Aufgabenblatt 2
Schema Matching

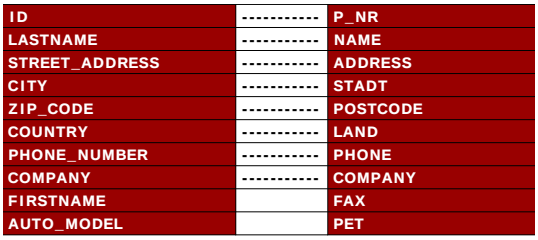
- Abgabetermin: **Dienstag, 07. Juni 2007**
- Das Aufgabenblatt gilt als bestanden, wenn beide Aufgaben bearbeitet und mindestens 20 der 40 Punkte erreicht werden.
- Die Aufgaben sollen in Zweiergruppen bearbeitet werden.
- Abgabe: Per E-Mail an Alexander Albrecht, Fachgebiet Informationssysteme.

In dieser Übung soll zwischen zwei Tabellen ein Matching automatisch gefunden werden. Für die zu matchenden Beispielrelationen wird eine Musterlösung, also die Menge der korrekten Matches, zur Verfügung gestellt. Die in Aufgabe 2 implementierten Matcher werden nach der Abgabe in einem Wettbewerb mit unbekannten Schemata und Datensätzen getestet und beurteilt. Die Qualität der Lösung messen wir mit den Maßen PRECISION und RECALL.

PRECISION mißt die Richtigkeit der gefundenen Matches, indem der Quotient aus den korrekt gefundenen Matches und allen gefundenen Matches gebildet wird.

RECALL mißt die Vollständigkeit der gefundenen Matches, indem der Quotient aus den korrekt gefundene Matches und allen korrekten Matches gebildet wird.

Für beide Maße wird das harmonische Mittel (also die F-Measure) berechnet, das maximiert werden soll.



Beispiel: Die Abbildung zeigt die Schemata zweier Relationen und das Matching zwischen Ihnen. Ein selbst implementierter Matcher gibt die folgenden Matches für die oben gezeigten Schemata aus:

<LASTAME ↔ NAME>
<STREET_ADDRESS ↔ ADDRESS>
<ZIP_CODE ↔ POSTCODE>
<PHONE_NUMBER ↔ PHONE>
<FIRSTNAME ↔ PET>

Es gibt insgesamt 8 korrekte Matches, von denen 4 gefunden wurden. Ein gefundener Match ist inkorrekt. Es ergibt sich also PRECISION = 4/5 und RECALL = 4/8. Das harmonische Mittel beträgt 8/13.

Das Team mit dem maximalen Wert für das harmonische Mittel aus PRECISION und RECALL ist Sieger des Schema Matching Contest 2007. Am Ende des Semesters wird der Gesamtsieger aller Wettbewerbe ermittelt. Die Preisverleihung findet am 19.07. im Rahmen der Vorlesung statt. Die Schemata mit Musterlösungen liegen im Verzeichnis R:\InfoInt07_Naumann\Uebung\0524\.

Aufgabe 1: Nur Schemata

In dieser Aufgabe werden ausschließlich die Schemata der zu matchenden Tabellen zur Verfügung gestellt. Es sind also keine Daten vorhanden. Ein Schema besteht lediglich aus einer Menge von Attributnamen und den zugehörigen Datentypen. Jedes Attribut passt zu höchstens einem anderen Attribut; es bestehen also nur 1:1 Matches. Implementiere einen Matcher, der ein solches Matching findet. **15 P**

Aufgabe 2: Schemata & Daten

In dieser Aufgabe stehen gegenüber Aufgabe 1 zusätzlich zu den Schemata auch entsprechende Datensätze zur Verfügung. Dabei kann u.a. ausgenutzt werden, dass einige Realweltobjekte in beiden Tabellen repräsentiert sind, d.h. es existieren Duplikate. Wie in Aufgabe 1 passt auch hier jedes Attribut zu höchstens einem anderen Attribut. Implementiere einen Matcher, der ein solches Matching findet. **25 P**