

OSWIR 2005 Workshop, Final Report

| | |
|---------------------------------|----------------------------------|
| Michel Beigbeder | Wai Gen Yee |
| Ecole de Mines de Saint Etienne | Illinois Institute of Technology |
| Saint Etienne, France | Chicago, IL USA |
| mbeig@emse.fr | yee@iit.edu |

October 6, 2005

1 Introduction

The goal of the Open Source Web Information Retrieval Workshop (OSWIR) is to bring together practitioners in development open source search technologies to share their recent advances and to coordinate their research plans. The ultimate goal of the Workshop is to foster community-based development and distribution of transparent Web search tools.

The first OSWIR Workshop took place on September 19, 2005, in conjunction with the Web Intelligence Conference in Compiègne, France. It was organized by Drs. Michel Beigbeder of the Ecole de Mines in St. Etienne, France and Wai Gen Yee of the Illinois Institute of Technology in Chicago, Illinois. The URL of the Workshop Web page is www.emse.fr/OSWIR05.

Of the 16 papers accepted (out of 25 submitted), 14 were presented by 13 participants. Two authors could not attend. One person was an author on two papers, and therefore presented two. There were also 3 non-presenting participants for a total of 17 participants.

2 Workshop Organization and Program

The Workshop spanned the entire day, and consisted first of paper presentation, followed by a break-out session into focus groups, followed by a summary by each of the groups, including plans for future research. The goal of this organization is to allow all participants to learn first about others' work ¹, then form focus groups where they could discuss their specializations in the context of what was learned in the morning and to decide on actionable items that could lead to fruitful research. Their conclusions are to be presented at the end of the Workshop to all participants to garner the opinions of others'.

¹Note that each participant was also asked to critique three of the accepted papers in advance of the Workshop.

Presentations roughly fell into four topics: systems (3 presentations), models (3), XML (4), and miscellaneous (4). The systems group spoke about complete systems that were being built to handle Web crawling and indexing. The models group spoke about ways of representing Web data to optimize retrieval quality. The XML group focused on search on semistructured XML documents. The miscellaneous group focused on other relevant topics, such as compound-word splitting and peer-to-peer information retrieval. The presentations lasted for an hour after lunch and were characterized by lively discussions.

3 Break Out Session and Group Conclusions

After the presentations, we formed three groups: the systems/models group, the XML group, and the miscellaneous group. We combined the systems and the models groups because of their relatively small sizes and their natural relationship. The breakout session lasted approximately 90 minutes and was followed by reports on the conclusions of each group.

3.1 XML Group

Fabien Laniel presented the summary of the XML group's conclusions, and began by making the distinction between data-centric and document-centric XML data. Data-centric XML data are generally for machine-to-machine interchange of machine-readable data. It is characterized as being highly structured. Document-centric XML data are generally human readable, and may contain some markup to help a person understand the text (e.g., markup indicating italics or footnotes).

They made this distinction, as three members of the group worked on document-centric XML (Hlaoua, Laniel, and Tannier), while one member worked on data-centric XML. This distinction is important, as generally, the techniques for processing and querying XML vary depending on its usage; more structured XML documents rely more on its structure to relay information, whereas document-centric XML relies more on content (i.e., the values of data).

With this in mind, Lobna Hlaoua is considering the use of Abdeslame Alilaouar's fuzzy matching of structures to group results in her relevance-feedback queries. This should give her the ability to better categorize her results, giving the user a better way of distinguishing his options for the second query.

Abdeslame Alilaouar, who worked on fuzzy logic to match XML schema, will consider using traditional IR techniques to match the textual part of data-centric XML documents.

Fabien Laniel, based on questions from the audience, will conduct experiments to test the sensitivity of the coefficients of his similarity equation over different INEX topics. Currently, a different set of coefficients is generated for each topic.

Xavier Tannier is looking into incorporating jWordSplitter, Sven Abel's project, into his XGTagger.

3.2 Miscellaneous Group

The Miscellaneous Group consisted of Sven Abels, David Parry, Katarzyna Wegrzyn-Wolska, and Wai Gen Yee. There was generally no unifying theme in this group, so each member just summarized what he will do based on what it learned during the Workshop.

Sven Abels presented work on jWordSplitter, a system that splits compound words (primarily in German). Through discussion, he has discovered many new initiatives. First, he will study the use of Bloom filters to test if a term is a member of a set, which should have strong space advantages over using regular hash tables. Second, he can use the inverse of word splitting to synthesize terms in Chinese (which may be made up of multiple characters). Third, he will study the reduction of his term dictionary to atomic words, which should save his system some space and decrease checking time. Fourth, he may consider different language specific rules in splitting terms. For example, some components of compound words may change their spelling when incorporated into another word. Finally, he will see if the incorporation of jWordSplitter into another system (e.g., Nutch, presented by Doug Cutting), would improve retrieval effectiveness.

David Parry presented work on automatic document similarity checking by using compression techniques. Based on comments from the audience about the short lengths of some of his test data, he realized that the compression tool he used has some similarity to n-grams, but the compression algorithms give more flexibility. This technique also may be useful in the clustering of data. If good centroids can be found, efficient compression can be done, and visa versa. To further his understanding of the scope of his work, he will look into the area of computational linguistics, which deals with topics such as writing style and authorship attribution. Finally, he is considering using compression dictionaries directly, making comparisons between dictionaries rather than the "black-box" approach he currently uses.

Katarzyna Wegrzyn-Wolska spoke about her measurements on the lifetimes of dynamic Web pages. In conjunction with this work, she will consider crawling algorithms that might take advantage of this information. She must also consider the definition of lifetime: how does one define that point at which a page is new?

Wai Gen Yee spoke about peer-to-peer information retrieval. One member of the audience suggests that he be careful to recognize that there exist elements in P2P networks that spoof content. He will consider how improved ranking can minimize the impact of such spoofing. Furthermore, the audience believes that P2P file-sharing is too limited in scope and that he might as well consider building a P2P Google. One system that has a similar spirit is Grub, and is worth further study.

3.3 Systems/Models Group

The Systems group was very active exchanging ideas for future work. Much of it was spurred by Wray Buntine's desire to create authorities for interest categories that could help in ranking searches within particular topics. The members settled on a discussion about DMOZ, the topic directory, which has 600,000 links to 4 million pages. Using DMOZ as a starting point, one might be able to create a list of anchor text for hyperlinks that have been shown to be effective in improving ranking when used in indexing. DMOZ might be the starting point for subject specific crawls. Finally, if a larger crawl were done, a reputation system for each of the DMOZ-linked Web sites might be created.

Another topic the Systems group discussed was the fact that many national libraries are doing Web crawls for their national archives. However, what these crawls lack is an effective search mechanism to access the archived data. There may be a market for searching these archives that the open source search community can address.

Finally, they discussed the need of an open service where researchers can have free access to code and data needed to run experiments. For example, a set of anchor text and links might be useful to small search engine companies that do not have the resources to do their own crawls. At the same time, this service might include the ability to collect anchor text data from network users who volunteer compute cycles, in a spirit similar to the one employed by Grub.

4 Conclusion

The scope of the Workshop seemed appropriate: there were some highly focused systems papers, and a smattering of odds-and-ends papers. Furthermore, the structure of the Workshop and the management of the program seemed to achieve the desired goals.

To improve the open nature of the Workshop, however, some participants suggested the requirement that submissions clearly state experimental technique, and include access to source code and experimental data to allow others to replicate results.

4.1 The Future

As a sign of the Workshop's success, most participants expressed that their experience was productive and that they are looking forward to future iterations. In fact, they expressed that, in the future, with better advertising, there should be much larger participation in the Workshop. Apparently, the open source community is larger than what was represented at the Workshop, and that there is no forum in which this community could meet. Also, this year's Workshop lacked sufficient advertisement and had too short a time frame to allow appropriate submissions. The question is, when and where should another OSWIR Workshop be held. The ACM SIGIR and CIKM conferences are natural candidates.

5 Appendix

5.1 Resources

The Workshop's Web site is at www.emse.fr/OSWIR05. On the Web site can be found the Workshop proceedings and the presentation slides.

Also available on the Workshop Web site are links Web sites and resources that are relevant to open source Web information retrieval.

5.2 Partial List of Participants

1. Sven Abels, University of Oldenburg, Germany
2. Abdeslame Alilaouar, IRIT, Toulouse, France
3. Michel Beigbeder, Ecole de Mines, St. Etienne, France
4. Wray Buntine, Helsinki Institute for Information Technology, Helsinki, Finland
5. Carlos Castillo, Universitat Pompeu Fabra, Barcelona, Spain
6. Doug Cutting, Internet Archive, CA, USA
7. Lobna Hlaoua, IRIT, Toulouse, France
8. Fabien Laniel, Ecole de Mines, St. Etienne, France
9. Jianchang Mao, Yahoo, CA, USA
10. Bruno Martins, Universidade de Lisboa, Lisbon, Portugal
11. David Parry, Auckland University of Technology, Auckland, New Zealand
12. Michael Stack, Internet Archives, CA, USA
13. Xavier Tannier, Ecole de Mines, St. Etienne, France
14. Katarzyna Wegrzyn-Wolska, Ecole d'Ingenieurs en Informatique et Genie des Telecommunications, Avon-Fontainebleau, France
15. Wai Gen Yee, Illinois Institute of Technology, Chicago, IL, USA
16. Jianhan Zhu, The Open University, Milton Keynes, United Kingdom