Inf Retrieval (2008) 11:473–475
DOI 10.1007/s10791-008-9064-x

BOOK REVIEW

Ellen Voorhees and Donna Harman (eds): TREC Experiment and Evaluation in Information Retrieval

MIT Press, Cambridge, 2005, 462 pp, Price: $45.00, ISBN: 0262220733

Vanessa Murdock

Published online: 14 June 2008

TREC Experiment and Evaluation edited by Ellen Voorhees and Donna Harman provides an overview of the body of work produced by TREC since NIST was tasked with building a testbed for information retrieval in 1990. The book is organized in three sections. The first section introduces TREC, discusses the research environment at the time TREC was started, and explains decisions about the document collections, task definitions, and evaluation of retrieval systems. The second section highlights selected track reports. The third section highlights the contributions of selected participants. The final chapter of the book, the Epilogue: "Metareflections on TREC" by Karen Sparck Jones, provides a summary of the major trends and findings in a crucial 10 years in the development of information retrieval systems from 1992 to 2002.

Books such as this risk becoming a catalog of material already available in the TREC Proceedings. While some of the chapters in the second and third sections take a more narrow view, and simply summarize what was learned from a given track or system, the book as a whole provides insight into decisions that shaped the evaluation of information retrieval systems. It gives perspective on the evolution of the state-of-the-art in information retrieval and the primary value of the book is that it provides a broad view it is not possible to glean from the individual track reports.

Chapters one through three provide an historical view of the evaluation of information retrieval systems, and the ways in which TREC standardized and modernized the evaluation methodology. It is clear that the current environment in academic research has been shaped for the better by the TREC paradigm. NIST provided the first large-scale collection of documents, a standard set of topics, and a level playing field for the comparison of systems, at a time when IR systems were not comparable because the reported evaluation metrics differed from system to system, and the collections were so small as to be practically useless for any real-world task. Of the goals stated in Chapter 1 (page 5), TREC has fallen short on goals two and three ("To increase communication among industry, academia, and government" and "To speed the transfer of technology from research labs into commercial products"). It is difficult to argue that the participation of an industrial lab in

V. Murdock (✉)
Yahoo! Research Barcelona, Barcelona, Spain
e-mail: vmurdock@yahoo-inc.com; vanessa@cs.umass.edu

TREC increases communication from industry to academia because it is most likely that commercial participants are participating with an academic system, and not a production system. By the same token, it is not clear that TREC has effected the technology transfer from research to commercial products, as the document collections and topics provided as part of TREC do not compare either in nature or in scale to those used by commercial systems. This could not have been predicted in 1991 when the goals were made concrete, as the commercial uses of information retrieval had not yet taken their modern form.

The second chapter explains the decisions made in the development of the test collections, which helps to understand why the collections look the way they do, and underscores some of the assumptions about the data that could potentially influence research using this data. The third chapter presents an explanation of the metrics used in TREC. A useful contribution of this chapter is a sensitivity analysis of the metrics, in Sect. 3.2.1, and the analysis of the stability of various metrics, presented in Sect. 3.2.2. These results were presented in the Proceedings of the Query Track from TREC 9 (Buckley 2001).

The second section of the book serves both as a reference for the details for each track, and an overview of the trends for a given task over the years the track ran. One surprising outcome apparent from the ad hoc, routing/filtering and interactive tracks was that manual intervention in the query formulation performed worse than automatic query processing. As put by Robertson and Callan (Chap. 5, Page 116), “Whatever skills a human may bring to searching, they are no match for the raw statistical power of learning from sufficient examples.”

Chapter nine looks at fundamental questions about the degree to which the TREC paradigm can be used to evaluate Web search, for example whether Internet searching can be evaluated by relevance judgments. People browse the Web for many reasons not related to information seeking, and they click on search results and advertisements for reasons other than topical relatedness. These issues were addressed in TREC by choosing the queries and defining the tasks carefully so that the information need was somewhat more clear, at the expense of realism or variety in the types of searches modeled. The authors are careful to note that the results from the Web tracks may not be extrapolated to actual Web search. The chapter makes clear the significant differences between Web search in general, and Web search in the TREC context, with an extensive analysis of how the TREC Web experiments are different than the actual Web.

Chapter nine also provides an overview of findings from different types of searches, recognizing that uses for search engines differ from the ad hoc document retrieval task. The TREC Web experiments were a good start, but the question remains whether any of the conclusions about the effectiveness of TREC systems compared to commercial search engines can be applied to the Web in general, given the task definitions, and the nature of the collections. The chapter makes the point that the definition of relevance in the context of search on the Web at large is not applicable: we look instead at utility, or importance, popularity, metrics that attempt to measure “clickability”.

The third section gives overviews of the efforts of specific groups. Research groups tend toward a unifying philosophy or approach, and the overviews provide an opportunity to see how a retrieval philosophy evolved over the years, as the tasks informed the approaches and the approaches informed the tasks. The University of Massachusetts (Chap. 11) which took an inference net approach (the INQUERY system (Callan et al. 1992)), the Okapi system (Chap. 12) which developed the successful BM25 ranking function (Robertson et al. 2003), and the PIRCS system (Chap. 14) provide overviews that are broader in scope, and thus more informative. Language modeling proved to be a flexible and popular framework for retrieval systems. Although pioneered at the University of Massachusetts (Ponte and Croft 1998)), Chapter 16, written by Djoerd Hiemstra and Wessel Kraaij, describes extensions to this by European Twenty One project.

In the epilogue, Karen Sparck Jones discusses the utility of TREC in an information retrieval landscape dominated by Web search engines. The chapter summarizes the trends indicated by TREC, and most importantly, suggests a way forward for IR research in the TREC setting, related to the Web. She recommends that TREC continue along its existing lines by promoting ad hoc retrieval and variations for new types of information need or material, and by tracking other individual information seeking tasks. She pushes for the use of the Web as a resource, as has been done by Brill et al. (2001). Sparck Jones also advocates for problem definitions to be more Web-like in nature, so that research in IR will more closely represent the tasks IR is used for in the real world. This final chapter is perhaps the most important and interesting chapter in the book as it provides at once a survey of what has been done from the early 1960s to the present day, and a window into the future of IR, with recommendations for where to take the state-of-the-art next.

References

Brill, E., Lin, J., Banko, M., Dumais, S., & Ng, A. (2001). Data-intensive question answering. In *Proceedings of TREC-2001* (pp. 393–400).

Buckley, C. (2001). The TREC-9 query track. In *Proceedings of TREC-9* (pp. 81–85).

Callan, J., Croft, W. B., & Harding, S. (1992). The INQUERY retrieval system. In *Proceedings of the 3rd international conference on database and expert systems application* (pp. 78–83).

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the twenty-first international ACM SIGIR conference* (pp. 275–281).

Robertson, S., Walker, S., Hancock-Beaulieu, M., Gull, A., & Lau, M. (2003). Okapi and TREC. In *Proceedings of TREC-1* (pp. 21–31).