Search Joins with the Web

Christian Bizer
Data and Web Science Research Group
University of Mannheim
B6, 26, D-68131 Mannheim, Germany
chris@informatik.uni-mannheim.de

ABSTRACT
The lecture discusses the concept of Search Joins. A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web in various formats. A Search Join takes as input a local table, a corpus of heterogeneous Web tables, and a description of the attributes that should be added to the local table. The challenges that Search Joins need to handle are threefold: 1. Determine the set of the top-k Web tables which are beneficial candidates for the join operation; 2. Join the local table with the top-k candidate tables given no external knowledge about key attributes; 3. Merge corresponding attributes and fuse attribute values in order to return a concise result table containing high-quality data.

Search Joins are useful in various application scenarios. They allow for example a local table about cities to be extended with an attribute containing the average temperature of each city for manual inspection [5]. They also allow tables to be extended with large sets of additional attributes as a basis for data mining, for instance to identify factors that might explain why the inhabitants of one city claim to be happier than the inhabitants of another [7].

Existing work on extending local tables with additional attributes from the Web mainly focused on corpora of HTML tables extracted from Web crawls [3][4][8][9]. The recent increase in the adoption of Linked Data publishing [2], Microdata and RDFa annotations [1] as well as the growth of public data repositories such as datahub.io and data.gov.uk make a wide range of larger tables available on the Web and enable Search Joins to exploit these more comprehensive data sets.

In the lecture, I will draw a theoretical framework for Search Joins and will survey the state of the art methods employed by Search Join systems to handle the challenges outlined above. Afterward, I will highlight how the recent developments in the context of Linked Data, RDFa and Microdata publishing, public data repositories, as well as crowd-sourcing integration knowledge [2][5][6] contribute to the feasibility of Search Joins in an increasing number of topical domains.

1. REFERENCES

(c) 2014, Copyright is with the authors. Published in Proc. 17th International Conference on Database Theory (ICDT), March 24-28, 2014, Athens, Greece: ISBN 978-3-89318066-1, on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.