system's interface; its broader contextual use; the compute efficiency, cost, and resource needs of search engines. A strong focus of IR research has been on measuring the *effectiveness* of an IR system: determining the *relevance* of items, retrieved by a search engine, relative to a user's information need.

The vast majority of published IR research assessed effectiveness using a resource known as a *test collection* used in conjunction with *evaluation measures*. Such is the importance of test collections that at the time of writing, there are many conferences and meetings devoted purely to their use: including three international conferences, TREC, CLEF, and NTCIR, which together have run more than 30 times since the early 1990s. This research focus is not just a feature of the past two decades but part of a longer tradition which was motivated by the creation and sharing of testing environments in the previous three decades, which itself was inspired by innovative work conducted in the 1950s. The classic components of a test collection are as follows:

- a collection of documents; each document is given a unique identifier, a *docid*;
- a set of topics (also referred to as queries); each given a query id (*qid*); and
- a set of *relevance judgments* (often referred to as *qrels* — query relevance set) composed of a list of qid/docid pairs, detailing the relevance of documents to topics.

In the possession of an appropriate test collection, an IR developer or researcher simply loads the documents into their system and in a batch process, submits the topics to the system one-by-one. The list of the docids retrieved for each of the topics is concatenated into a set, known as a *run*. Then the content of the run is examined to determine which of the documents retrieved were present in the qrels and which were not. Finally, an evaluation measure is used to quantify the effectiveness of that run.

Together, the collection and chosen evaluation measure provide a *simulation* of users of a searching system in an operational setting. Using test collections, researchers can assess a retrieval system in isolation helping locate points of failure, but more commonly, collections are used to compare the effectiveness of multiple retrieval systems. Either rival systems are compared with each other, or different configurations of the same system are contrasted. Such determinations, by implication, predict how well the retrieval systems will perform relative to each other if they were deployed in the operational setting simulated by the test collection.

A key innovation in the IR academic community was the early recognition of the importance of building and crucially sharing test collections.[2] Through sharing, others benefited from the initial (substantial) effort put into the creation of a test collection by re-using it in other experiments. Groups evaluating their own IR systems on a shared collection could make meaningful comparisons with published results tested on the same collection. Shared test collections provided a focus for many international collaborative research exercises. Experiments using them constituted the main methodology for validating new retrieval approaches. In short, test collections are a catalyst for research in the IR community.

Although there has been a steady stream of research in evaluation methods, there has been little survey of literature covering test collection based evaluation. Salton's evaluation section [210, Section 5] is one such document; a chapter in Van Rijsbergen's book [262] another; Spärck Jones's edited articles on IR experiments [242] a third. Since those works, no broad surveys of evaluation appear to have been written; though Hearst has recently written about usability evaluation in IR [116, Section 3]. The sections on evaluation in recent IR books provided the essential details on how to conduct evaluation, rather than reviewed

[2] Indeed, it would appear that the academic IR community is one of the first in the Human Language Technologies (HLT) discipline of computer science to create and share common testing environments. Many other areas of HLT, such as summarization, or word sense disambiguation did not start building such shared testing resources until the 1990s.

past work. There are notable publications addressing particular aspects of evaluation: Voorhees and Harman's book detailed the history of the TREC evaluation exercise and outlined evaluation methods used [280]; a special issue of Information Processing and Management reflected the state of IR evaluation in 1992 [98]; another special issue in the *Journal of the American Society for Information Science* provided a later perspective [253]. More recently, Robertson published his personal view on the history of IR evaluation [199]. However, there remains a gap in the literature, which this monograph attempts to fill.

Using test collections to assess the effectiveness of IR systems is itself a broad area covering a wide range of document types and forms of retrieval. IR systems were built to search over text, music, speech, images, video, chemical structures, etc. For this monograph, we focus on evaluation of retrieval from documents that are searched by their text content and similarly queried by text; although, many of the methods described are applicable to other forms of IR.

Since the initial steps of search evaluation in the 1950s, test collections and evaluation measures were developed and adapted to reflect the changing priorities and needs of IR researchers. Often changes in test collection design caused changes in evaluation measures and vice versa. Therefore, the work in these two distinct areas of study are described together and laid out in a chronological order. The research is grouped into three periods, which are defined relative to the highly important evaluation exercise, TREC.

- **Early 1950s–early1990s**, Section 2: the initial development of test collections and measures. In this time, test collection content was mostly composed of catalogue information about academic papers or later the full-text of newspaper articles. The evaluation measures commonly used by researchers were strongly focused on *high recall* search: finding as many relevant items as possible.
- **Early 1990s–early 2000s**, Section 3: the "TREC ad hoc" period. Scale and standardization of evaluation were strong themes of this decade. The IR research community collaborated to build a relatively small number of large test

collections mainly composed of news articles. Evaluation was still focused on high recall search.

- **Early 2000s–present**, Section 4: the post ad hoc period (for want of a better name). Reflecting the growing diversity in application of search technologies and the ever-growing scale of collections being searched, evaluation research in this time showed a diversification of content and search task along with an increasing range of evaluation measures that reflected user's more common preference for finding a small number of relevant items. Run data gathered by TREC and other similar exercises fostered of a new form of evaluation research in this period: studying test collection methodologies. This research is covered in Section 6.

The one exception to the ordering can be found in the section on the use of significance testing. Apart from a recent book [74], little has been written on the use of significance in IR evaluation and relatively little research has been conducted; consequently, I chose to describe research in this area, in Section 5, more as a tutorial than a survey.

Such an ordering means that descriptions of or references to evaluation measures are spread throughout the document. Therefore, we provide an index at the conclusion of this work to aid in their location.

Note, unless explicitly stated otherwise, the original versions of all work cited in this document were obtained and read by the author.

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, “Learning user interaction models for predicting web search result preferences,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10, New York, NY, USA: ACM, 2006.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, “Diversifying search results,” in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 5–14, ACM, 2009.

[3] A. Al-Maskari, M. Sanderson, and P. Clough, “The relationship between IR effectiveness measures and user satisfaction,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 773–774, New York, NY, USA: ACM Press, 2007.

[4] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio, “The good and the bad system: Does the test collection predict users' effectiveness?,” in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–66, New York, NY, USA: ACM, 2008.

[5] J. Allan, *Topic Detection and Tracking: Event-based Information Organization,* (The Kluwer International Series on Information Retrieval, vol. 12). Springer, 1st ed., 2002.

[6] J. Allan, B. Carterette, and J. Lewis, “When will information retrieval be “good enough”?,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433–440, New York, NY, USA: ACM, 2005.

[7] O. Alonso, D. E. Rose, and B. Stewart, “Crowdsourcing for relevance evaluation,” *ACM SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

[8] D. G. Altman, *Practical Statistics for Medical Research.* Chapman & Hall/ CRC, 1st ed., 1990.

[9] E. Amitay, D. Carmel, R. Lempel, and A. Soffer, “Scaling IR-system Evaluation using Term Relevance Sets,” in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, pp. 10–17, New York, NY, USA: ACM, 2004.

[10] T. Arni, P. Clough, M. Sanderson, and M. Grubinger, “Overview of the ImageCLEFphoto 2008 photographic retrieval task,” *Evaluating Systems for Multilingual and Multimodal Information Access*, Lecture Notes in Computer Science, *5706/2009*, 500-511. doi:10.1007/978-3-642-04447-2_62, 2009.

[11] J. Artiles, S. Sekine, and J. Gonzalo, “Web people search: Results of the first evaluation and the plan for the second,” in *Proceedings of the 17th International Conference on World Wide Web*, pp. 1071–1072, New York, NY, USA: ACM Press, 2008.

[12] J. A. Aslam, V. Pavlu, and R. Savell, “A unified model for metasearch, pooling, and system evaluation,” in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 484–491, New York, NY, USA: ACM, 2003.

[13] J. A. Aslam, V. Pavlu, and E. Yilmaz, “A statistical method for system evaluation using incomplete judgments,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 541–548, ACM, 2006.

[14] J. A. Aslam, E. Yilmaz, and V. Pavlu, “A geometric interpretation of r-precision and its correlation with average precision,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 573–574, New York, NY, USA: ACM, 2005.

[15] J. A. Aslam, E. Yilmaz, and V. Pavlu, “The maximum entropy method for analyzing retrieval measures,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27–34, New York, NY, USA: ACM, 2005.

[16] L. Azzopardi, M. de Rijke, and K. Balog, “Building simulated queries for known-item topics: An analysis using six European languages,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 455–462, New York, NY, USA: ACM, 2007.

[17] L. Azzopardi and V. Vinay, “Retrievability: An evaluation measure for higher order information access tasks,” in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 561–570, ACM Press: New York, NY, USA, 2008.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Addison Wesley, 1999.

[19] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, “Relevance assessment: Are judges exchangeable and does it matter,” in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 667–674, New York, NY, USA: ACM, 2008.

[20] M. Baillie, L. Azzopardi, and I. Ruthven, "A retrieval evaluation methodology for incomplete relevance assessments," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 4425, pp. 271–282, 2007.

[21] M. Barbaro and T. Zeller Jr, "A face is exposed for AOL Searcher No. 4417749," *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html, August 9 2006.

[22] J. R. Baron, D. D. Lewis, and D. W. Oard, "TREC-2006 legal track overview," in *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, NIST Special Publication, vol. 500, pp. 79–98, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2006.

[23] S. Bjørner and S. C. Ardito, "Online before the internet, Part 1: Early pioneers tell their stories," *Searcher: The Magazine for Database Professionals*, vol. 11, no. 6, 2003.

[24] D. C. Blair, "STAIRS redux: Thoughts on the STAIRS evaluation, ten years after," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 4–22, 1996.

[25] 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, no. 3, pp. 289–299, doi:10.1145/3166.3197, 1985.

[26] D. Bodoff and P. Li, "Test theory for assessing IR test collections," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 367–374, New York, NY, USA: ACM, 2007.

[27] T. Bompada, C. C. Chang, J. Chen, R. Kumar, and R. Shenoy, "On the robustness of relevance measures with incomplete judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 359–366, New York, NY, USA: ACM Press, 2007.

[28] A. Bookstein, "When the most "pertinent" document should not be retrieved — An analysis of the Swets model," *Information Processing & Management*, vol. 13, no. 6, pp. 377–383, 1977.

[29] H. Borko, *Evaluating The: Effectiveness of Information Retrieval Systems* (No. Sp-909/000/00). Santa Monica, California: Systems Development Corporation, 1962.

[30] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[31] P. Borlund and P. Ingwersen, "The development of a method for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[32] H. Bornstein, "A paradigm for a retrieval effectiveness experiment," *American Documentation*, vol. 12, no. 4, pp. 254–259, doi:10.1002/asi.5090120403, 1961.

[33] M. Braschler and C. Peters, "Cross-language evaluation forum: Objectives, results, achievements," *Information Retrieval*, vol. 7, no. 1–2, pp. 7–31, 2004.

[34] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, doi:10.1145/792550.792552, 2002.

[35] E. C. Bryant, "Progress towards evaluation of information retrieval systems," in *Information Retrieval Among Examining Patent Offices: 4th Annual*

Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices (ICIREPAT), pp. 362–377, Spartan Books, Macmillan, 1966.

[36] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, “Bias and the limits of pooling for large collections,” *Information Retrieval*, vol. 10, no. 6, pp. 491–508, 2007.

[37] C. Buckley and E. M. Voorhees, “Evaluating evaluation measure stability,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2000.

[38] C. Buckley and E. M. Voorhees, “Retrieval evaluation with incomplete information,” in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32, New York, NY, USA: ACM, 2004.

[39] C. Buckley and E. M. Voorhees, “Retrieval system evaluation,” in *TREC: Experiment and Evaluation in Information Retrieval*, pp. 53–75, MIT Press, 2005.

[40] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, Bonn, Germany, 2005.

[41] R. Burgin, “Variations in relevance judgments and the evaluation of retrieval performance,” *Information Processing & Management*, vol. 28, no. 5, pp. 619–627, doi:10.1016/0306-4573(92)90031-T, 1992.

[42] S. Büttcher, C. L. A. Clarke, and I. Soboroff, “The TREC 2006 terabyte track,” in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, vol. 500, pp. 128–141, Maryland, USA: Gaithersburg, 2006.

[43] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff, “Reliable information retrieval evaluation with incomplete and biased judgements,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 63–70, New York, NY, USA: ACM Press, 2007.

[44] F. Can, R. Nuray, and A. B. Sevdik, “Automatic performance evaluation of Web search engines,” *Information Processing and Management*, vol. 40, no. 3, pp. 495–514, 2004.

[45] B. Carterette, “Robust test collections for retrieval evaluation,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–62, New York, NY, USA: ACM Press, 2007.

[46] B. Carterette, “On rank correlation and the distance between rankings,” in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 436–443, ACM, 2009.

[47] B. Carterette, J. Allan, and R. Sitaraman, “Minimal test collections for retrieval evaluation,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275, New York, NY, USA: ACM, 2006.

[48] B. Carterette, P. Bennett, D. Chickering, and S. Dumais, "Here or There," in *Advances in Information Retrieval*, pp. 16–27, 2008. Retrieved from http://dx.doi.org/10.1007/978-3-540-78646-7_5.

[49] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 217–224, 2007.

[50] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," *Advances in Neural Information Processing Systems*, vol. 20, pp. 217–224, 2008.

[51] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 651–658, New York, NY, USA: ACM, 2008.

[52] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 621–630, New York, NY, USA: ACM Press, 2009.

[53] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 429–436, New York, NY, USA: ACM, 2006.

[54] C. L. A. Clarke, N. Craswell, and I. Soboroff, "Preliminary report on the TREC 2009 Web track," *Working notes of the proceedings of TREC 2009*, 2009.

[55] 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 the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 659–666, New York, NY, USA: ACM, 2008.

[56] C. L. A. Clarke, M. Kolla, and O. Vechtomova, "An effectiveness measure for ambiguous and underspecified queries," in *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10–12, 2009 Proceedings*, pp. 188–199, New York Inc: Springer-Verlag, 2009.

[57] C. W. Cleverdon, "The evaluation of systems used in information retrieval (1958: Washington)," in *Proceedings of the International Conference on Scientific Information — Two Volumes*, pp. 687–698, Washington: National Academy of Sciences, National Research Council, 1959.

[58] C. W. Cleverdon, *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems.* ASLIB Cranfield Research Project. Cranfield, UK, 1962.

[59] C. W. Cleverdon, *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages*, (Cranfield Library Report No. 3). Cranfield Institute of Technology, 1970.

[60] C. W. Cleverdon, "The significance of the cranfield tests on index languages," in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12,

Chicago, Illinois, United States: ACM Press New York, NY, USA, 1991. doi:10.1145/122860.122861.

[61] C. W. Cleverdon and M. Keen, "Factors Affecting the Performance of Indexing Systems," Vol 2. *ASLIB, Cranfield Research Project. Bedford, UK: C. Cleverdon*, pp. 37–59, 1966.

[62] P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artiles, and V. Peinado, "Large-scale interactive evaluation of multilingual information access systems — The iCLEF flickr challenge," in *Proceedings of Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pp. 33–38, Glasgow, UK, 2008.

[63] P. Clough, H. Muller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh, "The CLEF 2005 cross-language image retrieval track," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 535–557, 2006.

[64] P. Clough, M. Sanderson, and H. Muller, "The CLEF cross language image retrieval track (ImageCLEF) 2004," *Lecture notes in Computer Science*, pp. 243–251, 2004.

[65] W. S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems," *American Documentation*, vol. 19, no. 1, pp. 30–41, doi:10.1002/asi.5090190108, 1968.

[66] W. S. Cooper, "A definition of relevance for information retrieval," *Information storage and retrieval*, vol. 7, no. 1, pp. 19–37, 1971.

[67] W. S. Cooper, "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, no. 2, 1973.

[68] G. V. Cormack and T. R. Lynam, "Statistical precision of information retrieval evaluation," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–540, New York, NY, USA: ACM, 2006.

[69] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 282–289, New York, NY, USA: ACM, 1998.

[70] N. Craswell, A. de Vries, and I. Soboroff, "Overview of the trec-2005 enterprise track," in *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*, Gaithersburg, Maryland, USA, 2005.

[71] N. Craswell and D. Hawking, "Overview of the TREC 2002 Web track," in *The Eleventh Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251, pp. 86–95, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2003.

[72] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of the international conference on Web search and web data mining*, pp. 87–94, ACM, 2008.

[73] W. B. Croft, "A file organization for cluster-based retrieval," in *Proceedings of the 1st Annual International ACM SIGIR Conference on Information Storage and Retrieval*, pp. 65–82, New York, NY, USA: ACM, 1978.

[74] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice.* Addison Wesley, 1st ed., 2009.

[75] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin, "Fast, flexible filtering with phlat," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 261–270, New York, NY, USA: ACM, 2006.

[76] A. Davies, *A Document Test Collection for Use in Information Retrieval Research*, (Dissertation). Department of Information Studies. University of Sheffield, 1983.

[77] G. Demartini and S. Mizzaro, "A classification of IR effectiveness metrics," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 3936, pp. 488–491, 2006.

[78] B. K. Dennis, J. J. Brady, and J. A. Dovel, "Index manipulation and abstract retrieval by computer," *Journal of Chemical Documentation*, vol. 2, no. 4, pp. 234–242, 1962.

[79] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've seen: A system for personal information retrieval and re-use," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, ACM, 2003.

[80] G. E. Dupret and B. Piwowarski, "A user browsing model to predict search engine click data from past observations," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 331–338, ACM, 2008.

[81] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science*, vol. 1, no. 1, pp. 54–77, 1986.

[82] B. Efron and R. J. Tibshirani, "An introduction to the bootstrap," *Monographs on Statistics and Applied Probability*, vol. 57, pp. 1–177, 1993.

[83] R. A. Fairthorne, "Implications of test procedures," in *Information Retrieval in Action*, pp. 109–113, Cleveland, Ohio, USA: Western Reserve UP, 1963.

[84] E. M. Fels, "Evaluation of the performance of an information-retrieval system by modified Mooers plan," *American Documentation*, vol. 14, no. 1, pp. 28–34, doi:10.1002/asi.5090140105, 1963.

[85] E. A. Fox, *Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts*, (Computer Science Technical Reports). Cornell University. Retrieved from http://techreports.library.cornell.edu:8081/Dienst/UI/1.0/Display/cul.cs/TR83-561, 1983.

[86] E. A. Fox, *Virginia Disc One.* Blacksburg, VA, USA: Produced by Nimbus Records, 1990.

[87] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 2, pp. 147–168, 2005.

[88] H. P. Frei and P. Schäuble, "Determining the effectiveness of retrieval algorithms," *Information Processing and Management: An International Journal*, vol. 27, no. 2–3, pp. 153–164, 1991.

[89] N. Fuhr, "Optimum polynomial retrieval functions based on the probability ranking principle," *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 183–204, 1989.

[90] 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*, 2002.

[91] N. Fuhr and G. E. Knorz, "Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS)," in *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 391–408, UK: British Computer Society Swindon, 1984.

[92] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 107–130, Gaithersburg, Maryland, USA, 2000.

[93] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras, "GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 908–919, 2006.

[94] G. Gigerenzer, "Mindless statistics," *Journal of Socio-Economics*, vol. 33, no. 5, pp. 587–606, 2004.

[95] W. Goffman, "On relevance as a measure," *Information Storage and Retrieval*, vol. 2, pp. 201–203, 1964.

[96] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A brief history," in *Proceedings of the 16th Conference on Computational Linguistics — vol. 1*, pp. 466–471, Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=992628.992709, 1996.

[97] C. D. Gull, "Seven years of work on the organization of materials in the special library," *American Documentation*, vol. 7, no. 4, pp. 320–329, doi:10.1002/asi.5090070408, 1956.

[98] D. K. Harman, "Evaluation issues in information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 439–440, 1992.

[99] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," in *NIST Special Publication. Presented at the Second Text Retrieval Conference (TREC 2)*, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1993.

[100] D. K. Harman, "Overview of the third text retrieval conference (TREC-3)," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1994.

[101] D. K. Harman, "Overview of the fourth text retrieval conference (TREC-4)," in *The Forth Text Retrieval Conference (TREC-4), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1995.

[102] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," *Information Processing and Management*, vol. 31, no. 3, pp. 271–289, 1995.

[103] D. K. Harman, "Overview of the TREC 2002 novelty track," in *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, pp. 46–56, Gaithersburg, Maryland, USA, 2002.

[104] D. K. Harman, "Some Interesting Unsolved Problems in Information Retrieval," Presented at the Center for Language and Speech Processing, Workshop 2002, The Johns Hopkins University 3400 North Charles Street, Barton Hall Baltimore, MD 21218. Retrieved from http://www.clsp.jhu.edu/ws02/preworkshop/lecture_harman.shtml, July 2 2002.

[105] D. K. Harman, "The TREC ad hoc experiments," in *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, pp. 79–98, MIT Press, 2005.

[106] D. K. Harman and C. Buckley, "The NRRC reliable information access (RIA) workshop," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 528–529, New York, NY, USA: ACM, 2004.

[107] D. K. Harman and G. Candela, "Retrieving records from a gigabyte of text on a minicomputer using statistical ranking," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 581–589, 1990.

[108] V. Harmandas, M. Sanderson, and M. D. Dunlop, "Image retrieval by hypertext links," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 296–303, New York, NY, USA: ACM, 1997.

[109] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk, "Evaluating strategies for similarity search on the web," in *Proceedings of the 11th International Conference on World Wide Web*, pp. 432–442, New York, NY, USA: ACM Press, 2002.

[110] D. Hawking, "Overview of the TREC-9 Web track," in *NIST Special Publication*, pp. 87–102, 2001. Presented at the Ninth Text Retrieval Conference (TREC-9), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[111] D. Hawking, P. Bailey, and N. Craswell, "ACSys TREC-8 Experiments," in *NIST Special Publication*, pp. 307–316, 2000. Presented at the Eighth Text Retrieval Conference (TREC-8), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[112] D. Hawking, N. Craswell, and P. Thistlewaite, "Overview of TREC-7 very large collection track," in *The Seventh Text Retrieval Conference (TREC-7)*, pp. 91–104, NIST Special Publication, 1998. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[113] D. Hawking, F. Crimmins, and N. Craswell, "How valuable is external link evidence when searching enterprise Webs?," in *Proceedings of the 15th Australasian database conference*, vol. 27, pp. 77–84, Darlinghurst, Australia: Australian Computer Society, Inc, 2004.

[114] D. Hawking and S. E. Robertson, "On collection size and retrieval effectiveness," *Information Retrieval*, vol. 6, no. 1, pp. 99–105, 2003.

[115] D. Hawking and P. Thistlewaite, "Overview of TREC-6 very large collection track," in *The Sixth Text Retrieval Conference (TREC-6)*, pp. 93–106, NIST Special Publication, 1997. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[116] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 1st ed., 2009.

[117] M. A. Hearst and C. Plaunt, "Subtopic structuring for full-length document access," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–68, New York, NY, USA: ACM, 1993.

[118] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: An interactive retrieval evaluation and new large test collection for research," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192–201, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[119] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli, "TREC 2006 genomics track overview," in *The Fifteenth Text Retrieval Conference*, pp. 52–78, Gaithersburg, Maryland, USA, 2006.

[120] W. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, and C. Mosbaek et al., "Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions," *Journal of American Medical Informatics Association*, vol. 9, 2002.

[121] W. Hersh and P. Over, "TREC-9 Interactive Track Report," in *proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pp. 41–50, Gaithersburg, Maryland: NTIS, 2000.

[122] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, "Do batch and user evaluations give the same results?," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 17–24, New York, NY, USA: ACM Press, 2000.

[123] J. E. Holmstrom, "Section III. Opening plenary session," in *The Royal Society Scientific Information Conference, 21 June–2 July 1948: Report and papers submitted*, London: Royal Society, 1948.

[124] S. B. Huffman and M. Hochster, "How well does result relevance predict session satisfaction?," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 567–574, New York, NY, USA: ACM Press, 2007.

[125] D. Hull, "Using statistical testing in the evaluation of retrieval experiments," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, New York, NY, USA: ACM, 1993.

[126] S. Huuskonen and P. Vakkari, "Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine," *Journal of Documentation*, vol. 64, no. 2, pp. 287–303, 2008.

[127] E. Ide, "New experiments in relevance feedback," in *Report ISR-14 to the National Science Foundation*, Cornell University, Department of Computer Science, 1968.

[128] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer, 2005.

[129] M. Iwayama, A. Fujii, N. Kando, and A. Takano, "Overview of patent retrieval task at NTCIR-3," in *Proceedings of the third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[130] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, no. 1, pp. 248–263, 2006.

[131] B. J. Jansen, A. Spink, and S. Koshman, "Web searcher interaction with the Dogpile.com metasearch engine," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 744–744, 2007.

[132] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, New York, NY, USA: ACM, 2000.

[133] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[134] E. Jensen, *Repeatable Evaluation of Information Retrieval Effectiveness In Dynamic Environments*. Illinois Institute of Technology. Retrieved from http://ir.iit.edu/∼ej/jensen_phd_thesis.pdf, May 2006.

[135] E. C. Jensen, S. M. Beitzel, A. Chowdhury, and O. Frieder, "Repeatable evaluation of search services in dynamic environments," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 1, p. 1, doi:10.1145/1292591.1292592, 2007.

[136] T. Joachims, "Evaluating retrieval performance using clickthrough data," in *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, pp. 12–15, 2002.

[137] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, New York, NY, USA: ACM Press, 2002.

[138] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161, New York, NY, USA: ACM, 2005.

[139] T. Joachims and F. Radlinski, "Search engines that learn from implicit feedback," *Computer*, pp. 34–40, 2007.

[140] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. E. Robertson, "INEX 2007 evaluation measures," pp. 24–33, Retrieved from http://dx.doi.org/10.1007/978-3-540-85902-4_2, 2008.

[141] N. Kando, "Evaluation of information access technologies at the NTCIR workshop," in *Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF*, pp. 29–43, Springer, 2003.

[142] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka, "Overview of IR tasks at the first NTCIR workshop," in *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 11–44, 1999.

[143] P. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track," in *The Fifth Text Retrieval Conference (TREC-5)*, NIST Special Publication.

Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[144] P. B. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text," vol. 2, no. 2, pp. 165–176, 2000.

[145] R. V. Katter, "The influence of scale form on relevance judgments," *Information Storage and Retrieval*, vol. 4, no. 1, pp. 1–11, 1968.

[146] J. Katzer, M. J. McGill, J. A. Tessier, W. Frakes, and P. DasGupta, "A study of the overlap among document representations," *Information Technology: Research and Development*, vol. 1, no. 4, pp. 261–274, 1982.

[147] J. Katzer, J. A. Tessier, W. Frakes, and P. DasGupta, *A Study of the Impact of Representations in Information Retrieval Systems.* Syracuse, New York: School of Information Studies, Syracuse University, 1981.

[148] G. Kazai and M. Lalmas, "INEX 2005 evaluation measures," in *Advances in XML Information Retrieval and Evaluation*, Lecture Notes in Computer Science, vol. 3977, pp. 16–29, 2006.

[149] G. Kazai, M. Lalmas, and A. P. de Vries, "The overlap problem in content-oriented XML retrieval evaluation," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, New York, NY, USA: ACM, 2004.

[150] G. Kazai, N. Milic-Frayling, and J. Costello, "Towards methods for the collective gathering and quality control of relevance assessments," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 452–459, ACM, 2009.

[151] E. M. Keen, "Evaluation parameters," in *Report ISR-13 to the National Science Foundation*, Cornell University, Department of Computer Science, 1967.

[152] E. M. Keen, "Presenting results of experimental retrieval comparisons," *Information Processing and Management: An International Journal*, vol. 28, no. 4, pp. 491–502, 1992.

[153] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.

[154] A. Kent, *Encyclopedia of Library and Information Science.* CRC Press, 2002.

[155] A. Kent, M. M. Berry, F. U. Luehrs Jr, and J. W. Perry, "Machine literature searching VIII. Operational criteria for designing information retrieval systems," *American Documentation*, vol. 6, no. 2, pp. 93–101, doi:10.1002/asi.5090060209, 1955.

[156] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi, "Pooling for a large-scale test collection: An analysis of the search results from the first NTCIR workshop," *Information Retrieval*, vol. 5, no. 1, pp. 41–59, 2002.

[157] M. Lalmas and A. Tombros, "Evaluating XML retrieval effectiveness at INEX," *SIGIR Forum*, vol. 41, no. 1, pp. 40–57, doi:10.1145/1273221.1273225, 2007.

[158] F. W. Lancaster, *Evaluation of the MEDLARS Demand Search Service.* (No. PB-178-660) (p. 278). Springfield, VA 22151: Clearinghouse for Federal Scientific and Technical Information, 1968.

[159] F. W. Lancaster, *Information Retrieval Systems Characteristics, Testing, and Evaluation.* John Wiley & Sons, Inc, 1968.

[160] R. Ledwith, "On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases," *Information Processing & Management*, vol. 28, no. 4, pp. 451–455, doi:10.1016/0306-4573(92)90003-I, 1992.

[161] C. Léger, J. P. Romano, and D. N. Politis, "Bootstrap technology and applications," *Technometrics*, vol. 34, no. 4, pp. 378–398, 1992.

[162] M. E. Lesk, "SIG — The significance programs for testing the evaluation output," in *Report ISR-12 to the National Science Foundation*, Cornell University, Department of Computer Science, 1966.

[163] M. E. Lesk, D. Cutting, J. Pedersen, T. Noreault, and M. Koll, "Real life information retrieval (panel): Commercial search engines," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 333, New York, NY, USA: ACM, 1997.

[164] M. E. Lesk and G. Salton, "Relevance assessments and retrieval system evaluation*1," *Information Storage and Retrieval*, vol. 4, no. 4, pp. 343–359, doi:10.1016/0020-0271(68)90029-6, 1968.

[165] D. Lewis, "The TREC-5 filtering track," in *The Fifth Text Retrieval Conference (TREC-5)*, pp. 75–96, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[166] J. Lin and B. Katz, "Building a reusable test collection for question answering," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 851–861, 2006.

[167] H. Liu, R. Song, J. Y. Nie, and J. R. Wen, "Building a test collection for evaluating search result diversity: A preliminary study," in *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 31–32, 2009.

[168] T. Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li, "Letor: Benchmark dataset for research on learning to rank for information retrieval," in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pp. 3–10, 2007.

[169] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. De Rijke, and P. Rocha et al., "Overview of the CLEF 2004 multilingual question answering track," *Lecture notes in Computer Science*, vol. 3491, p. 371, 2005.

[170] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Mitra, and A. Sen et al., "Text collections for FIRE," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 699–700, ACM, 2008.

[171] R. Manmatha, T. Rath, and F. Feng, "Modeling score distributions for combining the outputs of search engines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–275, New York, NY, USA: ACM Press, 2001.

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

[173] M. Maron, J. Kuhns, and L. Ray, *Probabilistic Indexing: A Statistical Technique for Document Identification and Retrieval* (Technical Memorandum No. 3) (p. 91)*, Data Systems Project Office*. Los Angeles, California: Thompson Ramo Wooldridge Inc, 1959.

[174] D. Meister and D. Sullivan, "Evaluation of User Reactions to a Prototype On-line Information Retrieval System," Prepared under Contract No.

NASw-1369 by Bunker-Ramo Corporation, Canoga Park, CA. (No. NASA CR-918). NASA, 1967.

[175] M. Melucci, "On rank correlation in information retrieval evaluation," *ACM SIGIR Forum*, vol. 41, no. 1, pp. 18–33, 2007.

[176] T. Minka and S. E. Robertson, "Selection bias in the LETOR datasets," in *SIGIR Workshop on Learning to Rank for Information Retrieval*, pp. 48–51, 2008.

[177] S. Mizzaro and S. Robertson, "Hits hits TREC: exploring IR evaluation results with network analysis," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 479–486, New York, NY, USA: ACM, 2007.

[178] A. Moffat, W. Webber, and J. Zobel, "Strategic system comparisons via targeted relevance judgments," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–382, Amsterdam, The Netherlands, 2007.

[179] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transactions on Information Systems*, vol. 27, no. 1, p. Article No. 2, 2008.

[180] C. N. Mooers, *The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval System* (No. ZTB-132) (p. 20). Cambridge, Massachusetts: Zator Corporation, 1959.

[181] C. N. Mooers, "The next twenty years in information retrieval: Some goals and predictions," in *Papers Presented at the Western Joint Computer Conference*, pp. 81–86, ACM, 1959.

[182] M. Morita and Y. Shinoda, "Information filtering based on user behavior analysis and best match text retrieval," in *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 272–281, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[183] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," *Information Processing and Management*, vol. 42, no. 3, pp. 595–614, 2006.

[184] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, and B. Ramabhadran et al., "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, ACM, 2004.

[185] I. Ounis, M. De Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 blog track," in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, pp. 17–31, Gaithersburg, Maryland, USA, 2006.

[186] P. Over, "TREC-6 interactive report," in *Proceedings of the sixth Text Retrieval Conference (TREC-6)*, NIST Special Publication, vol. 500, pp. 73–82, Gaithersburg, Maryland, USA, 1997.

[187] P. Over, "The TREC interactive track: An annotated bibliography," *Information Processing and Management*, vol. 37, no. 3, pp. 369–381, 2001.

[188] P. Over, T. Ianeva, W. Kraaij, A. F. Smeaton, and S. Valencia, "TRECVID 2006-An overview," in *Proceedings of the TREC Video Retrieval Evaluation Notebook Papers*, 2006.

[189] W. R. Pearson, "Comparison of methods for searching protein sequence databases," *Protein Science: A Publication of the Protein Society*, vol. 4, no. 6, p. 1145, 1995.

[190] B. Piwowarski, G. Dupret, and R. Jones, "Mining user web search activity with layered bayesian networks or how to capture a click in its context," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 162–171, New York, NY, USA: ACM, 2009.

[191] S. M. Pollock, "Measures for the comparison of information retrieval systems," *American Documentation*, vol. 19, no. 4, pp. 387–397, doi:10.1002/asi.5090190406, 1968.

[192] F. Radlinski, M. Kurup, and T. Joachims, "How does clickthrough data reflect retrieval quality?," in *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pp. 43–52, 2008.

[193] A. M. Rees and D. G. Schultz, "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching," Final Report to the National Science Foundation. Volume II, Appendices. Clearinghouse for Federal Scientific and Technical Information, Springfield, VA. 22151 (PB-176-079), MF $0.65, HC $3.00), October 1967.

[194] A. Ritchie, S. Teufel, and S. Robertson, "Creating a test collection for citation-based IR experiments," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 391–398, Association for Computational Linguistics Morristown, NJ, USA, 2006.

[195] S. E. Robertson, "The parametric description of retrieval tests: Part II: Overall measures," *Journal of Documentation*, vol. 25, no. 2, pp. 93–107, 1969.

[196] S. E. Robertson, "On sample sizes for non-matched-pair IR experiments," *Information Processing and Management: An International Journal*, vol. 26, no. 6, pp. 739–753, 1990.

[197] S. E. Robertson, "Salton award lecture on theoretical argument in information retrieval," *SIGIR Forum*, vol. 34, no. 1, pp. 1–10, doi:10.1145/373593.373597, 2000.

[198] S. E. Robertson, "On GMAP: And other transformations," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 78–83, New York, NY, USA: ACM Press, 2006.

[199] S. E. Robertson, "On the history of evaluation in IR," *Journal of Information Science*, vol. 34, no. 4, pp. 439–456, doi:10.1177/0165551507086989, 2008.

[200] S. E. Robertson and D. A. Hull, "The TREC-9 filtering track final report," in *Proceedings of the Ninth Text REtrieval Conference (TREC-2001)*, pp. 25–40, Gaithersburg, Maryland, USA: NTIS, 2001.

[201] S. E. Robertson and H. Zaragoza, "On rank-based effectiveness measures and optimization," *Information Retrieval*, vol. 10, no. 3, pp. 321–339, 2007.

[202] G. Roda, J. Tait, F. Piroi, and V. Zenz, "CLEF-IP 2009: Retrieval experiments in the intellectual property domain," in *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[203] M. E. Rorvig, "The simple scalability of documents," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 590–598, doi: 10.1002/(SICI)1097-4571(199012)41:8<590::AID-ASI5>3.0.CO;2-T, 1990.

[204] D. Rose and C. Stevens, "V-twin: A lightweight engine for interactive use," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 279–290, 1996.

[205] T. Sakai, "New performance metrics based on multigrade relevance: Their application to question answering," in *NTCIR-4 Proceedings*, 2004.

[206] T. Sakai, "Evaluating evaluation metrics based on the bootstrap," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 525–532, Seattle, Washington, USA: ACM Press New York, NY, USA, 2006. doi:10.1145/1148170.1148261.

[207] T. Sakai, "Alternatives to bpref," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 71–78, ACM, 2007.

[208] T. Sakai, "Evaluating information retrieval metrics based on bootstrap hypothesis tests," *Information and Media Technologies*, vol. 2, no. 4, pp. 1062–1079, 2007.

[209] T. Sakai and N. Kando, "On information retrieval metrics designed for evaluation with incomplete relevance assessments," *Information Retrieval*, vol. 11, no. 5, pp. 447–470, doi:10.1007/s10791-008-9059-7, 2008.

[210] G. Salton, *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

[211] G. Salton, *The Smart Retrieval System. Experiments in Automatic Document Processing.* Prentice Hall, 1971.

[212] G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, New York, NY, USA: ACM, 1993.

[213] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, doi:10.1145/182.358466, 1983.

[214] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *Journal of the ACM (JACM)*, vol. 15, no. 1, pp. 8–36, 1968.

[215] G. Salton and C. T. Yu, "On the construction of effective vocabularies for information retrieval," in *Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval Table of Contents*, pp. 48–60, New York, NY, USA: ACM, 1973.

[216] M. Sanderson, "Accurate user directed summarization from existing tools," in *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pp. 45–51, New York, NY, USA: ACM, 1998.

[217] M. Sanderson and H. Joho, "Forming test collections with no system pooling," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2004.

[218] M. Sanderson and C. J. Rijsbergen, "NRT: News retrieval tool," *Electronic Publishing*, vol. 4, no. 4, pp. 205–217, 1991.

[219] M. Sanderson, T. Sakai, and N. Kando EVIA 2007: The First International Workshop on Evaluating Information Access, 2007.

[220] M. Sanderson and I. Soboroff, "Problems with Kendall's tau," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 839–840, New York, NY, USA: ACM, 2007.

[221] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What else is there? Search diversity examined," in *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pp. 562–569, Springer, 2009.

[222] M. Sanderson and J. Zobel, "Information retrieval system evaluation: Effort, sensitivity, and reliability," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169, New York, NY, USA: ACM, 2005.

[223] T. Saracevic, *An Inquiry into Testing of Information Retrieval Systems: Part II: Analysis of Results*. 1968. (No. CSL:TR-FINAL-II). Comparative Systems Laboratory: Final Technical Report. Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

[224] T. Saracevic, "RELEVANCE: A review of and a framework for the thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, no. 6, pp. 143–165, 1975.

[225] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 138–146, New York, NY, USA: ACM, 1995.

[226] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Information Processing and Management*, vol. 33, no. 4, pp. 495–512, 1997.

[227] Y. Shang and L. Li, "Precision evaluation of search engines," *World Wide Web*, vol. 5, no. 2, pp. 159–173, doi:10.1023/A:1019679624079, 2002.

[228] P. Sheridan, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, New York, NY, USA: ACM, 1997.

[229] M. Shokouhi and J. Zobel, "Robust result merging using sample-based score estimates," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 3, pp. 1–29, 2009.

[230] A. Smeaton and R. Wilkinson, "Spanish and Chinese document retrieval in TREC-5," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 57–64, Gaithersburg, Maryland, USA, 1997.

[231] A. F. Smeaton, W. Kraaij, and P. Over, "TRECVID-An overview," in *Proceedings of the TRECVID 2003 Conference*, Gaithersburg, Maryland, USA. Retrieved from http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/tv3overview.pdf, 2003.

[232] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330, New York, NY, USA: ACM, 2006.

[233] A. F. Smeaton, P. Over, and R. Taban, "The TREC-2001 video track report," in *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, pp. 52–60, 2001.

[234] C. L. Smith and P. B. Kantor, "User adaptation: Good results from poor systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 147–154, New York, NY, USA: ACM, 2008.

[235] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 623–632, New York, NY, USA: ACM, 2007.

[236] I. Soboroff, "On evaluating web search with very few relevant documents," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 530–531, 2004.

[237] I. Soboroff, "Dynamic test collections: Measuring search effectiveness on the live web," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 276–283, New York, NY, USA: ACM, 2006.

[238] I. Soboroff, C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgments," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 66–73, New Orleans, Louisiana, United States: ACM. doi:10.1145/383952.383961, 2001.

[239] I. Soboroff and S. E. Robertson, "Building a filtering test collection for TREC 2002," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 243–250, New York, NY, USA: ACM, 2003.

[240] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324–330, New York, NY, USA: ACM, 2002.

[241] K. Spärck Jones, "Automatic indexing," *Journal of Documentation*, vol. 30, no. 4, pp. 393–432, 1974.

[242] K. Spärck Jones, *Information Retrieval Experiment*. Butterworth-Heinemann Ltd, 1981.

[243] K. Spärck Jones, "Letter to the editor," *Information Processing & Management*, vol. 39, no. 1, pp. 156–159, doi:10.1016/S0306-4573(02)00026-2, 2003.

[244] K. Spärck Jones and R. G. Bates, *Report on a design study for the 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5428). Computer Laboratory, University of Cambridge, 1977.

[245] K. Spärck Jones and C. J. van Rijsbergen, *Report on the need for and the provision of an 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5266) (p. 43). Computer Laboratory, University of Cambridge, 1975.

[246] K. Spärck Jones and C. J. van Rijsbergen, "Information retrieval test collections," *Journal of Documentation*, vol. 32, no. 1, pp. 59–75, doi:10.1108/eb026616, 1976.

[247] L. T. Su, "Evaluation measures for interactive information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 503–516, 1992.

[248] L. T. Su, "The relevance of recall and precision in user evaluation," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 207–217, 1994.

[249] J. A. Swets, "Information retrieval systems," *Science*, vol. 141, no. 3577, pp. 245–250, 1963.

[250] J. A. Swets, "Effectiveness of information retrieval methods," *American Documentation*, vol. 20, no. 1, pp. 72–89, doi:10.1002/asi.4630200110, 1969.

[251] R. Tagliacozzo, "Estimating the satisfaction of information users," *Bulletin of the Medical Library Association*, vol. 65, no. 2, pp. 243–249, 1977.

[252] J. M. Tague and M. J. Nelson, "Simulation of user judgments in bibliographic retrieval systems," in *Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval: Theoretical Issues in Information Retrieval*, pp. 66–71, New York, NY, USA: ACM, 1981.

[253] J. M. Tague-Sutcliffe, "Some perspectives on the evaluation of information retrieval systems," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 1–3, doi:10.1002/(SICI)1097-4571(199601)47:1<1::AID-ASI1>3.0.CO;2-3, 1996.

[254] J. M. Tague-Sutcliffe and J. Blustein, "A statistical analysis of the TREC-3 data," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, pp. 385–398, NIST Special Publication, 1994. Department of Commerce, National Institute of Standards and Technology.

[255] J. A. Thom and F. Scholer, "A comparison of evaluation measures given how users perform on search tasks," *Presented at the Proceedings of the Twelfth Australasian Document Computing Symposium*, pp. 56–63, 2007.

[256] P. Thomas and D. Hawking, "Evaluation by comparing result sets in context," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 94–101, New York, NY, USA: ACM Press, 2006.

[257] R. Thorne, "The efficiency of subject catalogues and the cost of information searches," *Journal of Documentation*, vol. 11, pp. 130–148, 1955.

[258] A. Turpin and W. Hersh, "Why batch and user evaluations do not give the same results," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 225–231, New York, NY, USA: ACM, 2001.

[259] A. Turpin and W. Hersh, "User interface effects in past batch versus user experiments," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 431–432, New York, NY, USA: ACM, 2002.

[260] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–18, New York, NY, USA: ACM, 2006.

[261] C. J. van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[262] C. J. van Rijsbergen, *Information Retrieval.* Butterworth-Heinemann Ltd, 2nd ed., 1979.

[263] P. K. T. Vaswani and J. B. Cameron, *The National Physical Laboratory Experiments in Statistical Word Associations and Their Use in Document Indexing And Retrieval.* National Physical Laboratory Computer Science Division-Publications; COM.SCI.42 (p. 171). National Physical Lab., Teddington (Great Britain), 1970.

[264] J. Verhoeff, W. Goffman, and J. Belzer, "Inefficiency of the use of Boolean functions for information retrieval systems," *Communications of the ACM*, vol. 4, no. 12, pp. 557–558, 1961.

[265] B. C. Vickery, *On Retrieval System Theory.* Butterworths, 2nd ed., 1965.

[266] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, New York, NY, USA: ACM Press, 2004.

[267] E. M. Voorhees, "On expanding query vectors with lexically related words," in *The Second Text Retrieval Conference (TREC 2)*, NIST Special Publication 500-215, pp. 223–231, Department of Commerce, National Institute of Standards and Technology, 1993.

[268] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information retrieval*, pp. 315–323, New York, NY, USA: ACM Press, 1998.

[269] E. M. Voorhees, "The TREC-8 question answering track report," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 77–82, Gaithersburg, Maryland, USA, 1999.

[270] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing and Management*, vol. 36, no. 5, pp. 697–716, 2000.

[271] E. M. Voorhees, "Evaluation by highly relevant documents," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–82, New Orleans, Louisiana, United States. New York, NY, USA: ACM Press, 2001.

[272] E. M. Voorhees, "Overview of the TREC 2003 question answering track," in *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003),* vol. 142, 2003.

[273] E. M. Voorhees, "Overview of the TREC 2004 robust retrieval track," in *The Thirteenth Text Retrieval Conference (TREC 2004)*, NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2005.

[274] E. M. Voorhees, "On test collections for adaptive information retrieval," *Information Processing and Management*, 2008.

[275] E. M. Voorhees, "Topic set size redux," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 806–807, ACM, 2009.

[276] E. M. Voorhees and C. Buckley, "The effect of topic set size on retrieval experiment error," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 316–323, New York, NY, USA: ACM, 2002.

[277] E. M. Voorhees and D. K. Harman, "Overview of the seventh text retrieval conference," in *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pp. 1–24, NIST Special Publication, 1998.

[278] E. M. Voorhees and D. K. Harman, "Overview of the eighth text retrieval conference (TREC-8)," in *The Eighth Text Retrieval Conference (TREC-8)*, pp. 1–24, NIST Special Publication, 1999. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[279] E. M. Voorhees and D. K. Harman, "Overview of TREC 2001," in *NIST Special Publication 500-250*, pp. 1–15, Presented at the Tenth Text Retrieval Conference (TREC 2001), Government Printing Office, 2001.

[280] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press, illustrated ed., 2005.

[281] C. Wade and J. Allan, *Passage Retrieval and Evaluation* (CIIR Technical Report No. IR-396). Amherst, MA, USA: University of Massachusetts, Amherst Center for Intelligent Information Retrieval, 2005.

[282] W. Webber, A. Moffat, and J. Zobel, "Score standardization for inter-collection comparison of retrieval systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 51–58, New York, NY, USA: ACM, 2008.

[283] W. Webber, A. Moffat, and J. Zobel, "Statistical power in retrieval experimentation," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 571–580, ACM, 2008.

[284] R. W. White and D. Morris, "Investigating the querying and browsing behavior of advanced search engine users," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 255–262, New York, NY, USA: ACM Press, 2007.

[285] D. Williamson, R. Williamson, and M. E. Lesk, "The Cornell implementation of the SMART system," in *The SMART Retrieval System: Experiments in Automatic Document Processing*, (G. Salton, ed.), p. 12, Englewood Cliffs, New Jersey: Prentice-Hall, Inc, 1971.

[286] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes.* Morgan Kaufmann, 1999.

[287] S. Wu and F. Crestani, "Methods for ranking information retrieval systems without relevance judgments," in *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 811–816, New York, NY, USA: ACM, 2003.

[288] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 155–162, New York, NY, USA: ACM, 2008.

[289] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 102–111, New York, NY, USA: ACM Press, 2006.

[290] E. Yilmaz, J. A. Aslam, and S. E. Robertson, "A new rank correlation coefficient for information retrieval," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 587–594, New York, NY, USA: ACM, 2008.

[291] E. Yilmaz and S. E. Robertson, “On the choice of effectiveness measures for learning to rank,” in *Learning to Rank for Information Retrieval. Workshop in Conjunction with the ACM SIGIR Conference on Information Retrieval*, Boston, MA, USA: ACM Press New York, NY, USA, 2009.

[292] T. Zeller Jr, “AOL Moves to Increase Privacy on Search Queries,” *The New York Times*, Retrieved from http://www.nytimes.com/2006/08/22/technology/22aol.html, August 22 2006.

[293] C. X. Zhai, W. W. Cohen, and J. Lafferty, “Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 10–17, New York, NY, USA: ACM Press, 2003.

[294] Z. Zheng, K. Chen, G. Sun, and H. Zha, “A regression framework for learning ranking functions using relative relevance judgments,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 287–294, ACM, 2007.

[295] J. Zobel, “How reliable are the results of large-scale information retrieval experiments?,” in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307–314, New York, NY, USA: ACM Press, 1998.

Index