CIDR 2013 Paper
Data Integration and Data Exchange: It’s Really About Time, CIDR 2013
Authors: Mary Roth, Wang-Chiew Tan

Summary

The authors proposed a framework for data integration across time. This has a significant application in integrating electronic medical records in public health and integrating information from public reports. In the paper, the authors primarily used information from Securities and Exchange Commission about executives (information from forms 10Q and etc.).

How does it build upon, extend, or differ from other work
- There has been many research works on data integration in the area of temporal databases, but there is no general approach for combining heterogeneous data sources with temporal dimension.
- Previous work in mapping one schema to another is very application-specific. Also there need to be some data example training and/or ad-hoc manipulation function in the mapping process.

Key ideas in the paper
- Data integration is necessary, since each piece of information does not contain all the information we need.
- The main example in the paper is how we track a title (position in a company) when we have different pieces of information about their job at different points of time.
- There are more than 1 time dimension in a real world problem.
- Bi-temporal databases: For data integration, there are at least two different time dimensions: valid time and report time. The challenge was that if facts/updates arrive in different order, the result would be different if we track only the report time.
- Streaming data: contains valid time, occurrence-time and transaction-time.
- New model: time-aware data model where every element in an object record has a temporal context of n-dimensions where n is the number of attributes for that object.
- There is a need for some time-aware schema mapping: high-level language (predicate/constraints)
- A need for time-aware union: resolve conflict automatically at any time by fusing records but maintaining associativity and commutativity when combining them.

Positive/Strong Points
- S1: The authors pointed out deficiencies in current data integration techniques: versioning, archiving, annotation system and etc.
- S2: The use of linear temporal logic and modal operators can help generalize the constraints in data integration (although it is a question how we enforce/solve the predicate).
- S3: The fictitious example in the paper was simple and illustrative. It also demonstrate the use of how a time-aware data model would look like for a simple case.

Negative/Weak Points
- W1: For their method when we track each record/transaction, the authors assume that all sources of data provide data in a relational form; which is very unrealistic; converting unstructured data to structured requires a lot of additional processing and very prone to error.
- W2: The authors did not indicate whether an efficient time-aware union or mapping can really exist.
- W3: The complexity of a data record in the time-aware model is very high (n-dimension). Even though the record object might be sparse in the case of tracking a job title of a person, in other domains, objects can have many attributes and tracking temporal context of each dimension in each object appears to be computationally intractable and incur a lot of storage space.

Research Questions and Points for Discussion
- D1: How can we relax the assumption that each entity can only assume one position at a time?
- D2: Query in data integration/exchange is a very interesting topic. How do we determine whether materializing the target schema or converting the query to the target framework is more efficient?
- D3: What is the measure of correctness in data integration if all the time-aware unions, mapping are satisfied? Most of the time we do not have a training dataset and we cannot generalize any method across different domains.
- D4: How can we generate new time dimensions from raw data? What are all temporal features can we extract from them?