Konzepte zur Datenqualitätssicherung in analytischen Anwendungen

Mathias Körbs Eike Schallehn
Fakultät für Informatik
Universität Magdeburg, Postfach 4120, 39016 Magdeburg
mathias@koerbs.de,eike@iti.cs.uni-magdeburg.de

Zusammenfassung

Mit der zunehmenden Verwendung von Datenbanksystemen und stetig wachsenden Datenvolumen wird die Frage nach der Qualität der verwalteten Daten immer bedeutender. Entsprechend wurden die Fragen, was genau Datenqualität ist und wie man sie messen und bewerten kann, in den letzten Jahren zu einem aktuellen Thema in der Forschung und in der industriellen Praxis. Inbesondere bei analytischen Anwendungen ist eine Bewertung der Qualität der zu Grunde liegenden Daten wichtig, um eine Aussage über die Güte der abgeleiteten Analyseergebnisse machen zu können. Dies wird in Anwendungen, die auf einem Data Warehouse oder ähnlichen Ansätzen basieren, zunehmend problematisch, da Informationen zur Qualität integrierter Datenbestände nur schwer ableitbar sind.

Während grundlegende Ansätze zu spezifischen Problemen der Bewertung und Messung von Datenqualität existieren, ist die Integration entsprechender Funktionalität in Informationssystemen immer noch ein weitgehend ungelöstes Problem. Diese Aufgabenstellung wird in unserer aktuellen Forschung in Kooperation mit Industriepartnern angegangen.

1 Einleitung

Mit zunehmenden Einsatz von datenintensiven Anwendungen gewinnt auch der Begriff der Datenqualität an Bedeutung. Die Qualität im Allgemeinen ist ein Schlüsselfaktor bei der Herstellung beliebiger Produkte und dem Angebot von Dienstleistungen. Handelsketten, Automobilkonzerne und viele andere Unternehmen verlangen von ihren Zulieferern ein klar definiertes Mindestmaß an Qualität. Dabei beziehen sich Qualitätsanforderungen nicht nur auf die Produkte an sich, sondern auch auf alle Prozesse im Unternehmen. Vieles muss dokumentiert werden, wird nachgemessen und kontrolliert. Über die Qualität von Daten ist hingegen oftmals wenig bekannt. Häufig beschränken sich Aussagen über die Datenqualität auf Abschätzungen. Dennoch ist der Umgang mit Daten unter Qualitätsaspekten von Bedeutung, gerade dann, wenn auf Basis der Daten wichtige Entscheidungen getroffen werden sollen.

Die Datenqualität ist ein besonders kritischer Aspekt bei der Datenintegration, wo Daten aus verschiedenen autonomen Quellen zusammengeführt werden, und es häufig zu schwer abschätzbaren Fehlern und Inkonsistenzen kommt. Bei Data Warehouse-Systemen handelt es sich um einen speziellen und sehr erfolgreichen Ansatz zur Datenintegration. Dazu werden aus verschiedenen Systemen eines Unternehmens Daten extrahiert und materialisiert, um diese als Grundlage für Analysen in einem übergreifenden Kontext zu verwenden. Jedoch hat eine schlechte Qualität der integrierten Ausgangsdaten oft auch erheblichen Einfluß auf die Qualität der Analysergebnisse.

Für die Sicherstellung der Datenqualität haben Forschung und Praxis zu zahlreichen Lösungen für spezielle Probleme geführt. Jedoch ist Datenqualität jeweils nur durch anwendungsspezifische Kriterien und Verfahren zu erreichen. Ein weiteres Problem ist die einheitliche Er-

fassung und Verwaltung von Qualitätsdaten sowie die Interpretation entsprechend eines einheitlichen Modells.

Im Rahmen der hier dargestellten Forschungsarbeiten sollten am Beispiel eines Systems zur Durchführung von Analysen zur Fahrzeugsicherheit bei einem großen Automobilhersteller Konzepte zur Integration von Datenqualitätsfunktionalität untersucht werden. Dabei sollte einerseits eine enge Integration mit der bestehenden Anwendung erreicht werden, und andererseits soll die Lösung für zahlreiche relevante Datenqualitätsaspekte offen und erweiterbar sein.

2 Stand der Forschung

Angelehnt an die Qualittsdefiniton der ISO-Norm 9000:2000 [EN00] lässt sich die Qualität von Daten ähnlich wie bei natürlichen Produkten oder Dienstleistungen als eine Menge von Merkmalen dieser Daten ausdrücken. Solche Merkmale sind den Daten inhärente Eigenschaften, also unveränderlich mit ihnen verbunden. Bei einer Schraube ist die Tiefe ihres Gewindes ein Merkmal [Gie01], bei einer Menge von Daten ist dies zum Beispiel die Vollständigkeit. Ob eine konkrete Merkmalsausprägung hohe Qualität bedeutet hängt dabei von Anforderungen ab, die von interessierten Parteien, den Daten-Nutzern, Anbietern oder Produzenten, aufgestellt werden. Problematisch ist, dass in der Literatur viele Defintionen des Begriffes Datenqualität auf den Konzepten des Total Quality Management basieren, das im Gegensatz zur ISO 9000:2000 keine scharfe Definiton für Qualität kennt [GD03]. Die Folge ist eine uneinheitliche Terminologie. So werden die Begriffe Dimension, Attribut und Merkmal synonym verwendet. Lediglich in [Hin02] kommt eine Definition nach der ISO-Norm 9000:2000 zur Anwendung.

Um eine hohe Qualität von Daten zu erlangen ist es notwendig aktiv darauf Einfluss zu nehmen, das heißt, nicht auf Folgen mangelnder Datenqualität zu reagieren, sondern lenkend einzugreifen. Ein solcher Qualitätsmanagementprozess besteht im wesentlichen aus den vier Phasen

1. Anforderungen aufstellen,
2. Ist-Zustand ermitteln,
3. Ist-Zustand analysieren und
4. Verbesserungsmaßnahmen ableiten und durchführen.

Um eine ständige Qualitätskontrolle und -verbesserung zu erreichen werden die Phasen immer wieder iterativ ausgeführt. Selbst wenn ein gewünschtes Maß an Qualität erreicht ist, ist dieser Prozess zur Kontrolle notwendig. Auf Datenqualität angepasste Qualitätsmanagementprozesse werden unter anderen im Rahmen der Total Data Quality Management (TDQM) [Wan98] und Data Warehouse Quality (DWQ) [YV97, VBQ99] Projekte vorgestellt.

3 Konzepte zur Integration der Qualitätsicherung

Um die oben genannten Zielsetzungen des Qualitätsmanagements zu realisieren, sind im Folgenden vor allem Fragen der Integration des DQ-Managements in einem gegebenen Anwendungskontext sowie die flexible und erweiterbare Unterstützung verschiedener Metriken von Interesse.

3.1 Einbettung des DQ-Frameworks

Während sich für das Qualitätsmanagement Anforderungen und notwendige Verbesserungsmaßnahmen anwendungsspezifisch zumeist relativ einfach bestimmen lassen, ist die Ermittlung des

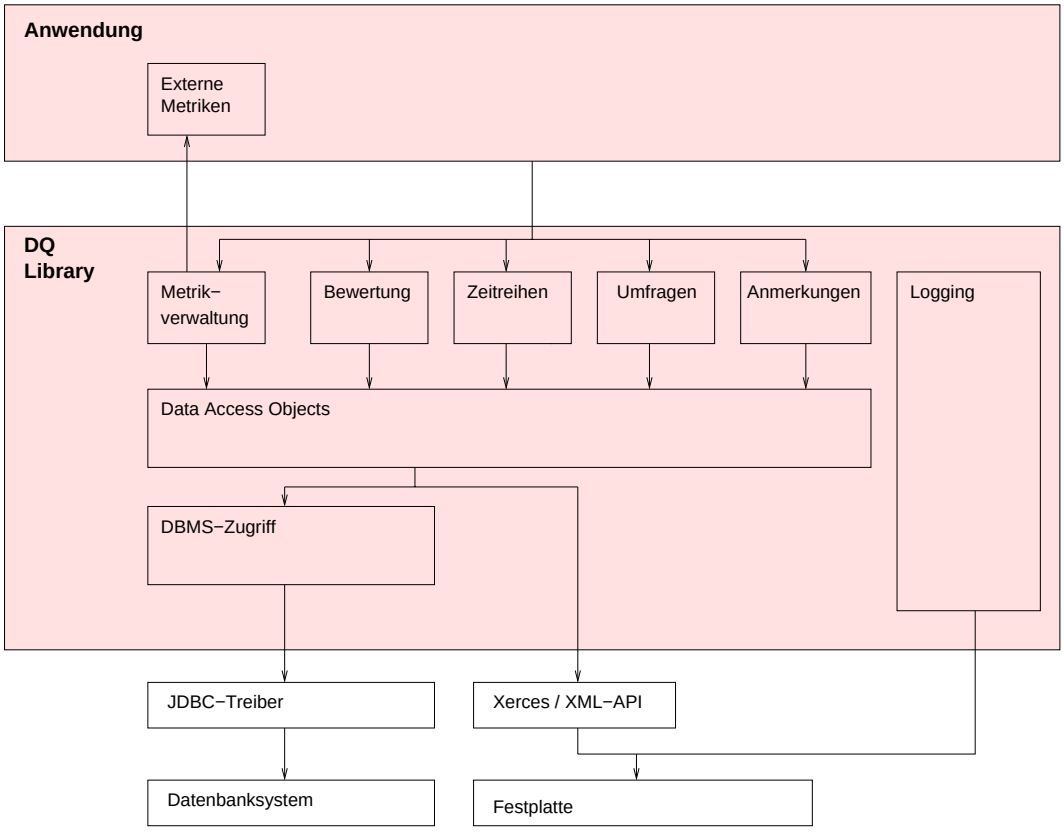

Abbildung 1: Architektur des DQ-Frameworks

Ist-Zustandes problematisch, weil Informationen über die Qualität aus den Daten gewonnen werden müssen.

Um die Qualität von Daten zu ermitteln, müssen die Ausprägungen relevanter Merkmale gemessen und die Messwerte daraufhin mit vorher aufgestellten Anforderungen verglichen und bewertet werden. Gerade die Messung ist schwierig, weil umfangreiches Wissen über die Daten notwendig ist, die Datenmenge wegen ihrer Größe häufig unhandlich ist, und weil viele Merkmale von der subjektiven Meinung eines Experten abhängen.

Um unter diesen Bedingungen dennoch Qualitätsaussagen zu erhalten, lassen sich Umfragen unter Angehörigen der interessierten Parteien durchführen oder Metriken speziell für einen bestimmten Anwendungsfall erstellen. Ein solches Vorgehen ist aber sehr teuer und daher, gerade durch den zyklischen QM-Prozess, häufig nicht durchführbar.

Dabei stellt sich die Frage, ob die Ermittlung der Qualität von Daten nicht mit einfachen und vor allem wiederverwendbaren Metriken ausreichend ist, gerade wenn das Verhältnis aus Kosten und Nutzen relevant ist und dadurch Ungenauigkeiten in Kauf genommen werden können. Ausgehend von dieser Problematik wurde ein Software-Framework entwickelt, um eine möglichst einfache Messung und Bewertung von Datenqualitätsmerkmalen zu ermöglichen. Die grundlegende Architektur des Frameworks ist in Abbildung 1 dargestellt.

3.2 Flexible Unterstützung von Metriken

Besondere Aspekte des Frameworks sind eine Metrikschnittstelle, die Normalisierung und Aggregation der ermittelten Messwerte um diese in einen Aggregationsbaum darstellen zu können, Zeitreihen um die Entwicklung der Merkmalsausprägungen über einen Zeitraum zu verfolgen, Umfragen um subjektive Merkmalsausprägungen zu ermitteln und die Möglichkeit der Markierung von Daten, damit den Daten-Nutzern ein Feedback über Fehler in den Daten gegeben

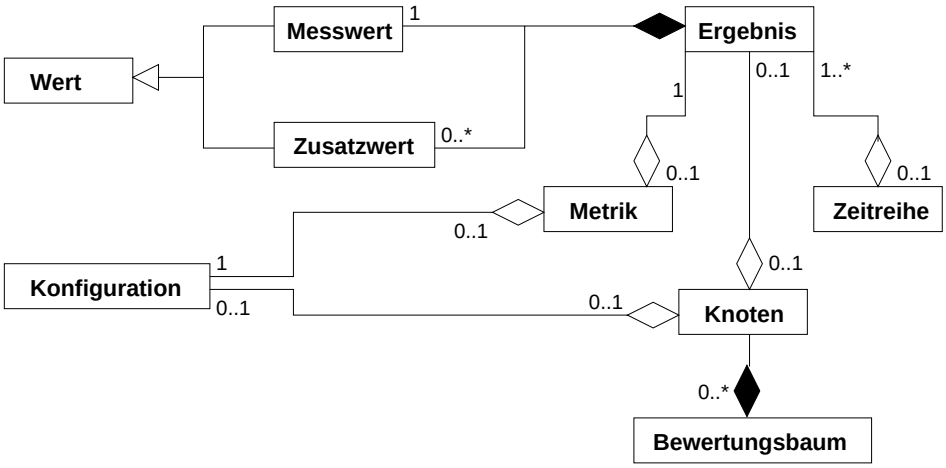

Abbildung 2: Konzeptionelles Schema des DQ-Frameworks

werden kann.

Im Einzelnen sind die betrachteten Daten enstprechend des in Abbildung 2 dargestellten Schemas strukturiert. Im Mittelpunkt stehen hierbei Metriken, durch deren Anwendungen eine konkrete Messung durchgeführt wird. Hierzu benötigt die Metrik eine Konfiguration, welche zum Beispiel festlegt, auf welche Daten (Relationen) die Metrik angewandt wird, und welche spezifischen Parameter gesetzt wurden. Das Ergebnis der Messung besteht nun aus einem einzelnen Messwert und gegebenenfalls weiteren Zusatzwerten, die zum Beispiel der Interpretation des Ergebnisses dienen können.

Zeitreihen werden zur kontinuierlichen und wiederholten Ausführung von Messungen verwaltet. Durch diese Sammlung von Messergebnissen und zugehörigen Zeitstempeln wird eine wichtige Grundlage für das Datenqualitätsmanagement gegeben, da auf diese Art und Weise die Bewertung durchgeführter Maßnahmen zur Verbesserung der Datenqualität möglich wird.

In einem Bewertungsbaum können die Ergebnisse verschiedener Messungen zusammengefasst werden und hierarchisch als Knoten dargestellt werden. Eine entsprechende Darstellung eines Bewertungsbaums in der Applikation zur Auswertung von Messeregebnissen ist in Abbildung 3 gegeben.

4 Zusammenfassung

Die hier skizzierten Ansätze wurden im Rahmen einer Diplomarbeit in der Forschungsabteilung eines Automobilherstellers implementiert und befinden sich momentan im Einsatz.

Durch eine Messung auf den Daten der dort eingesetzten analytischen Anwendung wurde ermittelt, inwieweit sich das Framework in diese Anwendung integrieren lässt und ob sich vorher benannte Daten quantifizieren lassen. Dabei wurden Messwerte für die Vollständigkeit auf Attribut und auf Tupelebene, der referentiellen Integrität und der Einhaltung von Konsistenzregeln ermittelt. Eine weitere Metrik wurde zur Messung eines in der Anwendung vorkommenden Spezialfalles der referentiellen Integrität verwendet.

Die Auswertung hat gezeigt, dass die Messwerte ausreichend genau waren, um das Ausmaß der vorher benannten Datenfehler zu bestimmen. Auch hat sich gezeigt, dass sich das Framework einfach in die Anwendung integrieren ließ.

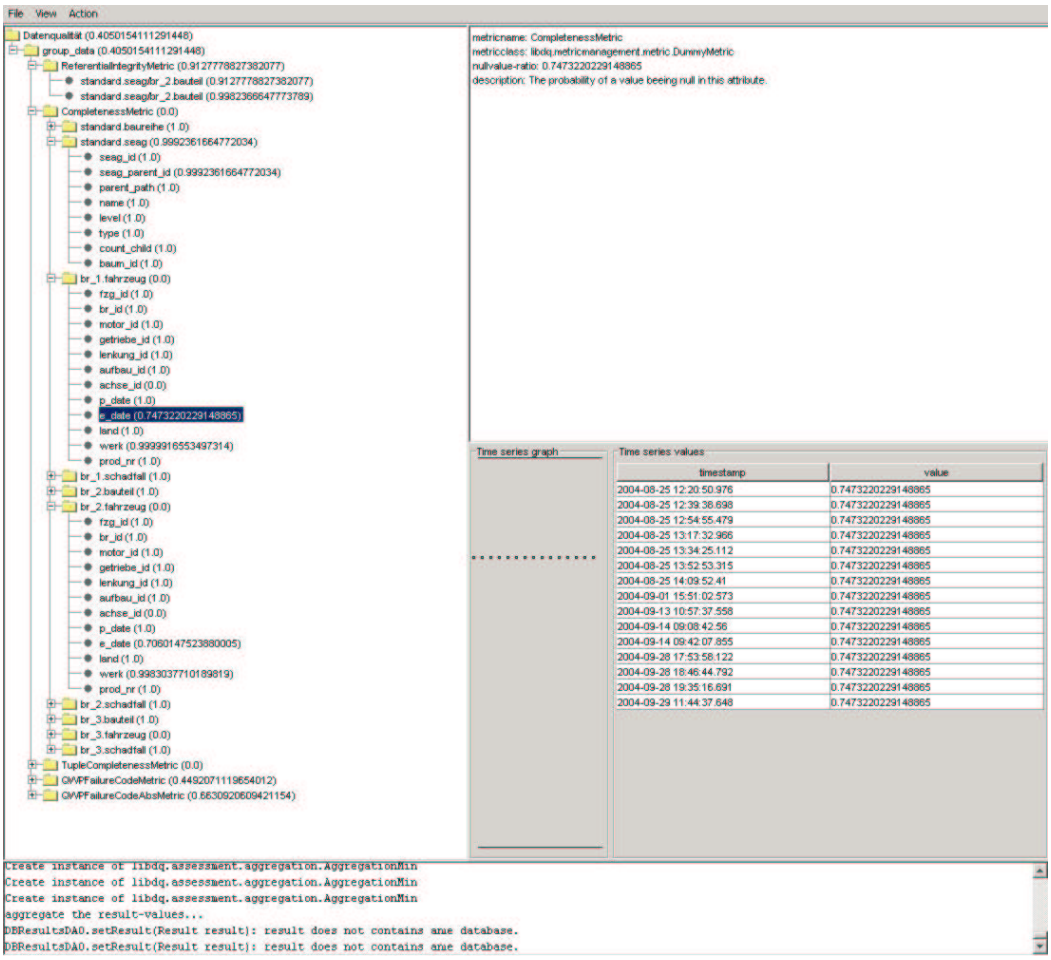

Abbildung 3: Applikation zur Auswertung und Kombination von Messergebnissen

Literatur

[EN00] DIN EN. ISO 9000 : 2000 Qualitätsmanagementsysteme Grundlagen und Begriffe., 2000.

[GD03] D.L. Goetsch and S. B. Davis. *Quality management: introduction to total quality management for production, processing and services.* Pearson Education, Inc., New Jersey, 2003.

[Gie01] D.G. Gietl. *Qualitätsmanagement: Begriffe und Definitionen.* Dr. Ingo Resch GmbH, 2001.

[Hin02] H. Hinrichs. *Datenqualitätsmanagement in Data-Warehouse-Systemen.* 2002.

[VBQ99] P. Vassiliadis, M. Bouzeghoub, and C. Quix. Towards quality-oriented data warehouse usage and evolution. pages 164–179, 1999.

[Wan98] R.Y. Wang. A product perspective on total data quality management. 41(2):58–65, 1998.

[YV97] M. Yarke and Y. Vassiliou. Data warehouse quality: A review of the dwq project. pages 299–313, 1997.