Structured Data on the Web

Alon Y. Halevy

Google Inc.,
1600 Amphitheatre Parkway,
Mountain View, California, 94043,
USA
halevy@google.com

Abstract of Plenary Talk

Though search on the World-Wide Web has focused mostly on unstructured text, there is an increasing amount of structured data on the Web and growing interest in harnessing such data. I will describe several current projects at Google whose overall goal is to leverage structured data and better expose it to our users.

The first project is on crawling the *deep web*. The deep web refers to content that resides in databases behind forms, but is unreachable by search engines because there are no links to these pages. I will describe a system that *surfaces* pages from the deep web by guessing queries to submit to these forms, and entering the results into the Google index [1]. The pages that we generated using this system come from millions of forms, hundreds of domains and over 40 languages. Pages from the deep web are served in the top-10 results on google.com for over 1000 queries per second.

The second project considers the collection of HTML tables on the web. The WebTables Project [2] built a corpus of over 150 million tables from HTML tables on the Web. The WebTables System addresses the challenges of extracting these tables from the Web, and offers search over this collection of tables. The project also illustrates the potential of leveraging the collection of schemas of these tables.

Finally, I'll discuss current work on computing aspects of queries in order to better organize search results for exploratory queries.

Keywords: Deep web, structured data, heterogeneous databases, data integration.

References

1. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep-web crawl. In: Proc. of VLDB, pp. 1241–1252 (2008)
2. Cafarella, M.J., Halevy, A., Zhang, Y., Wang, D.Z., Wu, E.: WebTables: Exploring the Power of Tables on the Web. In: VLDB (2008)

Y.A. Feldman, D. Kraft, and T. Kuflik (Eds.): NGITS 2009, LNCS 5831, p. 2, 2009.