Defining Evaluation Methodologies for Topical Crawlers

Padmini Srinivasan*
School of Library & Information Science
The University of Iowa
Iowa City, IA 52245

Filippo Menczer† and Gautam Pant
Department of Management Sciences
The University of Iowa
Iowa City, IA 52245

Abstract

Topical crawlers are becoming important tools to support applications such as specialized Web portals, online searching, and competitive intelligence. As the Web mining field matures, the disparate crawling strategies proposed in the literature will have to be evaluated and compared on common tasks through well-defined performance measures. We have argued that general evaluation methodologies are necessary for topical crawlers. Such methodologies should (1) characterize classes of crawling tasks that support applications of different nature and difficulty, and (2) define performance measures for fair comparative evaluations of crawlers with respect to notions of precision, recall, and efficiency that are appropriate and practical for the Web. We have proposed a general framework for the evaluation of topical crawlers that relies on independent relevance judgments compiled by human editors and available from public directories. Such a framework synthesizes a number of methodologies in the topical crawlers literature and many lessons learned from several studies conducted by our group. It is our desire to generate a debate that can ultimately lead to some consensus in how to standardize the evaluation of topical crawlers, and possibly even to a crawling task in the Web Track of TREC.

Background and Motivation

Topical crawlers, also known as topic driven, focused, or preferential crawlers, are an important class of crawler programs that complement search engines. Search engines serve the general population of Web users. In contrast, topical crawlers are activated in response to particular information needs. These could be from an individual user (query time or online crawlers) or from a community with shared interests (topical or vertical search engines and portals). The crawlers underlying search engines are designed to fetch as comprehensive a snapshot of the Web as is possible; topical crawlers are designed to target portions of the Web that are relevant to the triggering topic. Such crawlers have the advantage that they may in fact be driven by a rich context (topics, queries, user profiles) within which to interpret pages and select the links to visit. Today, topical crawlers have become the basis for many specialized services such as investment portals, competitive intelligence tools, and scientific paper repositories.

One research area that is gathering increasing momentum is the evaluation of topical crawlers. The rich legacy of information retrieval research comparing retrieval algorithms in the non-Web context offers many evaluation methods and measures that may be applied toward this end. However, given that the dimensions of the crawler evaluation problem are dramatically different, the design of appropriate evaluation strategies is a valid challenge.

In a general sense, a crawler may be evaluated on its ability to retrieve "good" pages. However, a major hurdle is the problem of recognizing these good pages. In an operational environment real users may judge the relevance of pages as these are crawled allowing us to determine if the crawl was successful or not. Unfortunately, meaningful experiments involving real users for assessing Web crawls are extremely problematic. For instance the very scale of the Web suggests that in order to obtain a reasonable notion of crawl effectiveness one must conduct a large number of crawls, i.e., involve a large number of users.

Crawls against the live Web also pose serious time constraints. Therefore crawls other than short-lived ones will seem overly burdensome to the user. We may choose to avoid these time loads by showing the user the results of the full crawl — but this again limits the extent of the crawl. Next we may choose indirect methods such as inferring

*Contact author. Tel: +1-319-335-5708, Fax: +1-319-335-5374. Email: padmini-srinivasan@uiowa.edu. Partially supported by National Library of Medicine grant No. RO1-LM06909

†Current affiliation: School of Informatics and Computer Science Department, Indiana University. Email: fil@indiana.edu. Partially supported by National Science Foundation CAREER grant No. IIS-0133124

crawler strengths by assessing the applications that they support. However this assumes that the underlying crawlers are openly specified, and also prohibits the assessment of crawlers that are new.

We argue that although obtaining user based evaluation results remains the ideal, at this juncture it is appropriate and important to seek user independent mechanisms to assess crawl performance. Moreover, in the not so distant future, the majority of the direct consumers of information is more likely to be Web agents working on behalf of humans and other Web agents than humans themselves. Thus it is quite reasonable to explore crawlers in a context where the parameters of crawl time and crawl distance may be beyond the limits of human acceptance imposed by user based experimentation.

A Proposed Framework

Our analysis of the Web information retrieval literature (e.g. [1, 4, 3, 2, 5, 11]) and our own experience [6, 8, 9, 7, 12, 14, 10, 13] indicate that in general, when embarking upon an experiment comparing crawling algorithms, several critical decisions are made. These impact not only the immediate outcome and value of the study but also the ability to make comparisons with future crawler evaluations. We have proposed a general framework for crawler evaluation research that is founded upon these decisions [15]. The framework has three distinct dimensions:

Crawl task nature: How topics are defined and how seed pages and target relevant pages are identified. We have proposed a class of tasks whose difficulty is determined by the number of links separating seed pages from known relevant targets. Topics and target pages can be selected with desired specificity and broadness from a directory such as the Open Directory Project.[1]

Evaluation metrics: How to analyze both effectiveness and efficiency of crawling algorithms. For the former we have proposed a set of metrics based on known relevant targets manually classified by ODP editors, as well as "softer" criteria such as similarity to descriptions of relevant pages. These criteria lead to generalized, approximate notions of precision and recall that can be applied without knowledge of full relevant sets. For the latter we have proposed ways to ensure fair evaluations across crawlers by enforcing equal use of resources (disk, memory), by discounting variability due to network congestion, and by monitoring CPU usage to gauge algorithmic complexity.

Topical analysis: How particular characteristics of topics affect different crawlers. Given a particular topic, it would be desirable to predict which crawling strategies may be most promising by analyzing the characteristics of the topic. We have shown that the performance or certain crawlers correlates significantly with topical characteristics such as popularity and authoritativeness.

The applicability of this framework was demonstrated through the evaluation of a number of off-the-shelf crawlers from the literature [15] as well as some novel crawlers that we have designed [10, 13]. We have been able to show that for short crawls (in the order of a few thousand pages), simple strategies such as a naive best-first search crawler are most effective; for longer crawls (several tens of thousand pages and up to million pages) more sophisticated techniques, which can learn to discriminate between links in a page and evolve over time to focus on more promising neighborhoods, have a significant advantage.

Conclusion

While our proposed framework has limitations, e.g. its dependence on the availability of a directory such as the ODP, it is powerful and flexible; it allows for automatic assessments; and it yields quantitative evaluations that can be easily analyzed for statistical significance. To our knowledge it is the only general framework that has been put forth to evaluate topical Web crawlers to date. Our results [15, 10] demonstrate that the framework is effective at evaluating, comparing, differentiating, and interpreting the performance of diverse crawlers.

More important that the framework itself is a recognition of the needs that such a framework can address. The formulation of standardized crawling tasks, the design of appropriate evaluation metrics, the systematic characterization of crawl topics and tasks, and a consensus on statistical analysis tools to compare crawling techniques are in our opinion sorely needed by the Web IR community. To foster a discussion on these needs we have made publicly

[1] A.k.a. ODP or DMOZ: `http://dmoz.org`

available under the terms of the Gnu GPL a script[2] that selects topics from the ODP based on a number of parametric specifications, and generates files containing topic keywords, descriptions, and target and seed URLs in order to facilitate comparative crawler evaluations.

It is our hope that the Web IR community can use our proposals and tools as a first step toward the definition of evaluation methodologies that will ultimately allow us to advance the state of the art in Web crawling and support the next generation of search tools. We believe this process could be significantly accelerated by introducing a crawling task in the Web Track of TREC.

References

[1] CC Aggarwal, F Al-Garawi, and PS Yu. Intelligent crawling on the World Wide Web with arbitrary predicates. In *Proc. 10th International World Wide Web Conference*, pages 96–105, 2001.

[2] S Chakrabarti, K Punera, and M Subramanyam. Accelerated focused crawling through online relevance feedback. In David Lassner, Dave De Roure, and Arun Iyengar, editors, *Proc. 11th International World Wide Web Conference*, pages 148–159, New York, NY, 2002. ACM Press.

[3] S Chakrabarti, M van den Berg, and B Dom. Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.

[4] J Cho, H Garcia-Molina, and L Page. Efficient crawling through URL ordering. *Computer Networks*, 30(1–7):161–172, 1998.

[5] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and Sigalit Ur. The shark-search algorithm — An application: Tailored Web site mapping. In *Proc. 7th Intl. World-Wide Web Conference*, 1998.

[6] F Menczer. ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery. In *Proc. 14th International Conference on Machine Learning*, pages 227–235, 1997.

[7] F Menczer. Complementing search engines with online Web mining agents. *Decision Support Systems*, 35(2):195–212, 2003.

[8] F Menczer and RK Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the Web. *Machine Learning*, 39(2–3):203–242, 2000.

[9] F Menczer, G Pant, M Ruiz, and P Srinivasan. Evaluating topic-driven Web crawlers. In Donald H. Kraft, W. Bruce Croft, David J. Harper, and Justin Zobel, editors, *Proc. 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 241–249, New York, NY, 2001. ACM Press.

[10] F Menczer, G Pant, and P Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, Forthcoming, 2003. http://dollar.biz.uiowa.edu/~fil/Papers/TOIT.pdf.

[11] M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. In *Proc. 10th International World Wide Web Conference*, 2001.

[12] G Pant and F Menczer. MySpiders: Evolve your own intelligent Web crawlers. *Autonomous Agents and Multi-Agent Systems*, 5(2):221–229, 2002.

[13] G Pant and F Menczer. Topical crawling for business intelligence. In *Proc. European Conference on Digital Libraries (ECDL)*, 2003.

[14] G Pant, P Srinivasan, and F Menczer. Exploration versus exploitation in topic driven crawlers. In *Proc. WWW-02 Workshop on Web Dynamics*, 2002.

[15] P Srinivasan, G Pant, and F Menczer. A general evaluation framework for topical crawlers. *Information Retrieval*, (Submitted), 2002.

[2] `http://dollar.biz.uiowa.edu/~fil/IS/Framework/`