Universität Rostock
Fakultät für Informatik und Elektrotechnik
Institut für Informatik

Provenance Management unter Verwendung von Schemaabbildungen mit Annotationen

Dissertation
von
Tanja Auge
zur
Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

Abgabedatum: 28.11.2022

Zusammenfassung

Die Verfolgung, Organisation und Archivierung von Daten, welche im Rahmen wissenschaftlicher Projekte, Experimente oder Beobachtungen gesammelt werden, ist Aufgabe des Forschungsdatenmanagements. Je nach Anwendungsfall unterscheiden sich dabei die Maßnahmen und Verfahren, welche hierfür zum Einsatz kommen. In allen Fällen soll jedoch der Weg von der Datenerhebung bis zur Veröffentlichung replizierbar, rekonstruierbar und plausibel gehalten werden. Das kontinuierliche Wachstum der Daten, häufige Schemaänderungen sowie die Vielfältigkeit der Daten und ihrer Auswertungsmethoden machen die Speicherung aller möglichen Datenbankzustände dabei zu einer sehr aufwendigen und langwierigen Aufgabe.

Mit Hilfe von Provenance lässt sich jedoch feststellen, welcher Teil der originalen Daten — meinst primäre oder sekundäre Forschungsdaten — langfristig gespeichert werden muss, um die Reproduzierbarkeit, Replizierbarkeit oder Plausibilität einer konkreten Auswertung zu gewährleisten. Auch sollen im Falle temporaler Datenbanken Änderungen am Schema berücksichtigt werden können, um alte Datenbestände nicht komplett archivieren zu müssen, sondern aus bekannten Beständen zurück berechnet werden können. Neben der Auswertungsanfrage sowie dem Anfrageergebnis benötigen wir somit zusätzliche Annotationen, um ein Anfrageergebnis auf eines der oben genannten Kriterien zu überprüfen. Außerdem dürfen die gespeicherten Daten nicht mit bestehenden Datenschutzrichtlinien kollidieren.

Durch die Kombination vom Chase — einer Familie von Algorithmen zur Transformation von Datenbanken — mit Data Provenance und zusätzlichen Annotationen können für eine gegebene Auswertungsanfrage eine anonymisierte (minimale) Teil-Datenbank einer originalen (Forschungs-)Datenbank berechnet werden. Um die Reproduzierbarkeit, Replizierbarkeit oder Plausibilität eines Anfrageergebnisses zu gewährleisten, müssen die auf der originalen Datenbank durchgeführten Auswertungen auch auf der rekonstuierten (minimalen) Teil-Datenbank durchführbar sein. Hierfür nutzen wir eine Version des Chase&Backchase, einer Erweiterung des Chase. Ziel des Promotionsprojekts *ProSA* (**Pro**venance Management using **S**chema Mappings with **A**nnotations) ist somit die Anwendung und Verallgemeinerung von Techniken des Provenance Managements im Bereich des Forschungsdatenmanagements unter Verwendung des mit zusätzlichen Provenance-Informationen erweiterten Chase&Backchase.

Abstract

Tracking, organizing, and archiving data collected in the course of scientific projects, experiments, or observations is the task of research data management. The measures and procedures used for this vary depending on the application. In all cases, however, the path from data collection to publication should be kept replicable, reconstructable and plausible. The continuous growth of the data, frequent schema changes, and the diversity of the data itself and its evaluation methods make the storage of all possible database states a very time-consuming and tedious task.

However, using provenance, it is possible to determine which part of the original data — primary or secondary research data, at least — need be stored in the long term in order to ensure the repeatability, replicability or plausibility of a concrete evaluation. Also, in the case of temporal databases, changes to the schema should be able to be taken into account in order not to have to archive old data sets completely. Thus, in addition to the evaluation query and its result, we need additional annotations to check a query result against one of the above criteria. Furthermore, the stored data must not collide with existing data protection policies.

By combining Chase — a family of algorithms for transforming databases — with data provenance as well as additional annotations, an anonymized (minimal) partial database of an original (research) database can be computed for a given evaluation query. To ensure the repeatability, replicability, or plausibility of a query result, the evaluations performed on the original database must also be feasible on the reconstructed minimal partial database. For this purpose, we use a version of Chase&Backchase, an extension of Chase. Thus, the goal of the PhD project *ProSA* (**Pro**venance Management using **S**chema Mappings with **A**nnotations) is to apply and generalize provenance management techniques in the field of research data management using Chase&Backchase enhanced with additional provenance.