The iCrawl System for Focused and Integrated Web Archive Crawling

Gerhard Gossen, Elena Demidova, Thomas Risse
L3S Research Center
Hannover, Germany
{gossen, demidova, risse}@l3s.de

ABSTRACT

The large size of the Web makes it infeasible for many institutions to collect, store and process archives of the entire Web. Instead, many institutions focus on creating archives of specific subsets of the Web. These subsets may be based around specific topics or events. Our iCrawl system provides a *focused crawler* that is able to automatically collect Web pages relevant to a topic based on content similarity. Recently, the archiving of Social Media platforms like Twitter has become relevant. Our system can conduct *integrated crawls* that collect Web pages and Social Media posts concurrently. During such a crawl newly discovered URLs are exchanged between the crawling subsystems. We built the system with the goal to enable domain experts to create archives for their topics of interest. Therefore the system is highly automated and provides support for specifying and conducting crawls. We will demonstrate an easy to use interface for crawl specification that allows users to find seed URLs as well as descriptive keywords using Web and Social Media search APIs. The iCrawl system is available as Open Source software.

PROBLEM

The large size of the Web makes it infeasible for many institutions to collect, store and process archives of the entire Web. Therefore many institutions have abandoned "collect all"-approaches [3] and target focused collections. These may be based around organizational responsibilities (e.g., to collect all Web documents for a given country) or around relevant events or topics. Typically these collections are created based on manual selections of relevant Web hosts. It is however also possible to conduct *focused Web crawls* that collect relevant Web pages using the similarity of the page content to a given *crawl specification*. We implemented this approach because it gives us a more user-friendly way to create the archives.

In recent years the relevance of Social Media platforms has been recognized [2]. These platforms give us a way to discover relevant and up-to-date content and URLs about a given topic [2]. The ephemeral nature especially of tweets makes a deep integration of Web and Social Media crawling necessary, as we otherwise increase the chance of missing relevant content. We implemented an integrated crawler that runs Web and Social Media crawls in parallel. URLs discovered in this process are shared between the subsystems so that Web URLs can be crawled shortly after they are posted on Twitter [2].

Our system is designed to make it easy for domain experts to specify and conduct crawls. In particular, we have developed and will demonstrate an easy to use interface for interactive crawl specifications [1]. The user can use search terms to find relevant seed URLs through Web and Twitter search APIs. Search result Web pages are analyzed using Natural Language Processing (NLP) methods to detect keywords that on the one hand help the user better understand the content of the pages, but on the other hand can also be used as descriptive keywords to help focus the crawl.

The iCrawl system is available as Open Source Software[1].

ACKNOWLEDGMENTS

This work was partially funded by the ERC under ALEXANDRIA (ERC 339233) and BMBF (Project "GlycoRec").

REFERENCES

[1] G. Gossen, E. Demidova, and T. Risse. The iCrawl Wizard – supporting interactive focused crawl specification. In *Proceedings of the European Conference on Information Retrieval (ECIR) 2015*, 2015.

[2] G. Gossen, E. Demidova, and T. Risse. iCrawl: Improving the freshness of web collections by integrating social web and focused web crawling. In *Proceedings of the Joint Conference on Digital Libraries 2015, JCDL 2015*. ACM, June 2015.

[3] T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 426–432. Springer, Berlin / Heidelberg, 2012. ISBN 978-3-642-33289-0. DOI:10.1007/978-3-642-33290-6 47.

[1] http://icrawl.l3s.uni-hannover.de/