Open Source Information Retrieval: a Report on the SIGIR 2012 Workshop

Andrew Trotman
University of Otago, New Zealand

Charles L. A. Clarke
University of Waterloo, Canada

Iadh Ounis
University of Glasgow, Scotland

Shane Culpepper
RMIT University, Australia

Marc-Allen Cartright
University of Massachusetts Amherst, USA

Shlomo Geva
Queensland University of Technology, Australia

Abstract

On August 16, 2012 the SIGIR 2012 Workshop on Open Source Information Retrieval was held as part of the SIGIR 2012 conference in Portland, Oregon, USA. There were 2 invited talks, one from industry and one from academia. There were 6 full papers and 6 short papers presented as well as demonstrations of 4 open source tools. Finally there was a lively discussion on future directions for the open source Information Retrieval community. This contribution discusses the events of the workshop and outlines future directions for the community.

1 Introduction

The open source Information Retrieval community has been strong for many years. Early search engines such as MG continue to be used in larger open source projects such as Greenstone. More recent open source search engines such as Apache Lucene are used to power the search facilities of some of the largest technology companies including IBM, AOL, and Apple. In the academic community, such search engines are routinely used to test ranking functions, compression algorithms, user interfaces, and so on. Open Source Information Retrieval is now an essential component of research and commerce.

Position papers, posters, and demos of open source Information Retrieval were sought. Topics included, but were not limited to: Software Engineering; Hardware Engineering; Evaluation; Needs, Desires, and Protocols. Of particular interest was how to work together to build OpenSearchLab, an open source, live and functioning, online web search engine for research purposes. The tools to build OpenSearchLab mostly already exist and by working together it may be possible build such a site and in doing so to transform the future of research in Information Retrieval.

In total 6 full papers, 6 short papers, and 4 demonstrations were presented at the workshop. There was lively discussion on OpenSearchLab and other ways that the open source movement could engage the community. This contribution outlines the events of the workshop

2 Invited Talks

Two invited talks were given. The first was given by from the commercial perspective of Grant Ingersoll from LucidWorks who discussed open software. The second was given from the academic perspective of Jamie Callan from Carnegie Mellon University who discussed open data.

Grant Ingersoll[1] outlined the Apache philosophy and in doing so highlighted the difference between open source and open development. Opening the source of a project is only one stop towards opening the entire development process including design decisions, corpora, evaluation, and research. Indeed, under the Apache philosophy every detail of the development process should happen in a public forum such as a publicly accessible discussion group.

He went on to discuss some of the issues in building OpenSearchLab including the securing of funding. Any ongoing concern requires funding from some source and obtaining such funds is difficult in an academic environment.

Next the Lucene ecosystem was presented. This included the Lucene search library on which the Solr search engine is built. The Nutch web crawler and the Hadoop implementation of MapReduce with distributed file system. The ecosystem also has the Mahout data mining tool supports clustering, classification, and collaborative filtering, the Tika toolkit for extracting content from different MIME types, and the openNLP natural language parser.

Finally, Ingersoll presented a pluggable services-based architecture for building OpenSearchLab. This was based on the proven Lucene tools; however, the pluggable architecture would support other tools from other open source vendors.

Jamie Callan[2] outlined the Lemur project and the ClueWeb collections including the soon to be released ClueWeb12.

Lemur is a joint collaboration between Carnegie Mellon University and University of Massachusetts Amherst (Callan, Croft, and their students). It includes software, datasets, and services to support research (including the Lemur Toolbar, and an interactive tool to search the ClueWeb collections). Problems with attracting funding for a long-term community oriented project were discussed. Many funding agencies have short-term funds available and this makes it difficult to secure skilled personnel to work on the project full-time. This means the focus of Lemur must constantly change, which is difficult for an established project.

The history of ClueWeb was outlined. The ClueWeb09 crawl was funded by the NSF's Cluster Exploratory (CLUE) program which provided labor and disk storage. The crawl was done on a dedicated Google/IBM cluster and took several months. A modified version of Nutch was used on a cluster of 100 machines running Hadoop. A total of 33TB of disk storage was available and a 1GTB/s network was used (although it proved problematic to saturate this). The crawl itself was performed between January and February 2009, is about 25TB

[1] Ingersoll's slides are available online: `http://www.slideshare.net/gsingers/opensearchlab-and-the-lucene-ecosystem`

[2] Callan's slides are available online: `http://opensearchlab.otago.ac.nz/SIGIR12-OSIR-callan.pdf`

in size and contains over a billion web pages. To date over 250 copies have been distributed worldwide.

The ClueWeb12 crawl was seeded with over 2 million URLs including the highest PageRank non-spam URLs from ClueWeb09, over 250 most popular English URLs from Alexa, and nearly 6,000 travel sites. During the crawl blacklisted (by URLBlacklist.com) URLs were removed as were sites that opted out. Non-text URLs were removed, but all objects in all crawled pages were stored, in other words images in <img> tags were downloaded but images pointed to by <a> tags were not. Files were truncated at 10MB.

The Heritrix crawler was chosen for ClueWeb12. It was run on a 7 node cluster in CMU between February and April 2012. At the same time a twitter feed was monitored for URLs and those were added to the crawl. In total 1.2 billion pages amounting to 37TB text (and 67 TB of other files) was crawled in 13 weeks. The collection is expected to be released in September 2012, to be released on disk, and to be distributed at a low cost (the cost of disks and shipping).

3 Papers

In total 6 full papers and 6 short papers were accepted for presentation. They are outlined in this section. The interested reader is referred to the proceedings[3] for full details.

3.1 Search Engines

Several search engines were presented. The presenters not only outlined the features of their search engines but also some of the engineering problems they addressed while building their systems. Only two programming languages were seen, Java and C++, but the problems seen by all authors were similar.

The Apache Lucene Java search library [4] on which Apache Solr is built is released under the Apache Software license version 2. The project was started in 1997 by Doug Cutting as a means to learn Java, but now at version 4 it has evolved into a feature-rich search system used by many commercial organizations. Lucene supports searching structured and semi-structured documents as well as proximity searching. The index supports addition and deletion of documents. The search engine supports ranked queries, Boolean queries, fielded queries, and indexes can be updated while in use; the updates are near real-time.

The Galago search engine [5] is also in Java. It was written at the University of Massachusetts Amherst as the successor to Indri. The aim was to build a search engine that worked well in a highly distributed environment while also being componentized at almost all parts of the search model (via JAR files). Galago has been released under the BSD license.

The Terrier search engine [9] from the University of Glasgow is already well known at SIGIR. It, too, is written in Java. It has been released under the Mozilla Public License. Terrier was designed to be flexible and extensible making it an ideal platform for academic search engine research. Terrier supports structured and semi-structured retrieval as well as proximity. It also works in a highly distributed environment and is currently being updated for near real-time indexing.

The SMART search engine [1] is a localized search engine designed to retrieve entitles from people interacting with social media. Such a query might be where can I get good a

[3] `http://opensearchlab.otago.ac.nz/FullProceedings.pdf`

good coffee now? to answer this the search engine must integrate social media information such as user recommendations as well as local data such as time and location. SMART is built on Terrier.

The ATIRE search engine [10] is written in C++ by staff and students at the University of Otago and Queensland University of Technology. It was build with the aim of producing a solid baseline for academic research. To do so it implements state of the art techniques. The authors have produced a fast and effective search engine that is not as feature rich as some of the other search engines, but does have tiny indexes, searches quickly, and has competitive precision. ATIRE is released under the BSD license and runs on Windows, Linux, and MacOS.

A comparison of Lucene and Indri was presented [11]. This comparison not only pitted the two head-to-head, but interestingly it also compared two different versions of the search engines and discussed how they had changed over time. The comparison was by researchers experienced in both search engines who were testing long and short queries on standard TREC test collections.

3.2 Beside Search Engines

Several tools to help in the building of search engine were discussed. These varied from phonetic matching algorithms to systems for measuring search engine performance.

TIRA [6] is an online system designed to simplify the process of reproducibly measuring the performance of information retrieval systems, and to build a community around doing so. It was recently used at the PAN plagiarism detection forum and the authors suggest such a system might be used by Information Retrieval forums. TIRA is web based, but can be downloaded and installed for use within a single institute.

The YeSQL web crawler [7] project was built to demonstrate the simplicity of creating a web crawler backed by a relational database (PostgreSQL in this case). The crawler is distributed and uses the database as backing store for URLs and to prioritize those URLs for crawling. The YeSQL crawler is fewer than 1,000 lines of C code and was tested during the 2012 French presidential elections where it was able to crawl 20 million URLs in a few days. YeSQL is released under the GNU public license.

Japanese phonetic matching extensions to the PostgreSQL relational database were presented [12]. Such a tool can be used to identify spelling errors in Japanese and in doing so to conflate such works in a search engine index. Experiments with the NTCIR-9 Japanese Intent collection demonstrate that the system does, indeed, conflate such words. Showing an improvement in search engine performance is left for future work.

3.3 Beyond Search Engines

The WikiQuery system [13] was designed for archiving complex CNF Boolean queries. Users are often faced with long search sessions resulting in one online-document that fulfills their information need. WikiQuery is a place to store those queries along with links to the resultant documents; and also to the search engine that found them. The system is collaborative and online. The authors showed that using CNF Boolean queries was more effective than keyword searching.

ezDL [3] is a system for searching in digital libraries and a framework for developing custom interactive search systems. It comes from the University of Duisburg-Essen and is a re-engineering of the Daffodil system. The system is modular and supports different search engines and different user interface components; it has a modular service-oriented architecture. ezDL was used in the INEX 2010 interactive experiments, but also in some domain specific libraries including biological sciences and computer sciences.

YouSeer [8] is an integration of the Internet Archive's open-source Heritrix web crawler and the Apache Lucene/Solr search engine. The system from the Pennsylvania State University was build for vertical search, and in building such a system the authors were faced with the challenges of how to build a complete system (from crawler to user interface), problems not faced by the authors of each individual component. The authors propose a pluggable architecture for communicating between modules.

4 Demonstrations

Four demonstrations were given. The ezDL[4] digital library (see Section 3.3) was demonstrated by Beckers, Dungs, Fuhr, Jordan, Kriewel, & Tran. Of particular note was image searching as well as text searching. The demonstration ran over the wireless Internet provided at the conference and connected to a server in Germany. SDSL[5], the Succinct Data Structure Library, was demonstrated by Gog, Petri, Culpepper & Moffat. SDSL is a library implementing several succinct data structures and algorithms that are being used by researchers at RMIT University and University of Melbourne for their work on self-indexing structures for information retrieval. Gollub demonstrated TIRA[6] (see Section 3.2) the online tool for capturing, tracking, and building a community around information retrieval performance measurement. Finally, Trotman, Jia & Crane demonstrated the ATIRE[7] search engine (see Section 3.1) searching ClueWeb09 Category A on their laptop (and separately the searching of NTCIR and INEX collections).

5 The Workshop Discussion Session

SIGIR Workshops provide a unique forum in which participants can discuss common interests, and the Open Source Information Retrieval workshop was no exception. An hour was allocated at the end of the workshop for participants to discuss a common future, including the building of an open source web search engine. The session started with a very brief discussion-stimulating presentation by Trotman and Cartright, and was cut short by the close of day. Discussion centered on many topics.

The difficulty of reproducing experiments was considered to be a problem the community needed to urgently address perhaps a topic of focus due to the SIGIR 2012 panel on Proprietary Data. The community might address this by creating a repository in which experimental configurations (including source code) could be deposited once a paper had been accepted. This would make it possible for others to download and exactly reproduce results and to build on them, especially if links from papers to the repository were given. Standard reference implementations could be added such as lexical analyzers, stemmers, and

[4] http://ezdl.de & http://www.is.inf.uni-due.de/projects/khresmoi/index.html.en

[5] https://github.com/simongog/sdsl

[6] http://tira.webis.de

[7] http://atire.org

implementations of ranking functions such as BM25. Placing these on a wiki (or a public version control system) could lower the entry barrier to search engine research. It might also allow experiments in which the effect of pluggable components (such as the lexical analyzer) could be measured.

Discussion on the difficulty of performing search engine experiments under increasing collection sizes and decreasing budgets was had. For example, it is not-trivial to obtain ClueWeb09, to index it, and to create a run for TREC. The hardware necessary is expensive and the skills in managing large collections are not easily obtained (debugging a program that takes many hours to run is time-consuming). This difficulty might be overcome if some standard document collection indexes were provided online along with executables and source code to search them. Details of how to re-index (if necessary) might be added for those with the hardware. If the open source search engines included an auto-tuning mechanism then this would decrease the likelihood of papers reporting results on mis-configured baselines.

Discussion moved to the question of how to build a community. There was a clear indication by those in the room that this would be beneficial to all. Indeed, to progress from the current open source silos to something larger than any one group can achieve on their own requires cooperation. Ingersoll reminded those present of the Apache open development approach and indicated that several simple steps could be made including the creation of a public discussion group.

The greater engagement of the research community and the commercial community came under discussion. It was suggested that researchers attend the Lucene conference to gain greater insights into the commercial needs and driving forces. It was also suggested that industry players (including Lucene) consider setting grand challenges for researchers to address, while themselves focusing on implementation. Doing so would take advantage of the research skills of the research community and the implementation (and customer facing) skills of the commercial open source industry.

A brief discussion of the Community Evaluation Service outlined in Section 5.5 of the SWIRL-12 report [2] resulted in the conclusion that funding would be an issue. It would require external on-going infrastructure which is hard to attract, or alternatively draw funds from participants which would only attract well-funded groups. However, there are some behind-the-scenes initiatives addressing these issues.

Little discussion was had on building a web scale open source search engine. It is, perhaps, premature to contemplate such a service while the main players are currently working independently.

A show of hands demonstrated very strong support for a follow-up workshop.

References

[1] M-D. Albakour, C. Macdonald, I. Ounis, A. Pnevmatikakis, and J. Soldatos. Smart: An open source framework for searching the physical world. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 48–51, 2012.

[2] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. In *SIGIR Forum*, pages 2–32, 2012.

[3] T. Beckers, S. Dungs, N. Fuhr, M. Jordan, and S. Kriewel. ezdl: An interactive search and evaluation system. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 9–16, 2012.

[4] A. Bialecki, R. Muir, and G. Ingersoll. Apache lucene 4. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 17–24, 2012.

[5] M.-A. Cartright, S. Huston, and H. Field. Galago: A modular distributed processing and retrieval system. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 25–31, 2012.

[6] T. Gollub, S. Burrows, and B. Stein. First experiences with tira for reproducible evaluation in information retrieval. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 52–55, 2012.

[7] P. Jourlin, R. Deveaud, E. Sanjuan-Ibekwe, J.-M. Francony, and F. Papa. Design, implementation and experiment of a yesql web crawler. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 56–59, 2012.

[8] M. Khabsa, S. Carman, S. R. Choudhury, and C. L. Giles. A framework for bridging the gap between open source search tools. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 32–39, 2012.

[9] C. Macdonald, R. McCreadie, R. L.T. Santos, and I. Ounis. From puppy to maturity: Experiences in developing terrier. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 60–63, 2012.

[10] A. Trotman, X. Jia, and M. Crane. Towards an efficient and effective search engine. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 40–47, 2012.

[11] H. Turtle, Y. Hegde, and S. Rowe. Yet another comparison of lucene and indri performance. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 64–67, 2012.

[12] M. Yasukawa, J. S. Culpepper, and F. Scholer. Phonetic matching in japanese. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 68–71, 2012.

[13] L. Zhao, X. Liu, and J. Callan. Wikiquery - an interactive collaboration interface for creating, storing and sharing effective cnf queries. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 1–8, 2012.