QuanTD

Quantifying Trustworthiness of Data

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft - 10. Ausschreibung (2021)	**Status**	laufend
Projektstart	01.12.2022	**Projektende**	31.05.2026
Zeitraum	2022 - 2026	**Projektlaufzeit**	42 Monate
Keywords	trustworthiness, quality metric, uncertainty, quantifying		

Projektbeschreibung

Fehlende, falsche und inkonsistente Daten verursachen viele Probleme und hohe Kosten. Datenqualität (DQ) ist also eine wichtige Voraussetzung für datenbasierte Anwendungen. Zum Messen von DQ existieren viele verschiedene Metriken für unterschiedliche DQ Dimensionen, z.B. Vollständigkeit oder Korrektheit. Da die globale Datenqualität eines Informationssystems auf Basis einzelner Metriken in der Praxis oft schwer zu erfassen ist, streben Organisationen nach aggregierten DQ-Scores, welche die Ergebnisse mehrere Metriken zu einzelnen Ergebniswerten kombinieren.

Die Verdichtung auf einen einzelnen Ergebniswert erhöht den Bedarf nach der Bewertung der Vertrauenswürdigkeit solcher DQ-Scores. In einigen Fällen können sie genau den „realen“ Wert der Datenqualität widerspiegeln, in anderen kann große Unsicherheit involviert sein, z.B. durch eine große Anzahl von Standardwerten. Daher ist es unser Ziel, die Vertrauenswürdigkeit von Daten und deren DQ-Scores unter Berücksichtigung der Unsicherheit messbar zu machen. Zu diesem Zweck schlagen wir vor, die Entwicklung einzelner DQ-Scores zu untersuchen, indem Werte aus geeigneten DQ Metriken für verschiedene DQ Dimensionen kombiniert werden, wobei zusätzliche Informationen über ihre Unsicherheit berücksichtigt und in eine neue Darstellung der Unsicherheit solcher kombinierter einzelner DQ-Scores propagiert werden.

Die Unsicherheit wird basierend auf Wahrscheinlichkeiten dargestellt und gemäß der Wahrscheinlichkeitstheorie propagiert. Unter anderem schlagen wir vor, Intervalle zu untersuchen, um die Unsicherheit eines einzelnen DQ-Scores darzustellen, z.B. unter Verwendung von Konfidenzintervallen. Zusätzlich schlagen wir einen Ansatz mittels maschinellem Lernen vor, um mehrere Werte von verschiedenen Metriken automatisch zu einzelnen DQ-Scores zu kombinieren.

Zur Erklärung der Datenqualitätsanalysen soll ein einzelner DQ-Score mittels einer Visualisierungskomponente in die aggregierten Metriken mit den zugehörigen Unsicherheiten zerlegt werden können. Das ermöglicht es den Benutzer:innen eine mehrdimensionale Analyse der Datenqualität vorzunehmen und im Rahmen von Data Governance Maßnahmen eine Verbesserung der Datenqualität zu erreichen.

Wir werden unsere neuen Ansätze in zwei verschiedenen Organisationen bewerten, der Robert Bosch AG und der

Österreichischen Post AG. Ziel ist es, kombinierte DQ-Scores − mit Unsicherheiten − in verschiedenen Anwendungsfällen zu validieren, z.B. für Stammdatenmanagement und für Prozess- sowie Projektdaten.

Dadurch wird es für Organisationen in der Praxis und in der Gesellschaft einfacher sein, die Qualität von strukturierten Daten anhand einzelner DQ-Scores und ihre Vertrauens¬würdigkeit durch zusätzliche Informationen zur Unsicherheit zu analysieren und zu verstehen. Dies erleichtert die automatische Verwendung der Informationen über DQ für Berichte, Vorhersagen und Verbesserungen, sowie für den externen Austausch von Daten. Der Aufbau solchen Know-hows in Österreich leistet auch einen wichtigen Beitrag zur nationalen und europäischen Technologiesouveränität.

Abstract

Missing, incorrect, and inconsistent data cause a lot of problems and costs. Hence, data quality (DQ) is an essential prerequisite for data-based applications. For measuring DQ, many different DQ metrics exist for various DQ dimensions, e.g., completeness or correctness. Since the overall data quality of an information system is often hard to perceive in practice based on the single DQ metrics, organizations strive for aggregating DQ scores that combine the results of several metrics into single numbers.

The aggregation to a single number increases the need to evaluate the trustworthiness of such DQ scores. In some cases, they may accurately reflect the ‘real’ value of data quality, in others, a lot of uncertainty may be involved, e.g., due to a large number of standard values. Hence, we aim at making the trustworthiness of data and their DQ scores measurable by taking into account uncertainty. To this end, we propose to study the development of single DQ scores by combining values from specific DQ metrics for different DQ dimensions, where additional information about their uncertainty is taken into account and propagated to a new representation of the uncertainty of such combined single DQ scores.

The uncertainty will be represented based on probabilities and propagated according to probability theory. Among several approaches, we propose to study intervals for representing the uncertainty of a single DQ score, e.g., using confidence intervals. Additionally, we propose a machine learning approach to automatically combine multiple values from different metrics into single DQ scores.

To explain the data quality analysis, a single DQ score should be decomposable into the values from the aggregated metrics with their associated uncertainties, using a visualization component. This enables users to analyze multidimensional data quality and to improve the quality of this data in the course of data governance processes.

We will evaluate our new approaches in two different organizations, Robert Bosch AG and Österreichische Post AG. The aim is to validate combined DQ scores − with their uncertainties assigned − in different use cases, e.g., for master data management and for process and project data.

In effect, it will be easier for organizations in practice and the society in general to analyze and understand the quality of structured data through single DQ scores, and their trustworthiness through additional information on uncertainty. This will facilitate automated use of the information on DQ for reporting, predictions, and data improvements, as well as external data exchange. Creating such know-how in Austria makes an important contribution to national and European technology sovereignty.

Projektkoordinator

- Technische Universität Wien

Projektpartner

- Österreichische Post Aktiengesellschaft
- Software Competence Center Hagenberg GmbH
- Wirtschaftsuniversität Wien
- Robert Bosch Aktiengesellschaft