Disambiguating Databases

Rendl Roland

27. Mai 2015

1 Einleitung

Data ambiguity oder zu deutsch Daten Mehrdeutigkeit bezeichnet den Umstand, dass Wörter, Messwerte, Ergebnisse, etc.. verschieden interpretiert werden können, oder nicht eindeutig definiet sind. Teilweise ist dieses Problem schon von Menschen nicht leicht zu beheben, noch schwieriger wird es, wenn ein Computer damit konfrontiert wird.
Die folgende Seminararbeit gibt einen kurzen Überblick über solche Mehrdeutigkeiten in Datenbanken und wie man diese handhaben und beheben kann.

2 Wörter mit mehreren Bedeutungen

Es gibt Wörter, die bei gleicher Schreibweise oder Aussprache je nachdem in welchem Kontext sie stehen, unterschiedliche Bedeutungen haben. Diese Wörter nennt man auch Homonyme. Hier kommt es vor allem zu Problemen, wenn geschriebener Text in eine sprachliche Ausgabe umgewandelt werden soll oder umgekehrt, da auf den ersten Blick häufig nicht klar ist, welche Bedeutung nun gemeint ist.
Beispiele hierfür sind:

Homophone mit verschiedener Schreibweise:

- die Ahle (Werkzeug), Aale (Fische)
- der Arm (Körperteil), arm (mittellos)
- Lose (Plural von Los), lose (nicht angebunden)
- malen (abbilden), mahlen (zerreiben)
- das Meer (Gewsser), mehr (Gegensatz von weniger)

Homophone mit verschiedener Aussprache:

- die Lache (Art des Lachens), die Lache (Pfütze)
- mdern (verrotten), modern (neumodisch, neuzeitlich)
- die Hochzeit (Vermählung), die Hochzeit (Höhepunkt)
- übersetzen (in eine andere Sprache bertragen), übersetzen (einen Fluss berqueren)
- Sie rasten (fuhren schnell) über die Autobahn Hier könnten wir rasten (ausruhen)

Wie kann man also dieses Problem beheben?

2.1 Strategien zur Behebung dieses Problems

- **Leichte Methode:**
  Verwendung von umstehenden Wörtern oder Wortgruppen zur Feststellung der Bedeutung des Wortes. Diese Methode ist verhältnismäßig einfach zu implementieren, allerdings ist sie auch fehleranfälliger als die zweite Methode, vor allem bei mehreren Worten mit verschiedenen Bedeutung in einem Text oder Datensatz.
- **Tiefschürfende Methode**
  Bei dieser Methode werden für die Wörter alle Bedeutungen aus Wörterbüchern, Lexika und anderen Quellen geladen um alle möglichen Bedeutungen abzudecken. Diese Methode ist genauer, erfordert dafür aber auch einen höheren Aufwand, vor allem da eine Datenbank, die eine zufriedenstellende Genauigkeit ermöglicht, sehr groß und schwierig zu erstellen ist.

[3]

3 Mehrdeutigkeiten in Datenbanken nach einem JOIN

Ein Join in einer Datenbank fügt 2 Tabellen zu einer gemeinsamen Tabelle zusammen. Dies kann bei entsprechenden Spaltennamen dazu führen, dass in der neuen Tabelle mehrere Spalten mit dem selben Namen vorkommen.

Mit Hilfe eines Beispiels wird dies nocheinmal genauer beleuchtet.
Untenstehend sind 2 Tabellen aufgelistet, die mit Hilfe eines Equi-JOIN zusammengefügt werden:

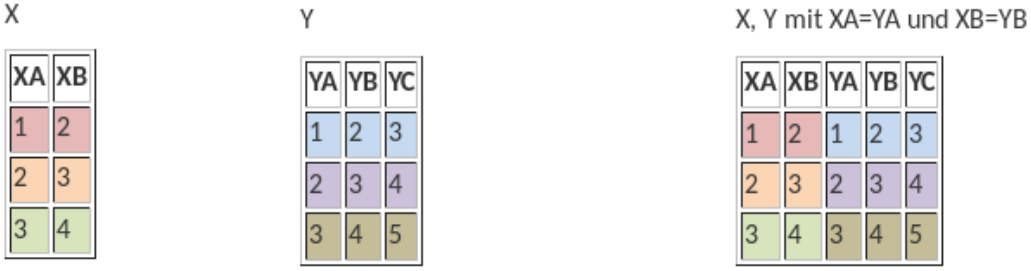

Abbildung 1: Equi-JOIN [1]

Man sieht, dass die so entstandene Tabelle 2 Spalten namens A und B enthält. Hier ist dies kein Problem, da die Spalten dieselben Einträge besitzen, aber was kann man tun, um die richtige Spalte abzufragen, wenn die Einträge nicht ident sind?

3.1 Alle Spaltennamen eindeutig vergeben

Man kann alle Spalten in allen Tabellen so benennen, dass sich keine Überschneidungen ergeben, was für kleine Datenbanken noch ganz gut funktioniert, aber spätestens bei etwas größeren Projekten schnell kompliziert und nicht mehr rentabel ist. Daher ist diese Methode (zumindest für größere Datenbanken) nicht empfehlenswert.

3.2 Zugriff über den Namen der Ausgangstabellen

Wie in Abbildung1 schon angedeutet, kann auf die jeweilige Spalte mit dem Namen der Ausgangstabelle und dem jeweiligen Spaltennamen zugegriffen werden (z.B. x.A). Dies verhindert zwar nicht das prinzipielle Vorhandensein von gleichnamigen Spaltennamen in der neu erstellten Tabelle, wird allerdings häufig in der Praxis verwendet. Nachteil ist, dass es bei größeren Datenbank schnell zu einer gewissen Unübersichtlichkeit kommt, auf welche Tabelle man nun tatsächlich verweist oder verweisen möchte.

3.3 Name Aliasing

Wie wir bereits festgestellt haben, sind die bisher beschriebenen Methoden zwar durchaus funktionsfähig, jedoch haben beide gewisse Nachteile, die sie für größere Projekte nicht empfehlenswert machen. // Abhilfe schafft hier das sogenannte 'Name Aliasing' oder zu deutsch Namensersetzung, das mit Hilfe des 'AS'-Operators durchgeführt wird.

Hierbei wird jedem Spaltennamen aus den ursprünglichen Tabellen ein neuer Name zugewiesen, über den dann auf die jeweilige Spalte zugegriffen werden kann. **Beispiel** (übernommen von [2]):

shop_id	shop
1	Zurich
2	New York
3	London

id	shop_id	gender	name	salary
1	1	m	Jon Simpson	4500
2	1	f	Barbara Breitenmoser	(NULL)
3	2	f	Kirsten Ruegg	5600
4	3	m	Ralph Teller	5100
5	3	m	Peter Jonson	5200
15	2	f	Kirsten Ruegg	5550

```
SELECT emp.shop_id AS 'Shop ID',
       emp.name    AS 'Full Name',
       emp.gender  AS 'Gender',
       emp.salary  AS 'Rate',
       sh.shop     AS 'Store'
FROM   employees   AS emp,
       shops       AS sh
WHERE  emp.shop_id = sh.shop_id;
```

Shop ID	Full Name	Gender	Rate	Store
1	Jon Simpson	m	4500	Zurich
1	Barbara Breitenmoser	f	(NULL)	Zurich
2	Kirsten Ruegg	f	5600	New York
3	Ralph Teller	m	5100	London
3	Peter Jonson	m	5200	London

Wie man sehen kann, wurden die Namen der einzelnen Spaltennamen entsprechend geändert und in die neue Tabelle eingefügt.

4 Schlussfolgerung

Mehrdeutigkeiten in Datenbanken bereiten mehr Probleme, als man am Anfang denkt. Dies hat diese kurze Seminararbeit gezeigt, wobei aufgrund des Umfangs nur auf einige wenige eingegangen werden konnte.

Literatur

[1] Abbildung1: Equi-join. `http://wiki.selfhtml.org/wiki/Datenbank/Einf%C3%BChrung_in_Joins`. [Online; aufgerufen 26-Mai-2015].

[2] Rob Gravelle. Disambiguating between duplicate column names in mysql. `http://www.databasejournal.com/features/mysql/article.php/3904451/Disambiguating-between-Duplicate-Column-Names-in-MySQL.htm`. [Online; aufgerufen 26-Mai-2015].

[3] Margaret Rouse. Disambiguation definition. `http://searchdatamanagement.techtarget.com/definition/disambiguation`. [Online; aufgerufen 26-Mai-2015].