The Lucene for Information Access and Retrieval Research (LIARR) Workshop at SIGIR 2017

Leif Azzopardi,[1] Matt Crane,[2] Hui Fang,[3] Grant Ingersoll,[4] Jimmy Lin,[2]
Yashar Moshfeghi,[5] Harrisen Scells,[6] Peilin Yang,[3] and Guido Zuccon[6]
[1] University of Strathclyde [2] University of Waterloo [3] University of Delaware
[4] Lucidworks [5] University of Glasgow [6] Queensland University of Technology

1 INTRODUCTION

As an empirical discipline, information access and retrieval research requires substantial software infrastructure to index and search large collections. In the spirit of collaboration, many researchers have shared their systems with the community. An incomplete list includes SMART [18], Lemur/Indri [15, 16], Galago [8], Terrier [14, 17], ATIRE [20], Ivory [12], JASS [13], MG4J [6], Wumpus [21], and Zettair [25]. Academic IR systems are primarily designed to advance some particular research objective—in most cases, better retrieval effectiveness as measured by standard test collections, but in others cases, more efficient query evaluation. By their nature, these toolkits are often not as well engineered or feature complete as production search engines; common issues include scalability challenges when indexing large web collections, the inability to fully take advantage of modern multi-core processors, and limitations in ingesting heterogeneous content. Although some academic systems enjoy adoption across multiple institutions, many researchers use only the systems they have developed.

On the other hand, with the exception of a small number of companies (e.g., commercial web search engines), the open-source Lucene system and its derivatives such as Solr and Elasticsearch have become the *de facto* platform for deploying search applications in industry. For convenience, we refer to software in the broader Lucene ecosystem simply as "Lucene" here. Examples of prominent deployments include LinkedIn, Twitter, Bloomberg, as well as a number of online retailers and many large companies in the financial services space. Lucene has achieved broad adoption, successes in production deployments, and a vibrant open-source community. However, it is poorly suited for information retrieval research (for a variety of reasons, discussed below) and hence has been under-utilized by the research community.

This workshop is motivated by the desire to better align information retrieval research with the practice of building search applications from the perspective of open-source information retrieval systems. We believe that better alignment can lead to richer academic–industrial collaborations, more efficient knowledge transfer of research innovations, and greater reproducibility of research results. In principle, there are two approaches to achieving this goal: by promoting greater use of academic IR systems in industry or by adapting Lucene to better support information retrieval research. The second option seems far more realistic, and therefore the goal of this workshop is to promote the use of Lucene for information access and retrieval research. Specifically:

- We wish to gain a better understanding of "the barriers to entry" of using Lucene for information retrieval research. Why are researchers currently *not* using Lucene?
- We aim to address each one of these barriers through sharing code, documentation, guidelines, best practices, and experiences.
- We hope to develop a community roadmap of what needs to be accomplished to further facilitate broader adoption of Lucene by the research community.

2 BACKGROUND AND RELATED WORK

The IR community has a longstanding interest in sharing systems to support research, which can be traced back to the version of Cornell's SMART system [7] from the mid 1980s.

In 2005, a workshop on Open Source Web Information Retrieval (OSWIR) was held in association with the 2005 IEEE/WIC/ACM International Conferences on Web Intelligence & Intelligent Agent Technology [5]. This was followed up by a workshop on Open Source Information Retrieval (OSIR) at SIGIR 2006, which provided a "forum that allows open source developers, consumers, and researchers to interact to coordinate their efforts" [24]. A follow-up workshop was held at SIGIR 2012 [19].

More recently, Azzopardi et al. [4] organized Lucene4IR,[1] a workshop that brought together researchers and developers to discuss, plan, and develop a common set of teaching and training resources for students and researchers wishing to use Lucene for information retrieval research. The event included hands-on sessions illustrating how to use Lucene to perform typical IR operations (i.e. indexing, retrieval, etc.) as well as how to extend and modify Lucene to extract term statistics, implement different ranking models, etc.

A related thread is the community's aspirations toward reproducible research. Armstrong et al. [3] previously identified the prevalent problem of weak baselines in experimental IR papers. Proposed solutions to this problem include common environments for sharing research results [2, 9] and competitive baselines that are open source and easily replicable. The latter thread ties directly into the goals of this workshop. The workshop on Reproducibility, Inexplicability, and Generalizablity of Results (RIGOR) [1] was held at SIGIR 2015, a part of which focused on reproducing results in different open-source IR systems. These efforts were expanded into the open-source reproducibility challenge [11] that brought together developers of open-source search engines to provide reproducible baselines of their systems in a common cloud-based

SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan
© 2017 Copyright held by the owner/author(s). 978-1-4503-5022-8/17/08.
DOI: http://dx.doi.org/10.1145/3077136.3084374

[1] https://sites.google.com/site/lucene4ir/home

execution environment. A total of seven systems participated, and the product of the exercise is a repository that contains all code necessary to generate competitive *ad hoc* retrieval baselines on the gov2 test collection, such that with a single script, anyone with a copy of the collection can reproduce the submitted runs. Further work along these lines includes the RISE platform [22] built on top of Indri, and discussions around reproducibility continued with a 2016 Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science" [10].

This workshop represents a natural continuation of dialogue within the community around the issues discussed above, but we advocate a specific path forward: that the IR community adopt Lucene as the default toolkit for research studies.

3 PERCEIVED AND REAL BARRIERS

To jumpstart the discussion, we have compiled a list of "complaints" about using Lucene for information retrieval research. We are quick to emphasize that these are entirely anecdotal and reflect the experiences of the workshop organizers. Some issues are merely matters of perception, but they are nevertheless important to highlight.

1. Lucene cannot run *ad hoc* retrieval experiments right out of the box. Much work in IR research is organized around test collections from TREC, CLEF, NTCIR, etc., for which Lucene does not provide any built-in support.
2. Lucene is not effective. For the longest time, Lucene severely lagged behind in providing modern ranking functions. For example, Okapi BM25 was not added to Lucene until 2011.[2]
3. Lucene is not efficient. Because Lucene is written in Java, there is the perception that it is slow, particularly when scaling up to modern web collections.
4. Lucene is difficult to use and has poor documentation for system internals. The low-level abstractions for document scoring, postings traversal, and accessing terms statistics are confusing and poorly documented.

To a large extent, the first issue has been alleviated by Lucene4IR and Anserini [23], a project that grew out of the reproducibility challenge discussed above. With respect to the second and third points, these perceptions are no longer accurate today: empirical studies [11, 23] have shown that the effectiveness of Lucene's baseline retrieval models is at least as good as those in academic IR systems, and that Lucene is capable of high-throughput multi-threaded inverted indexing as well as low-latency query evaluation.

4 WORKSHOP AGENDA

In bringing together researchers and developers for our workshop, it is not our intention to organize a "mini-conference"-style event with a set of talks around refereed contributions, but to rather foster direct interactions among the participants. The workshop will be organized more along the lines of a hackathon where attendees work with Lucene in a hands-on capacity, for example, to explore its feature or to rapidly prototype new functionalities.

Nevertheless, we believe that some presentations remain necessary to structure the discussion, and to that end, we have (short) talks on the following topics planned:

- An overview of the Lucene4IR effort.[3]
- A presentation about Anserini, a recent effort to build a research IR toolkit around Lucene.[4]
- An overview of using Elasticsearch for IR experiments.
- A walkthrough of Lucene internals from the ranking "inner loop" to Solr's learning-to-rank capabilities.

The intended outcomes of this workshop are manyfold: We hope to build documentation, resources, and reusable code on how Lucene can be used for IR research today—to help researchers get started with as minimal effort as possible. We plan to develop a community roadmap to outline features that will increase usage moving forward. Finally, the workshop will build momentum and enthusiasm for rallying around Lucene as the research toolkit of choice.

REFERENCES

[1] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. 2016. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum* 49, 2 (2016), 107–116.

[2] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. EvaluatIR: An Online Tool for Evaluating and Comparing IR Systems. In *SIGIR*. 833.

[3] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In *CIKM*. 601–610.

[4] L. Azzopardi, Y. Moshfeghi, M. Halvey, R. Alkhawaldeh, K. Balog, E. Di Buccio, D. Ceccarelli, J. Fernández-Luna, C. Hull, J. Mannix, and S. Palchowdhury. 2016. Lucene4IR: Developing Information Retrieval Evaluation Resources using Lucene. *SIGIR Forum* 50, 2 (2016), 58–75.

[5] M. Beigbeder and W. Yee. 2015. OSWIR 2005 Workshop, Final Report.

[6] P. Boldi and S. Vigna. 2005. MG4J at TREC 2005. In *TREC*.

[7] C. Buckley. 1985. *Implementation of the SMART Information Retrieval System*. Department of Computer Science TR 85-686. Cornell University.

[8] M.-A. Cartright, S. Huston, and H. Feild. 2012. Galago: A Modular Distributed Processing and Retrieval System. In *SIGIR 2012 Workshop on Open Source IR*.

[9] H. Fang, H. Wu, P. Yang, and C. Zhai. 2014. VIRLab: A Web-based Virtual Lab for Learning and Studying Information Retrieval Models. In *SIGIR*. 1249–1250.

[10] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum* 50, 1 (2016), 68–82.

[11] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. 2016. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In *ECIR*. 408–420.

[12] J. Lin, D. Metzler, T. Elsayed, and L. Wang. 2009. Of Ivory and Smurfs: Loxodontan MapReduce Experiments for Web Search. In *TREC*.

[13] J. Lin and A. Trotman. 2015. Anytime Ranking for Impact-Ordered Indexes. In *ICTIR*. 301–304.

[14] C. Macdonald, R. McCreadie, R. Santos, and I. Ounis. 2012. From Puppy to Maturity: Experiences in Developing Terrier. In *SIGIR 2012 Workshop on Open Source IR*.

[15] D. Metzler and W. Croft. 2004. Combining the Language Model and Inference Network Approaches to Retrieval. *IP&M* 40, 5 (2004), 735–750.

[16] D. Metzler, T. Strohman, H. Turtle, and W. Croft. 2004. Indri at TREC 2004: Terabyte Track. In *TREC*.

[17] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *SIGIR 2006 Workshop on Open Source IR*.

[18] G. Salton. 1971. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, New Jersey.

[19] A. Trotman, C. Clarke, I. Ounis, S. Culpepper, M.-A. Cartright, and S. Geva. 2012. Open Source Information Retrieval: A Report on the SIGIR 2012 Workshop. *SIGIR Forum* 46, 2 (2012), 95–101.

[20] A. Trotman, X.-F. Jia, and M. Crane. 2012. Towards an Efficient and Effective Search Engine. In *SIGIR 2012 Workshop on Open Source IR*.

[21] Wumpus. http://www.wumpus-search.org. Accessed: 2017-05-30.

[22] P. Yang and H. Fang. 2016. A Reproducibility Study of Information Retrieval Models. In *ICTIR*. 77–86.

[23] P. Yang, H. Fang, and J. Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *SIGIR*.

[24] W. Yee, M. Beigbeder, and W. Buntine. 2006. SIGIR06 Workshop Report: Open Source Information Retrieval Systems (OSIR06). *SIGIR Forum* 40, 2 (2006), 61–65.

[25] Zettair. http://www.seg.rmit.edu.au/zettair/. Accessed: 2017-05-30.

[2] https://issues.apache.org/jira/browse/LUCENE-2959

[3] https://www.github.com/leifos/lucene4ir

[4] http://anserini.io/