Best-Effort Modeling of Structured Data on the Web

Alon Halevy

Google Research,
1600 Amphitheatre Parkway,
Mountain View, California, USA

The World-Wide Web provides access to millions of data tables with high-quality content, formatted either in HTML tables, HTML lists, or other structured formats, or stored in on-line data management services. These tables contain data about virtually every domain of interest to mankind. Several reasearch projects aim at enabling search over these data sets and ultimately the ability to answer queries and to combine data from multiple sources.

In addition to the challenges involved in extracting the high-quality data sets from the Web, there is a fundamental challenge concerning how and whether to create a conceptual model of the data that can be used by the higher-level services. Creating a conceptual model, in the traditional sense, for such a collection of data is impractical because of (1) the breadth of the data, (2) the fact that domains overlap in complex ways, and (3) that modeling assumptions differ depending on the level of detail and cultural context.

Several projects at Google have the goal of leveraging this collection of data and to make it easier to create and share new data sets. In each case, interesting challenges arise from the lack of a conceptual model. In the WebTables Project [1,3] we collected over 100 million high-quality HTML tables, developed search over this collection. We used information from text on the Web to recover some of the semantics of these tables. In Google Fusion Tables [2], we make it easy for data owners to upload and manipulate their data, create visualizations and discover other data sets that may be relevant to them, all this without requiring them to a priori create a model of their data. The experiences from these projects suggest that we may require a fundamentally different approach to data modeling in the context of the Web.

References

1. Cafarella, M.J., Halevy, A.Y., Wang, D.Z., Wu, E., Zhang, Y.: WebTables: Exploring the Power of Tables on the Web. PVLDB 1(1), 538–549 (2008)
2. Gonzalez, H., Halevy, A., Jensen, C., Langen, A., Madhavan, J., Shapley, R., Shen, W., Goldberg-Kidon, J.: Google Fusion Tables: Web-Centered Data Management and Collaboration. In: SIGMOD (2010)
3. Venetis, P., Halevy, A., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. In: PVLDB (2011)

M. Jeusfeld, L. Delcambre, and T.W. Ling (Eds.): ER 2011, LNCS 6998, p. 32, 2011.