Self-Extending Peer Data Management

Ralf Heese        Sven Herschel        Felix Naumann        Armin Roth
{rheese|herschel}@informatik.
                hu-berlin.de
Humboldt-Universität zu Berlin
Databases and Information Systems
Berlin, Germany

{naumann|aroth}@informatik.
                hu-berlin.de
Humboldt-Universität zu Berlin
Information Integration Group
Berlin, Germany

Abstract: Peer data management systems (PDMS) are the natural extension of integrated information systems. Conventionally, a single integrating system manages an integrated schema, distributes queries to appropriate sources, and integrates incoming data to a common result. In contrast, a PDMS consists of a set of peers, each of which can play the role of an integrating component. A peer knows about its neighboring peers by mappings, which help to translate queries and transform data. Queries submitted to one peer are answered by data residing at that peer and by data that is reached along paths of mappings through the network of peers.

The only restriction for PDMS to cover unbounded data is the need to formulate at least one mapping from some known peer to a new data source. We propose a Semantic Web based method that overcomes this restriction, albeit at a price. As sources are dynamically and automatically included in a PDMS, three factors diminish quality: The new source itself might store data of poor quality, the mapping to the PDMS might be incorrect, and the mapping to the PDMS might be incomplete. To compensate, we propose a quality model to measure this effect, a cost model to restrict query planning to the best paths through the PDMS, and techniques to answer queries in such Web-scale PDMS efficiently.