Building Test Collections

An Interactive Guide for Students and Others Without Their Own Evaluation Conference Series

Ian Soboroff
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899-8940
ian.soboroff@nist.gov

This is a full-day tutorial on building and validating test collections.

The intended audience is advanced students who find themselves in need of a test collection, or actually in the process of building a test collection, to support their own research. Not everyone can talk TREC, CLEF, INEX, or NTCIR into running a track to build the collection you need. The goal of this tutorial is to lay out issues, procedures, pitfalls, and practical advice.

Attendees should come with a specific current need for data, and/or details on their in-progress collection building effort. The structure of the tutorial will include a lecture component covering history, techniques, and research questions, and an interactive discussion component during which we will collaboratively work through problems the attendees are currently working on.

The presenter is Dr. Ian Soboroff. Ian is a computer scientist and the leader of the Retrieval Group at the National Institute of Standards and Technology (NIST). The Retrieval Group organizes the Text REtrieval Conference (TREC), the Text Analysis Conference (TAC), and the TREC Video Retrieval Evaluation (TRECVID). These are all large, community-based research workshops that drive the state-of-the-art in information retrieval, video search, web search, text summarization and other areas of information access. He has co-authored many publications in information retrieval evaluation, test collection building, text filtering, collaborative filtering, and intelligent software agents. Ian has built test collections for search, filtering, novelty, web, social media, intranet access and other domains. His current research interests include building test collections for social media environments and nontraditional retrieval tasks.

1 MOTIVATION

Test collections are vital to information retrieval research and experiment. However, test collections only exist for a limited number of genres, data types, and search tasks. Since many graduate students would like to explore IR in novel areas, it's likely they will need to build their own test collection to get the best measurements for their research.

Authors rolling their own test collections have a high bar to overcome because reviewers prefer test collections that emerge from large community evaluations like TREC. However, a significant body of recent research has made it possible for small teams to not only build their own test collections, but to support their use by measuring the properties of the test collection and including those figures in their work. Reviewers can then rely on observstional properties of test collections rather than requiring the blessing of a major venue in order to trust a test collection experiment.

The community only has so much capacity for community evaluations, and a growing field of such venues means that resources are increasingly limited for each. You might think that the presenter is trying to put himself out of business. You might be right.

2 OBJECTIVES

Upon completion of this tutorial, attendees will:

- be familiar with the history of the test collection evaluation paradigm;
- understand the process of beginning from a concrete user task and abstracting that to a test collection design;
- understand different ways of establishing a document collection;
- understand the process of topic development;
- understand how to operationalize the notion of relevance, and be familiar with issues surrounding elicitation of relevance judgments;
- understand the pooling methodologies for sampling documents for labeling, and be familiar with sampling strategies for reducing effort;
- be familiar with procedures for measuring and validating a test collection; and
- be familiar with current research issues in this area.

3 RELEVANCE

This tutorial is highly relevant to information retrieval researchers, especially students nearing the experimental phase of their research. While numerous test collections have already been built and are ready for them to use, it seems increasingly common that researchers wish to explore an information retrieval task for which no test collection yet exists. The choice is to adapt an existing collection, or design a new one. Publication of research based on an independently-developed test collection would seem to be fraught with peril without following best practices drawn from the literature while building that test collection.

4 MATERIALS

Materials for the course will include lecture slides and an annotated bibliography of papers drawn from the current literature for further reading. These materials will be hosted on a website, and students with internet connectivity at the tutorial will be able to access them

SIGIR '17, August 07–11, 2017, Shinjuku, Tokyo, Japan
2017. 978-1-4503-5022-8/17/08
DOI: http://dx.doi.org/10.1145/3077136.3082064

and follow along during the tutorial itself. The bibliography is not complete or comprehensive, but is intended to reflect the current practice and research in the field.

5 TUTORIAL OUTLINE

(1) Introduction to test collections
- basic concepts: task, documents, topics, relevance judgments, and measures.
- history of Cranfield paradigm.

(2) Task
- a task-centered approach to conceiving test collections.
- metrics as an operationalization of task success.
- understanding the user task and the role of the system.

(3) Documents
- the relationship between documents and task.
- naturalism vs constructivism.
- opportunity sampling and bias.
- distribution and sharing.

(4) Topics
- designing topics from a task perspective.
- sources for topics.
- exploration or topic development.
- extracting topics from logs.
- topics and queries.
- topic set size.

(5) Relevance
- defining relevance and utility starting from the task.
- obtaining labels, explicit and implicit elicitation (highlighting).
- Interface considerations.
- inter-annotator agreement.
- errors.
- crowdsourcing for relevance judgments.
- validation, process control, quality assurance.
- annotator skill set.

(6) Pooling
- problem of scale and bias.
- breadth of pools, multiple systems.
- completeness vs. samples.
- methods.
- estimating pool depths.
- leave-one-out (LOO) test, reusability.
- double pooling.

(7) Analysis
- bounding the score resolution (Voorhees, Sakai, etc).
- resolution with respect to evaluation scope – contrasting many different systems or variations of a single system.
- PCA of topic:run scores, bounding of topic breadth.
- anova, factor analysis, regression.
- statistical significance and meaningful differences.
- calibration by user pilot study.
- per-topic analysis and failure analysis.
- bias.

(8) Test collection diagnosis
- LOO test revisited.
- unjudged == irrelevant.
- differentiating poor systems from collection bias.

(9) Validation
- user study.
- side-by-side comparison.
- a/b testing.
- interleaving.

(10) Pooling and sampling
- pooling as a sampling method.
- pooling as optimization.
- move-to-front pooling.
- uniform sampling, stratified sampling, measure sampling.
- minimal test collections.

(11) Advanced task concepts
- filtering, supporting system adaptation.
- sessions, time, user adaptation.
- context, feedback.
- exploration and fuzzy tasks.
- novelty, differential relevance.
- fundamental limits of Cranfield.

The last section is intended to be a bridge from the tutorial material to collection ideas that the attendees bring with them. Depending on the class, the final hour of the tutorial will be spent discussing and solving specific problems brought by students.

REFERENCES

[1] Alonso, O., and Mizzaro, S. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manage. 48*, 6 (Nov. 2012), 1053–1066.

[2] Amitay, E., Carmel, D., Lempel, R., and Soffer, A. 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* (New York, NY, USA, 2004), SIGIR '04, ACM, pp. 10–17.

[3] Aslam, J. A., and Savell, R. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (New York, NY, USA, 2003), SIGIR '03, ACM, pp. 361–362.

[4] Bailey, P., Moffat, A., Scholer, F., and Thomas, P. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2015), SIGIR '15, ACM, pp. 625–634.

[5] Baillie, M., Azzopardi, L., and Ruthven, I. A retrieval evaluation methodology for incomplete relevance assessments. In *Proceedings of the 29th European Conference on IR Research* (Berlin, Heidelberg, 2007), ECIR'07, Springer-Verlag, pp. 271–282.

[6] Baillie, M., Azzopardi, L., and Ruthven, I. Evaluating epistemic uncertainty under incomplete assessments. *Inf. Process. Manage. 44*, 2 (Mar. 2008), 811–837.

[7] Beheshti, J., Bilal, D., Druin, A., and Large, A. Testing children's information retrieval systems: Challenges in a new era. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47* (Silver Springs, MD, USA, 2010), ASIS&T '10, American Society for Information Science, pp. 55:1–55:4.

[8] Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., and Tran Duc, T. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2011), SIGIR '11, ACM, pp. 923–932.

[9] Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. Bias and the limits of pooling for large collections. *Inf. Retr. 10*, 6 (Dec. 2007), 491–508.

[10] Buckley, C., and Voorhees, E. M. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2004), SIGIR '04, ACM, pp. 25–32.

[11] Büttcher, S., Clarke, C. L. A., Yeung, P. C. K., and Soboroff, I. 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 (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 63–70.

[12] Carterette, B. Robust test collections for retrieval evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 55–62.

[13] Carterette, B. On rank correlation and the distance between rankings. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 436–443.

[14] Carterette, B., and Allan, J. Incremental test collections. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2005), CIKM '05, ACM, pp. 680–687.

[15] Carterette, B., Allan, J., and Sitaraman, R. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2006), SIGIR '06, ACM, pp. 268–275.

[16] Carterette, B., Gabrilovich, E., Josifovski, V., and Metzler, D. Measuring the reusability of test collections. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2010), WSDM '10, ACM, pp. 231–240.

[17] Carterette, B., Kanoulas, E., Pavlu, V., and Fang, H. Reusable test collections through experimental design. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), SIGIR '10, ACM, pp. 547–554.

[18] Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., and Allan, J. If i had a million queries. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (Berlin, Heidelberg, 2009), ECIR '09, Springer-Verlag, pp. 288–300.

[19] Carterette, B., and Soboroff, I. The effect of assessor error on IR system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), SIGIR '10, ACM, pp. 539–546.

[20] Clarke, C. L., Culpepper, J. S., and Moffat, A. Assessing efficiency—effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. *Inf. Retr. 19*, 4 (Aug. 2016), 351–377.

[21] Cormack, G. V., Palmer, C. R., and Clarke, C. L. A. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 282–289.

[22] Diaz, F. Condensed list relevance models. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (New York, NY, USA, 2015), ICTIR '15, ACM, pp. 313–316.

[23] Efron, M. Using multiple query aspects to build test collections without human relevance judgments. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (Berlin, Heidelberg, 2009), ECIR '09, Springer-Verlag, pp. 276–287.

[24] Eickhoff, C., Harris, C. G., de Vries, A. P., and Srinivasan, P. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2012), SIGIR '12, ACM, pp. 871–880.

[25] Hauff, C., and de Jong, F. Retrieval system evaluation: Automatic evaluation versus incomplete judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), SIGIR '10, ACM, pp. 863–864.

[26] Hauff, C., Hiemstra, D., de Jong, F., and Azzopardi, L. Relying on topic subsets for system ranking estimation. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (New York, NY, USA, 2009), CIKM '09, ACM, pp. 1859–1862.

[27] Hosseini, M., Cox, I. J., Milic-Frayling, N., Vinay, V., and Sweeting, T. Selecting a subset of queries for acquisition of further relevance judgements. In *Proceedings of the Third International Conference on Advances in Information Retrieval Theory* (Berlin, Heidelberg, 2011), ICTIR'11, Springer-Verlag, pp. 113–124.

[28] Imhof, M., and Braschler, M. Are test collections "real"? mirroring real-world complexity in IR test collections. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283* (New York, NY, USA, 2015), CLEF'15, Springer-Verlag New York, Inc., pp. 241–247.

[29] Jayasinghe, G. K., Webber, W., Sanderson, M., and Culpepper, J. S. Extending test collection pools without manual runs. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2014), SIGIR '14, ACM, pp. 915–918.

[30] Jayasinghe, G. K., Webber, W., Sanderson, M., and Culpepper, J. S. Improving test collection pools with machine learning. In *Proceedings of the 2014 Australasian Document Computing Symposium* (New York, NY, USA, 2014), ADCS '14, ACM, pp. 2:2–2:9.

[31] Kamps, J., Koolen, M., and Trotman, A. Comparative analysis of clicks and judgments for IR evaluation. In *Proceedings of the 2009 Workshop on Web Search Click Data* (New York, NY, USA, 2009), WSCD '09, ACM, pp. 80–87.

[32] Kazai, G., Milic-Frayling, N., and Costello, J. 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* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 452–459.

[33] Lease, M. On quality control and machine learning in crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation* (2011), AAAIWS'11-11, AAAI Press, pp. 97–102.

[34] Lipani, A., Lupu, M., and Hanbury, A. Splitting water: Precision and anti-precision to reduce pool bias. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2015), SIGIR '15, ACM, pp. 103–112.

[35] Losada, D. E., Parapar, J., and Barreiro, A. Feeling lucky?: Multi-armed bandits for ordering judgements in pooling-based evaluation. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (New York, NY, USA, 2016), SAC '16, ACM, pp. 1027–1034.

[36] Lu, X., Moffat, A., and Culpepper, J. S. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr. 19*, 4 (Aug. 2016), 416–445.

[37] Sakai, T. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2006), SIGIR '06, ACM, pp. 525–532.

[38] Sakai, T. Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 71–78.

[39] Sakai, T., and Kando, N. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr. 11*, 5 (Oct. 2008), 447–470.

[40] Soboroff, I., Nicholas, C., and Cahan, P. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2001), SIGIR '01, ACM, pp. 66–73.

[41] Urbano, J., Schedl, M., and Serra, X. Evaluation in music information retrieval. *J. Intell. Inf. Syst. 41*, 3 (Dec. 2013), 345–369.

[42] Voorhees, E. M. 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* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 315–323.

[43] Voorhees, E. M. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2001), SIGIR '01, ACM, pp. 74–82.

[44] Webber, W., and Park, L. A. F. Score adjustment for correction of pooling bias. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 444–451.

[45] Yilmaz, E., and Aslam, J. A. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2006), CIKM '06, ACM, pp. 102–111.

[46] Yilmaz, E., Aslam, J. A., and Robertson, S. 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* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 587–594.

[47] Yilmaz, E., Kanoulas, E., and Aslam, J. A. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 603–610.

[48] Zobel, J. 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* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 307–314.