Klassifikation von Integrationskonflikten

Inhaltsverzeichnis

Inhaltsverzeichnis

1. Was bedeutet Integration?

Definition: Integration

Die Datenintegration erfordert die **einheitliche** Verwaltung **aller** von Anwendungen benötigten Daten. Hier verbirgt sich die Möglichkeit der kontrollierten nicht-redundanten Datenhaltung des gesamten relevanten Datenbestandes.[1]

[1]Andreas Heuer, Gunter Saake, Datenbanken: Konzepte und Sprachen, S. 7

2. Strukturelle Heterogenitätskonflikte

Es liegen unterschiedliche Datenstrukturen in unterschiedlichen Informationssystemen vor

Grund:
- Unterschiedliche Informationssysteme werden für unterschiedliche Anwendungen konzipiert
- Persönliche Vorlieben des Designers

Voraussetzung zur Beseitigung des Konflikts:
Informationen müssen semantisch äquivalent sein

2.1 Konflikte bei bilateralen Korrespondenzen

Bei bilateralen Korrespondenzen vergleicht man zwei Strukturelemente miteinander.

- Bezeichnerkonflikte
- Datentypkonflikte
- Integritätskonflikte

  Integritätsconstraint-Konflikt, Schlüsselkonflikt

2.1 Konflikte bei bilateralen Korrespondenzen

2.1.1 Bezeichnerkonflikte:
2.1.1.1 Synonyme

Beispiel für ein Synonym:
Bank A nennt die Relation Person
Bank B nennt die Relation Kunde
-> Semantische Äquivalenz, aber unterschiedliche Bezeichner

Konflikt ist einfach durch Umbenennung lösbar.

2.1 Konflikte bei bilateralen Korrespondenzen

2.1.1 Bezeichnerkonflikte:
2.1.1.2 Homonyme

Beispiel für ein Homonym:
Bank A hat eine Relation namens Person
Bank B hat ein Relation namens Person
Beide Relationen unterscheiden sich aber in ihren Attributmengen
-> syntaktisch gleiche Bezeichner, aber semantische Heterogenität

Konflikt ist einfach durch Umbenennung lösbar.

2.1 Konflikte bei bilateralen Korrespondenzen

2.1.2 Datentypkonflikte: String vs. Integer

Bank A speichert den Familienstatus eines Kunden als String (ledig, verheiratet, geschieden, verwitwet)

Bank B speichert den Familienstatus eines Kunden als Integer (1=ledig, 2=verheiratet, 3=geschieden, 4=verwitwet)

Sofern eine problemlose Konvertierung der Datentypen möglich ist, ist der Konflikt einfach zu lösen.

Wiederholung: Integritätsbedingung

Definition: Integritätsbedingung

Allgemein wird als **Integritätsbedingung** eine Bedingung für die “Zulässigkeit” oder “Korrektheit” bezeichnet. In Bezug auf Datenbanken kann diese Bedingung

- (einzelne) Datenbankzustände
- Zustandsübergänge oder auch
- langfristige Datenbankentwicklungen

betreffen[2]

[2]Andreas Heuer, Gunter Saake, Datenbanken: Konzepte und Sprachen, S. 496

2.1 Konflikte bei bilateralen Korrespondenzen

2.1.3 Integritätskonflikt:

1. Default-Wert-Konflikt
2. Schlüsselkonflikt
3. Integritätsconstraint-Konflikt

1. **Bank A** speichert Newsletter-Attribut in Boolean
   Bank B erlaub NIL-Werte in Newsletter-Attribut
2. Funktionale Abhängigkeit ist durch verschiedene Attributmengen gegeben
3. Ist eine Person einmal als verheiratet eingetragen kann sie nie wieder den Zustand ledig annehmen

2.2 Konflikte bei multilateralen Korrespondenzen

Vergleich zweier Mengen gleicher Strukturelemente; Was passiert, wenn eine Information auf mehrere Strukturelemente verteilt ist?

Konflikte bei

- multilateralen Attribut Korrespondenzen
- multilateralen Entität Korrespondenzen
- fehlenden Informationen

2.2 Konflikte bei multilateralen Korrespondenzen

2.2.1 Konflikt bei multilateralen Attribut Korrespondenzen

Bank A speichert den Namen eines Kunden in zwei Attributen, Name und Vorname

Bank B speichert den Namen eines Kunden in einem Attribut Name, wobei Nachname und Vorname durch Komma getrennt sind

Konfliktlösung durch Funktion, die aus einem Attribut zwei, oder die aus zwei Attributen eins macht.

2.2 Konflikte bei multilateralen Korrespondenzen

2.2.2 Konflikt bei multilateralen Entität Korrespondenzen

Bank A speichert die kompletten Daten eines Kunden in einer Relation

Bank B speichert die Daten eines Kunden in mehreren Relationen, wobei die jeweiligen Datensätze über den Schlüssel KundenId identifiziert werden
-> Ein Kunde kann mehrere Konten besitzen

Lösung des Konflikts durch Funktion, die Informationen aus verschiedenen Informationsquellen selektiert und zusammenfasst -> relationale Vollständigkeit.

2.2 Konflikte bei multilateralen Korrespondenzen

2.2.3 Konflikt bei fehlenden Informationen

Bank A verteilt Newsletter mit neuen Inhalten über ihren Status

Bank B verteilt Newsletter mit aktuellen Börsendaten der Frankfurter Börse

-> Bei der Integration beider Systeme muss herausgefunden werden, welcher Kunde zu welcher Bank gehört, damit der richtige Newsletter verschickt werden kann

Lösung des Konfliktes ist möglich genau dann, wenn die fehlende Information ermittelt werden kann.

2.3 Metaebenen-Diskrepanzen

Ungleiche Strukturelemente in unterschiedlichen Informationssystemen, bei semantisch gleicher Information.

Konflikte bei

- Daten-Attribute-Korrespondenzen
- Daten-Entitäten-Korrespondenzen
- Attribute-Entitäten-Korrespondenzen

2.3 Metaebenen-Diskrepanzen

2.3.1 Konflikt bei Daten-Attribute Korrespondenzen

Bank A gibt bei Angabe von KundenId und KtoNr das Guthaben in Euro aus

Bank B gibt nur bei Angabe von KundenId, KtoNr und Datum das Guthaben in Dollar zu dem angegebenen Zeitpunkt aus

2.3 Metaebenen-Diskrepanzen

2.3.2 Konflikt bei Daten-Entitäten Korrespondenzen

Bank A selektiert über die KundenId den Familienstatus in eine eigenen Relation namens Familienstatus

Hierbei korrespondiert der entsprechende Attributname der ersten Entität mit dem Namen der Zielrelation.

2.3 Metaebenen-Diskrepanzen

2.3.3 Konflikt bei Attribute-Entitäten Korrespondenzen

Bestimmung mehrerer Datensätze durch ein Attribut.

Bank A kann durch die KundenId mehrere Datensätze mit Informationen eines bestimmten Kunden selektieren, z.B. KtoNr

3. Semantische Heterogenitätskonflikte

Beziehung auf die in einem System enthaltenden Informationen und nicht auf die des Systems zugrunde gelegene Struktur

- Semantische Datenheterogenität
- Semantische Domänenheterogenität

Konflikte sind schwer zu finden, da es zumeist an einer ausreichenden Dokumentation mangelt.

3.1 Semantische Datenheterogenitäten

Man nimmt an, dass verschiedene Daten semantisch übereinstimmen, aber sich in ihrer Syntax unterscheiden.

- Skalierungs- und Einheitskonflikt
- Repräsentationskonflikt
- surjektiver Abbildungskonflikt

3.1 Semantische Datenheterogenitäten

3.1.1 Skalierungs- und Einheitenkonflikte

Zwei semantisch identische numerische Werte können sich in ihrem Wert unterscheiden.

Bank A speichert den Kontostand in Euro
Bank B speichert den Kontostand in Dollar, die Angabe erfolgt hier allerdings in Cent

Die Aufdeckung dieser Konflikte und die Behebung durch eine Konvertierungsfunktion ist zwingend notwendig.

3.1 Semantische Datenheterogenitäten

3.1.2 Repräsentationskonflikte

Unterschiedliche Darstellung semantisch gleicher Werte.

Bank A speichert das Guthaben als Absolutbetrag und vermerkt das Vorzeichen an der letzten Stelle

Bank B gibt das Vorzeichen des Guthabens nur dann an der ersten Stelle an, wenn das Guthaben < 0 ist.

Die Aufdeckung und Behebung dieser Konflikte durch eine Konvertierungsfunktion ist zwingend notwendig.

3.1 Semantische Datenheterogenitäten

3.1.2 Surjektive Abbildungskonflikte

Abbildung von Wertemengen unterschiedlicher Grösse, wobei die Bijektivität verloren geht

Bank A hält einen Kunden für kreditwürdig, wenn er ledig, sein Kontostand > 0 und sein monatliches Einkommen > 3000 Euro ist.

Bank B klassifiziert einen Kunden als kreditwürdig, wenn sein Kontostand und sein monatliches Einkommen > 0 ist.

Wenn die von **Bank B** integrierten Daten teilweise keine Auskunft über den Familienstatus geben, liegt ein surjektiver Abbildungskonflikt vor, weil **Bank A** diesen Kunden automatisch als kreditwürdig anerkennt.

Lösung dieses Konfliktes ist schwierig, aber möglich, wenn eine Abbildungsfunktion gefunden wird.

3.2 Semantische Domänenheterogenitäten

Die zu integrierenden Informationssysteme unterscheiden sich in ihren Konzeptualisierungen.

- Subsumptionskonflikt
- Überlappungskonflikt
- Inkompatibilität
- Aggregationskonflikt

3.2 Semantische Domänenheterogenitäten

3.2.1 Subsumptionskonflikte

Bei einem Subsumptionskonflikt ist eine Repräsentantenmenge Untermenge einer anderen Repräsentantenmenge.

Bank A benötigt zu einer KundenId neben dem Kontostand auch immer den Familienstand um eine mögliche Kreditwürdigkeit auszuschliessen.
Bank B liefert zu der KundenId jedoch nur den Kontostand, um eine mögliche Kreditwürdigkeit in Betracht zu ziehen.
-> Bank B braucht nicht alle Informationen von Bank A

Lösung des Konfliktes ist relativ einfach.

3.2 Semantische Domänenheterogenitäten

3.2.2 Überlappungskonflikte

Bei den Überlappungskonflikten müssen sich die jeweiligen Repräsentantenmengen überschneiden.

Bei einer Anfrage an das Informationssystem gibt
Bank A bei der KundenId Name, Vorname, KtoNr und Guthaben aus
Bank B hingegen gibt durch die KundenId Name, KtoNr, PLZ und Guthaben aus.

Lassen sich die Datensätze selektieren, so ist dieser Konflikt lösbar.

3.2 Semantische Domänenheterogenitäten

3.2.3 Inkompatibilitäten

In diesem Fall liegen unterschiedliche Repräsentantenmengen zweier Entitiäten vor.

Falls die Repräsentantenmengen semantsich identisch sind, liegt der Konflikt sehr wahrscheinlich in den unterschiedlichen Abstraktionsebenen begründet.

Bank A versteht unter Ort das Bundesland
Bank B versteht unter Ort den zur PLZ passenden Kreis

3.2 Semantische Domänenheterogenitäten

3.2.4 Aggregationskonflikte

Zwei Informationssysteme unterscheiden sich bei der Konzeptualisierung einer Domäne im Detaillierungsgrad.

Bank A hält in ihrem Relationenschema die Kontoart fest (Girokonto, Sparkonto, etc.).
Bank B kann keine Auskunft über die Kontoart geben.

4. Inkonsistenz- und Redundanzprobleme

Definition: Konsistenzerhaltung

Die Konsistenzerhaltung fordert, dass Regeln und Einschränkungen, die im Eingabedokument gewährleistet wurden, auch in der neuen Modellierung respektiert werden.[3]

[3]Andreas Heuer, Gunter Saake, Datenbanken: Konzepte und Sprachen, S.170

4. Inkonsistenz- und Redundanzprobleme

- Ungenauigkeiten der Datenwerte
- Temporale Abhängigkeiten von Informationssystemen
- Fehlende Daten

Diese Art von Problemen gilt überwiegend als unlösbar.

4. Inkonsistenz- und Redundanzprobleme

4.1 Datenungenauigkeitskonflikte

Zwei Werte, die gleich sein sollten, unterscheiden sich in unterschiedlichen Informationssystemen.

Bank A speichert Euro auf 2 Nachkommastellen genau
Bank B speichert Euro auf 8 Nachkommastellen genau
-> Erheblicher Wertunterschied bei Umrechnung von Euro in Dollar

Diese Art von Konflikt ist sehr schwer zu lösen.

4. Inkonsistenz- und Redundanzprobleme

4.2 Temporale Inkonsistenzkonflikte

Die Daten in dem einen System sind aktueller als die Daten in dem anderen System.

Bank A berechnet die aktuellen Währungskurse minütlich
Bank B berechnet die aktuellen Währungskurse stündlich
-> Diskrepanzen bei Umrechnung von Euro in Dollar

Diese Art von Konflikt ist sehr schwer zu lösen.

4. Inkonsistenz- und Redundanzprobleme

4.3 Konflikte durch fehlende Daten

In einem Datensatz fehlen Datenwerte.

Bank A lässt NIL-Werte für Familienstatus zu
Bank B braucht Familienstatus um die Kreditwürdigkeit eines Kunden eindeutig bestimmen zu können

Konflikt ist lösbar, wenn das Fehlen von Datenwerten vorhersehbar ist und diese durch Defaultwerte ersetzt werden können (Deduktion).

4. Inkonsistenz- und Redundanzprobleme

4.4 Redundanzprobleme

Zwei zu integrierende Informationssysteme beinhalten gleiche Informationen.

Beseitigung der Redundanz durch eindeutige Identifizierung.

Problem: Finden einer eindeutigen Identifizierung.

Fazit

Je inhomogener die zu integrierenden Informationssysteme, desto mehr Integrationskonflikte treten auf. Daher wird zur Lösung eine Kombination von Integrationsansätzen benötigt.
Diese Integrationsansätze müssen äußerst komplex sein, um alle auftretenden Konflikte abfangen zu können.

Welche Ansätze hier konkret vorgenommen werden, sehen wir nächste Woche...

Habt Ihr Fragen oder Anmerkungen.....

..... ?

Vielen Dank für Eure Aufmerksamkeit !

Quellen- und Literaturverzeichnis

- Holger Wache (2003): Semantische Mediation für heterogene Informationsquellen. Akademische Verlagsgesellschaft Aka GmbH, Dissertationen zur Künstlichen Intelligenz, Berlin.
- Andreas Heuer, Gunter Saake (2000): Datenbanken: Konzepte und Sprachen. Bonn: mitp-Verlag, 2000 ISBN: 3-8266-0619-1