

Seminar Imperfektion und Datenbanken  
Wintersemester 2003/2004

# Ein Überblick über imperfekte Daten in Datenbanken

Oliver Schöll  
Betreuer: Heiko Schepperle

Universität Karlsruhe  
Fakultät für Informatik  
Institut für Programmstrukturen und Datenorganisation  
Lehrstuhl für Systeme der Informationsverwaltung

# Reale Welt - Datenbank

Eine Datenbank ist eine Sammlung von Daten aus einem bestimmten Bereich der realen Welt.

Ziel einer Datenbank ist es die reale Welt möglichst getreu abzubilden. Es werden also alle relevanten Daten aus dem Bereich erfasst und in die Datenbank eingebracht.

Ein großer Faktor für die Imperfektion ist bereits hier in der Entscheidung, welches die „alle relevanten Daten“ sind angelegt, jedoch wird dies hier nicht behandelt.

# Perfekte Daten

Was sind perfekte Daten?

Der Begriff der perfekten Daten meint die exakte Übereinstimmung der Abbildung der Daten in der Datenbank mit den tatsächlichen in der realen Welt.

Da sich die reale Welt jedoch ständig verändert und die Datenbank auf die Bedürfnisse von Nutzern ausgelegt ist, die sich ebenfalls ändern können, sind perfekte Daten eine Idealisierung des wirklichen Zustandes von Datenbanken.

# Imperfekte Daten

Nimmt man Abstand von der Idealisierung der perfekten Daten, so kommt man zum Begriff der imperfekten Daten, welcher für die Einsicht steht, dass eine Datenbank praktisch zu keiner Zeit ein perfektes Bild der realen Welt darstellt.

Die Behandlung von Imperfektion in Datenbanken liegt somit nahe, jedoch stellt sie höhere Anforderungen an:

- Ausdrucksstärke der Datenbank
- Leistungsfähigkeit der Datenbank und der Rechner
- Folgerungen der Anwendungen aus imperfekten Daten

# Wo kann Imperfektion auftreten?

Imperfektion kann an folgenden Stellen auftreten:

- in der Datenbasis
  - innerhalb eines Datums
  - zwischen mehreren Daten
- im Datenbankschema

# Imperfektion in der Datenbasis

Zunächst soll die Imperfektion in der Datenbasis betrachtet werden und die auftretenden Arten unterschieden werden.

# Imperfektion innerhalb eines Datums

Hierbei handelt es sich um Imperfektion, die in einem Datum selbst liegt, ohne Bezug auf andere Daten der Datenbasis.

Die folgenden Arten der Imperfektion innerhalb eines Datums können unterschieden werden:

- Unvollständigkeit
- Unsicherheit
- Ungenauigkeit
- Vagheit

# Unvollständigkeit

Unvollständigkeit bezeichnet das Fehlen eines (relevanten) Datums.

Hierbei können zwei Arten unterschieden werden:

- **existentiell:** Es existiert ein Wert, aber er ist unbekannt.
- **universell:** Alle Elemente einer Menge haben denselben Wert, aber der Wert ist unbekannt.

siehe: „*null*-Werte im Relationalen Modell“

# Beispiel zu Unvollständigkeit

Beispiel (existenzielle Unvollständigkeit):

- Es ist keine Angabe zum Alter von A vorhanden (aber A lebt noch und hat somit ein Alter).

Beispiel (universelle Unvollständigkeit):

- Alle Personen (einer bestimmten Menge) haben dieselbe Religion, aber es ist unbekannt, um welche es sich handelt.

# Unsicherheit

Unsicherheit bezeichnet das Fehlen von Information über den wahren Wert, aber ein unsicherer und mit einem Maß versehener, Wert ist vorhanden.

Möglichkeiten zur Quantifizierung von Unsicherheit sind:

- Wahrscheinlichkeit
- Möglichkeit
- Glaubwürdigkeit und Plausibilität

siehe: „Theorien für die Darstellung von Unsicherheit - Ein Vergleich der Wahrscheinlichkeits-, Möglichkeits- und Dempster-Shafer Theorie“ und „Der probabilistische Ansatz“

# Beispiel zu Unsicherheit

Beispiel (Unsicherheit):

- Es ist nicht sicher, dass das Alter von A 30 Jahre ist.

Durch Verwendung eines Maßes für die Unsicherheit erhält man eine Maßzahl, die die Unsicherheit des Wertes beschreibt.

Beispiel (quantifizierte Unsicherheit):

- Das Alter von A ist mit Wahrscheinlichkeit 0,8 30 Jahre.

# Ungenauigkeit

Ungenauigkeit meint, dass das Datum nicht genau bekannt ist, aber eine Näherung - ein ungenaues Datum - existiert.

Es ist zu beachten, dass dies von einer völligen Unbrauchbarkeit des Datums bis zur problemlosen Verwendung je nach Zweck eine große Breite an Auswirkungen haben kann.

Unterscheidbar sind die folgenden Arten:

- intervallwertig
- diskret
- negativ

# Beispiel zu Ungenauigkeit

Beispiel (Ungenauigkeit, intervallwertig):

- A ist zwischen 20 und 40 Jahre alt.

Beispiel (Ungenauigkeit, diskret):

- A ist entweder 20, 25 oder 30 Jahre alt.

Beispiel (Ungenauigkeit, negativ):

- Es ist bekannt, dass A nicht mit B verheiratet ist. Der Familienstand von A ist dadurch jedoch nicht bekannt.

# Vagheit

Vagheit bezeichnet das Fehlen einer genauen Bestimmung eines natürlichsprachlichen Begriffs, etwa jung, schnell, hoch, für den dann eine entsprechende linguistische Variable benutzt wird, deren Bedeutung jedoch nicht exakt ist und die je nach Zusammenhang verschiedene Wertigkeit haben kann.

Eine solche linguistische Variable kann etwa als Fuzzy-Wert dargestellt werden.

siehe: „Fuzzy-Ansatz und Fuzzy-Architekturen“ und „Rough Sets“

# Beispiel zu Vagheit

Beispiel (Vagheit):

- A ist „jung“.

Diese Information ist für die Kundensegmentierung bereits ausreichend.

- Das Verkehrsmittel ist „schnell“.

# Imperfektion zwischen Daten

Bisher wurde die Imperfektion jeweils eines einzelnen Datums behandelt. Imperfektion kann jedoch auch zwischen zwei oder mehreren Daten vorhanden sein.

Eine Datenbank wird hier als in sich konsistent aufgefasst, so dass eine Imperfektion zwischen Daten dann vorkommen kann, wenn sich die Bedeutung der Daten im Zeitablauf ändert oder wenn Daten aus unterschiedlichen Quellen zusammengeführt werden sollen.

# Unvereinbarkeit

Unvereinbarkeit von Daten meint, dass zwei oder mehr Daten vorhanden sind, die dieselbe Bedeutung besitzen, sich jedoch widersprechen, so dass sie nicht zu einem gemeinsamen Datum zusammengefasst werden können.

Imperfektion zwischen Daten kann entstehen durch:

- zusammenführen von Daten aus unterschiedlichen Quellen
- sich im Lauf der Zeit ändernde Umstände

# Beispiel zu Unvereinbarkeit

Beispiel (Unvereinbarkeit, unterschiedliche Quellen) 1:

- Datenbank 1: Alter von A ist 30
- Datenbank 2: Alter von A ist 32

Beispiel (Unvereinbarkeit, durch Zeitablauf):

- Das Alter 30 und Geburtsjahr 1973 stimmt im Jahr 2010 nicht mehr.
- Die Mehrwertsteuer wird erhöht, so dass die Preise „inklusive MwSt“ nicht mehr stimmen.

# Beispiel zur Vereinbarkeit

Es ist bereits erkennbar, dass Daten, die nicht identisch sind, durchaus vereinbar sein können.

Beispiel (Vereinbarkeit):

Datenbank 1: Alter zwischen 20 und 40

Datenbank 2: Alter zwischen 30 und 50

Hier ist als gemeinsames Datum etwa „Alter zwischen 30 und 40“ verwendbar, also ein Datum, welches in keiner der Datenbanken zuvor vorhanden war.

Weitere Möglichkeiten der Zusammenführung sind vorhanden.

# Imperfektion im Datenbankschema

Die Gründe für eine imperfekte Datenbank können nicht nur in der Datenbasis (und der Datenerfassung) begründet sein, sondern können bereits in der Formulierung des Datenbankschemas angelegt sein.

Die Imperfektion kann entstehen durch:

- Struktur der Daten ist unvollständig spezifiziert
- Daten passen nicht in die spezifizierte Struktur

siehe: „Schemaerweiterung zur Abbildung von imperfekten Daten“

# Zitat

„Imperfektion durchdringt unser Begreifen der realen Welt. Die reale Welt nachzubilden, ist der Zweck von Informationssystemen. Folglich müssen Informationssysteme mit Imperfektion umgehen können.“

(nach A. Motro, 1993)

# Literaturverzeichnis

BERZTISS, Alfs T.: Uncertainty Management. In: CHANG, Shi-Kuo (Hrsg.): *Handbook of Software Engineering and Knowledge Engineering*. Bd. 2 : *Emerging Technologies*. River Edge : World Scientific, 2002  
<ftp://cs.pitt.edu/chang/handbook/01UNb.pdf>

PARSONS, Simon: Current approaches to handling imperfect information in data and knowledge bases. In: *IEEE Transactions on Knowledge and Data Engineering* 8 (1996), Nr. 3, S. 353-372

ZIMÁNYI, Esteban ; PIROTTE, Alain: Imperfect knowledge in databases. In: MOTRO, Amihai (Hrsg.) ; SMETS, Philippe (Hrsg.): *Uncertainty Management in Information Systems : From Needs to Solutions*. Boston : Kluwer, 1996, S. 35-87  
<http://cs.ulb.ac.be/publications/P-97-01.pdf>



Vielen Dank für die  
Aufmerksamkeit



Für Fragen stehe ich  
zur Verfügung