Provenance for Database Transformations

Val Tannen

University of Pennsylvania, USA

Database transformations (queries, views, mappings) and the languages in which they are expressed are of obvious interest in information management. They take apart, filter and recombine source data in order to populate warehouses, views, and analysis tool inputs. As they do so, we need to track the relationship between parts and pieces of the sources and parts and pieces of the transformations' output. This relationship is what we call database provenance.

This talk will present an approach to database provenance that relies on three observations. First, provenance definitions follow the constructs of the language in which queries/views/mappings are expressed. Second, provenance is a kind of annotation, and there exist approaches to annotated data that we can relate to. In fact, it can be argued that provenance is the most general kind of annotation, when properly viewed. Third, the propagation of annotation through most language constructs seems to rely on just two annotation operations: one when annotations are jointly used and one when they are used alternatively. We will see that this leads to annotations forming a specific algebraic structure, a commutative semiring.

The semiring approach works for annotations on standard relations, but also on nested relations (complex values), and unordered XML. It works for the positive fragment of relational algebra, nested relational calculus, unordered XQuery, and even for languages with recursion (Datalog). It turns out that specific semirings correspond to the approaches to provenance presented in previous work. Other semirings yield applications to incomplete/probabilistic data, and to access control in databases.

This is joint work with J. N. Foster, T. J. Green, and G. Karvounarakis.

J. Freire, D. Koop, and L. Moreau (Eds.): IPAW 2008, LNCS 5272, p. 1, 2008.