

SIGIR06 Workshop Report: Open Source Information Retrieval Systems (OSIR06)

Wai Gen Yee

Information Retrieval Laboratory
Illinois Institute of Technology
10 W 31st Street, Chicago, IL 60616
USA
yee@iit.edu

Michel Beigbeder

Ecole Nationale Supérieure des Mines de Saint-Etienne
159 cours Fauriel, 42023 Saint-Etienne Cedex 2
France
mbeig@emse.fr

Wray Buntine

Complex Systems Computation Group
Helsinki Institute of Information Technology
P.O. Box 9800, FIN-02015 HUT
Finland
wray.buntine@hiit.fi

Abstract

This paper summarizes the activities and discussions of the 2006 Open Source Information Retrieval Workshop, held in conjunction with the 2006 ACM SIGIR Conference. It summarizes the technical program and panelist discussion on the Workshop.

1 Introduction

Open source systems have the recognized advantages over those of closed source in that their internal workings are subject to public scrutiny and can be adapted to particular commercial or experimental uses. In theory, more robust and better-performing systems could be built as a result.

These advantages are recognized by academia, industry, the government, and the private consumer. The availability of open source decreases development time and makes functionality transparent. For example, researchers enjoy the ability to replicate experimental results and private industrial users can economically customize these systems to their business needs, such as the search of their databases. These factors are demonstrated by the significant adoption of open source search tools within specialized search communities, such as Wikipedia, Technorati, and University Web sites. Community support is also demonstrated by a Web site that specializes in open source search technologies [4]. What has been missing, however, is a forum that allows open source developers, consumers, and researchers to interact to coordinate their efforts.

The goals of the 2006 Open Source Information Retrieval Work-shop (OSIR06) are to offer exposure to openly available systems, data, and methodologies for the practice and research of information retrieval, to increase the use of such systems, and to encourage collaboration among the developers and consumers of such systems [2]. OSIR06 was held on August 10, 2006 in conjunction with the 2006 ACM SIGIR Conference. It was a follow-on to the successful 2005 OSWIR Workshop held in conjunction with the 2005 Web Intelligence Conference [1].

Over 30 people attended the Workshop, representing academia (e.g., University of Glasgow, University of Massachusetts), industry (e.g., AOL, Monster, Yahoo!), and the government (e.g., the NSA). Interaction among the participants was lively and, we believe, relationships were initiated or reinforced.

The OSIR06 program consisted of two parts, paper presentations and a panel. The presented papers described recent advances to well-known systems, experiences, new applications, metrics, and standards. These systems represent a general cross-section of the research directions and focus in system development.

The panel, made up of experts from academia and industry, gave the audience an opportunity to discuss the day's progress and ask burning questions to the panelists as well as other audience members.

2 Summary of Technical Presentations

The paper presentations were divided into three broad categories: Systems, Distributed IR and Users, and Miscellaneous Open Source IR issues.

The Systems session featured three talks. The first was *Low Latency Index Maintenance* in Indri, by Trevor Strohman and W. Bruce Croft, presented by Trevor Strohman of the University of Massachusetts. This paper addresses the problem of the unavailability of search engine querying facilities during index maintenance. It suggests that the important metric is not the speed of indexing a set of new documents but rather the time it takes for a new document to be available for querying – its “latency.” Reducing latency is achieved by two basic methods: creating a main memory index for recently inserted items, and not allowing blocking due to disk I/Os. The former reduces latency for new documents, and the latter is implemented by separating the process of document parsing from document insertion into an index. Furthermore, disk data is read-only, so they never need to be locked. One outstanding question is the scheduling of index updates given user and document change behavior.

The second talk in the Systems session was *PF/Tijah: Text Search in an XML Database System*, presented by Henning Rode from the University of Twente. PF/Tijah is the integration of the PathFinder (PF for short) XQuery system with the Tijah XML information retrieval system and is part of the MonetDB/XQuery system. PF/Tijah allows information retrieval on XML documents that rank based on text as well as document structure. PF/Tijah also allows the combination of database and information retrieval queries into an integrated result. In adding ranking XQuery results, PF/Tijah is attempting to extend the emerging XQuery standards.

The final talk in the Systems session was *Terrier: A High Performance and Scalable Information Retrieval Platform*, by Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma, presented by Vassilis Plachouras. This talk described Terrier, the search system developed at the University of Glasgow, and focused on its open source subset. Terrier has a modular design that supports the indexing of many file formats. Supporting a format (e.g., Microsoft Word) merely requires the implementation of a plug-in. Terrier features index compression, modular design, and flexible ranking models. The Terrier deployment includes ready-to-run applications, such as a desktop search tool.

The Distributed IR and Users session featured four talks. The first of these talks, *Distributing Lucene on the Grid*, by Edgar Meij and Maarten de Rijke, was presented by Edgar Meij. This talk described an experience with deploying a search engine on a widely distributed grid. Middleware was developed to allow different grid nodes to handle the indexing and search load. However, it turns out

that the communication overhead overwhelmed of the grid made it slightly worse for indexing and significantly worse for search than a centralized system. More tests are required, however, to demonstrate the true scalability of GridLucene.

The second talk in the Distributed IR and Users session was *IR-Wire: A Research Tool for P2P Information Retrieval*, by Shefali Sharmi, Linh Thai Nguyen, and Dongmei Jia, presented by Linh Thai Nguyen. This talk described the emerging area of peer-to-peer information retrieval and showed how a commercial P2P file-sharing system (LimeWire's Gnutella) has been adapted to be an information retrieval research tool. The system, IR-Wire implements IR-style ranking and can crawl P2P networks for data useful to research. Some initial results indicate that many of the query properties are similar to those of the Web (e.g., query length distribution, query popularity distribution).

The third talk in the Distributed IR and Users session was *Tagging in Peer-to-Peer Wikipedia, A Method to Induce Cooperation*, by Jenneke Fokker, Wray Buntine, and Johan Pouwelse, presented by Jenneke Fokker. This talk described a way of merging the information richness of a system like Wikipedia with the sharing technology afforded by P2P file-sharing systems. They help each other by increasing the search comprehensiveness of file-sharing systems and enriching the Wikipedia experience with multimedia. Some of the unique questions addressed are sociological: how to encourage people to annotate their data.

The final talk in the Distributed IR and Users session was *Web Recommender System Implementations in Multiple Flavors: Fast and (Care-)Free for All*, by Olfa Nasraoui, Zhiyong Zhang, and Esin Saka, presented by Olfa Nasraoui. This talk described how to adapt an open source Web search engine, Nutch, to provide navigation recommendations to Web users. Being able to design such a system puts the power of recommendations into the hands of any user. Presented were promising results of a system for content-based filtering recommendations. Notably, this work epitomized the power of open source systems: it demonstrated how open source can aid in the rapid development of very practical systems.

The third session, on miscellaneous topics, featured three talks. The first talk was *An Effectiveness Measure for Evaluating Open Retrieval Systems*, by Hsieh-Chang Tu and Jieh Hsiang, presented by Hsieh-Chang Tu. This talk aimed to address a problem with traditional information retrieval metrics (e.g., a recall-precision curve) when testing over "open" document repositories (e.g., the Web): one does not know the true number of relevant pages for each query on the Web, so it is difficult to measure recall accurately. The proposed "relnum-precision" plot addresses this problem; it measures precision at a given number of top relevant documents, not requiring knowledge of the true number of relevant documents.

The second talk of the miscellaneous session was *Show and Tell: A Seamlessly Integrated Tool for Searching with Image Content and Text*, by Zhiyong Zhang, Calos Rojas, Olfa Nasraoui, and Hichem Frigui, presented by Olfa Nasraoui. This talk described how Nutch/Lucene was used to create an image search engine that combined keyword based and content-based search for the images. Color properties of the images were mapped to "color words," which could be added to the queries like ordinary terms either manually or with a query-by-example-like interface. Using both keywords and image information improves the precision of the results.

The final talk on of the miscellaneous session was *Standards for Open Source Retrieval*, by Wray Buntine, Michael P. Taylor, and Francois Lagunas, presented by Wray Buntine. This talk rounds out the day, describing standards relevant to information querying, collecting, and publishing. It argued the benefits of standardization and described how some standards arise. It also presented the case for the increasing importance of information extraction to information retrieval, and for common standards in query and result formats to support distributed information retrieval.

3 Summary of Panelist Discussion

The panelists were Doug Cutting, of Yahoo!, Vassilis Plachouras, of the University of Glasgow, Trevor Strohman, of the University of Massachusetts, and Hugo Zaragoza, of Yahoo!. They discussed with the audience issues of open source project sustainability, application to industry and academia, as well as various experiences. Industrial participants, who run specialized search engines in their enterprises, were well-represented in the Workshop. A significant part of the discussion during the panel was devoted to hearing their needs.

What emerged from the discussion was that sustainability of open source requires many factors. First, the public must be assured that there will be continuing support for the product either by the developers or by the community. For industry acceptance, relaxed licensing (e.g., Apache public license) is necessary to allow companies to build and incorporate open source systems. (For a discussion on license decision-making, see [3].) One audience member mentioned that he would have liked to incorporate some open source code he found that was developed by an academic institution. However, the bureaucratic terms of the license discouraged him: each license had to be individually negotiated. Easier, he added, would have been for the license to simply ask for a lump sum of money to take a portion of the code for proprietary use.

Academic groups also represent a community for open source systems development. Given the multiple demands of the student, student turnover, and the unpredictable nature of funding of such initiatives, relying on an academic community may be risky. However, academic developers also represent a substantial opportunity for innovation, as the Indri and Terrier systems attest.

Another factor contributing to the wide acceptance of a system is the recognition of the particular needs of the intended “customer.” Certainly, open source must be well-designed and well-documented, but, for industry, overall throughput and functionality is the key factor, whereas research tools tend to focus in certain proof-of-concept features, while minimally implementing others. The fact that several open source search engines exist attests to existence of design goals and therefore various implementation directions. As noted by Doug Cutting, the heavy users of open source systems tend to be key developers, so practical systems should be conscious of their needs.

A third factor in the adoption of an open source system is its adherence to standards. The adoption of standards is particularly important to open source projects as they tend to be developed several independent groups simultaneously. For example, in the OSIR talks, PF/Tijah argued about its decision to use or extend the XQuery standards. Terrier is designed to be able to read from several standard data formats.

Lucene, as a case study, has wide adoption because it uses the permissive Apache license and is built with industrial applications in mind. Moreover, because of its popularity, it has an active community that ensures its continued evolution and maintenance. Furthermore, given its stature, Lucene has been the building block of several industrial and academic applications, and has created some de facto standards (e.g., its query syntax).

Open source is also amenable to education. Its open nature allows students to play with certain functionality on a working system. The University of Glasgow, for example, has used their Terrier system for formal instruction. We expect education to be an increasingly important application of open source systems.

Certain issues not touched on are the transfer of intellectual property contained in open source. Some licenses allow the use of known intellectual property (e.g., patented algorithms) within code, but what would happen if open source unwittingly violated patents?

Another topic not covered was the development of an open source based search engine with global coverage, as discussed earlier [1][5]. Several of the attendees represented companies running specialized search engines, and this is the application where open source search tools are having their greatest impact. A global search engine would be yet another application and its development may be conducive to the global, distributed nature of open source projects.

4 Conclusion

OSIR06 was a successful gathering of academia, industry, and government, characterized by high quality presentations and lively discussions. The presented work represented a good cross section of current information retrieval system and standards development to be of use to the community of researchers, developers, and customers. Furthermore, the panelist discussion on experiences and the costs and benefits of open source system development and maintenance was enlightening. Finally, it is clear from our discussions and the participating attendees that there is a healthy commercial interest in open source search, and thus a continuing need for research and development in the area.

5 References

- [1] 2005 Open Source Web Information Retrieval Workshop Web Site, www.emse.fr/OSWIR05.
- [2] 2006 Open Source Information Retrieval Workshop Web Site, www.emse.fr/OSIR06.
- [3] Choose a License, Java.net. http://java.net/choose_license.csp.
- [4] Open Source Search, www.opensourcesearch.org.
- [5] Doug Cutting in Outer-Court Blog. http://blog.outer-court.com/archive/2004_05_28_index.html.