Siphoning Hidden-Web Data through Keyword-Based Interfaces: Retrospective

Luciano Barbosa[1], Juliana Freire[2]

[1] AT&T Labs Research
lbarbosa@research.att.com
[2] School of Computing, University of Utah
juliana@cs.utah.edu

Categories and Subject Descriptors: Information Systems [**Miscellaneous**]: Databases

Keywords: Online Databases, Hidden-Web Crawler

In this paper, we proposed the first, fully-automatic approach to crawling the Hidden Web through keyword-based interfaces. Our crawler uses an algorithm for automatically deriving a series of keyword-based queries whose goal is to obtain high coverage while minimizing the costs. In other words, our goal is to retrieve as much of the hidden contents as possible while minimizing the number of required queries. The intuition behind our algorithm is that, by obtaining samples of the hidden contents in a online database or document collection, we are able to discover keywords that have high frequency. Then, by using these high-frequency keywords we are able to construct queries that return a large number of answers.

Since our paper was published in the Proceedings of the Brazilian Database Sypomsium in 2004, it has been cited sixty six times[1]. Other hidden-Web crawlers were later proposed which make use of our algorithm [Madhavan et al. 2008] and that present different crawling strategies [Ntoulas et al. 2005]. Ntoulas et al. [Ntoulas et al. 2005] proposed a crawling strategy that selects a query which is most likely to retrieve the largest number of new pages according to an estimator. This work differs from our approach in many ways. First, whereas our approach initially builds a sample of the database before issuing crawling queries, Ntoulas et al. start the crawling process without any knowledge about the database. As a result, several iterations (submissions) may be required before effective queries are derived. Second, whereas the initial keywords issued in their method are set manually, our approach automatically selects them. Third, their strategy does not take advantage of keyword interfaces that accept disjunctive queries, which considerably reduces the number of queries used by the crawler. Madhavan et al. [Madhavan et al. 2008] present a system for crawling hidden-Web content to be incorporated into Google search engine index. Although their approach is targeted to both structured forms and keyword-based interfaces, they apply our algorithm for generating queries for the latter as well as to derive inputs for open-ended elements in structured forms. More recently, we have extended our algorithm to adapt the query generation and selection to the characteristics of the underlying index, and in a preliminary experimental evaluation, we have shown that the adaptive strategy obtains higher coverage than the original, fixed strategy [Vieira et al. 2008].

[1]Citation count provided by Google Scholar on May 17, 2010

One of the main contributions of this paper was to show that it is feasible to automatically crawl online databases through Web forms. To perform this task, however, it is necessary, first, to locate the online databases on the Web. This motivated a series of subsequent papers whose focus was on discovering and organizing hidden-Web sites [Barbosa 2009], including: focused crawlers specialized for finding Web interfaces [Barbosa and Freire 2005; 2007a; Barbosa et al. 2007], a classification method to identify relevant Web forms for a particular domain [Barbosa and Freire 2007b], and a clustering algorithm that groups together forms beloging to the same database domain [Barbosa et al. 2007].

Our hidden-Web crawler is currently being used in a production setting to retrieve large volumes of hidden content. Recently, we have used it to generate a query for retrieving the contents from the Pubmed database (http://www.ncbi.nlm.nih.gov/pubmed). While our sampling method produced a coverage estimate of 90%, the actual crawl downloaded more than 90% of the actual Pubmed database (roughly 17.2 million pages). This indicates that our approach can also be used to produce reliable estimates of query coverage.

REFERENCES

Barbosa, L. *Uncovering the Hidden Web*. Ph.D. thesis, University of Utah, 2009.

Barbosa, L. and Freire, J. Searching for hidden-web databases. In *Proceedings of the Workshop on Web and Databases*. Baltimore, USA, pp. 1–6, 2005.

Barbosa, L. and Freire, J. An adaptive crawler for locating hidden-web entry points. In *Proceedings of the International World Wide Web Conferences*. Banff, Canada, pp. 441–450, 2007a.

Barbosa, L. and Freire, J. Combining classifiers to identify online databases. In *Proceedings of the International World Wide Web Conferences*. Banff, Canada, pp. 431–440, 2007b.

Barbosa, L., Freire, J., and Silva, A. Organizing hidden-web databases by clustering visible web documents. In *Proceedings of the International Conference on Data Engineering*. Istanbul, Turkey, pp. 621–633, 2007.

Barbosa, L., Tandon, S., and Freire, J. Automatically constructing a directory of molecular biology databases. In S. Istrail, P. Pevzner, and M.Waterman (Eds.), *Data Integration in the Life Sciences*. Lecture Notes in Computer Science, vol. 4544. Springer, pp. 6–16, 2007.

Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A., and Halevy, A. Google's Deep Web crawl. *Proceedings of the VLDB Endowment archive* 1 (2): 1241–1252, 2008.

Ntoulas, A., Zerfos, P., and Cho, J. Downloading textual hidden web content through keyword queries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries*. Denver, USA, pp. 100–109, 2005.

Vieira, K., Barbosa, L., Freire, J., and Silva, A. Siphon++: a hidden-webcrawler for keyword-based interfaces. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, pp. 1361–1362, 2008.