Lösung - Übungsblatt 5

(Duplikaterkennung)

Fabian Panse

panse@informatik.uni-hamburg.de

Universität Hamburg

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Aufgabe 1-2: DQ-Maß für Duplikate

Frage: Was bedeutet verdoppeln der Duplikate?

- doppelte Anzahl Duplikate bei gleicher Anzahl Tupel?
  (d.h. Nichtduplikat in Duplikat ändern)
- doppelte Anzahl Duplikate bei gleicher Anzahl Nichtduplikate?
  (d.h. Duplikat hinzufügen)
- Neuen Duplikate im selben Cluster oder in verschiedenen?

Beispiel (Nichtduplikat in Duplikat ändern):

- Gegeben: 10 Tupel, Clusterverteilung: $8 \times 1, 1 \times 2$ (1 Dup)
- Selbe Cluster $\Rightarrow 7 \times 1, 1 \times 3$ (2 Dup)
- Verschiedene Cluster $\Rightarrow 6 \times 1, 2 \times 2$ (2 Dup)

Beispiel (Duplikat hinzufügen):

- Selbe Cluster $\Rightarrow 8 \times 1, 1 \times 3$ (2 Dup)
- Verschiedene Cluster $\Rightarrow 7 \times 1, 2 \times 2$ (2 Dup)

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 1: Anzahl an Duplikaten

- $|Dups|$ = Anzahl überschüssigen Duplikaten

$$QM_1 = |Dups|$$

- Nicht normiert und steigt anstatt zu sinken (hoher Wert = schlechte DQ)
- skaliert kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 2: Inverse Anzahl an Duplikaten

- $|Dups|$ = Anzahl überschüssige Duplikate

$$QM_2 = \frac{1}{|Dups| + 1}$$

- Normiert auf $[0, 1]$ und 1 wenn $|Dups| = 0$
- skaliert nicht kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 3: Durchschnittliche Clustergröße

- $Cluster$ = Menge aller Duplikatcluster
- $|C|$ = Größe von Duplikatcluster $C \in Cluster$

$$QM_3 = \frac{\sum_{C \in Cluster} |C|}{|Cluster|}$$

- Nicht Normiert und steigt statt zu sinken, aber 1 wenn $\sum_{C \in Cluster} |C| = |Cluster|$ was genau dann eintritt, wenn kein Cluster mehr als ein Tupel beinhaltet (d.h. keine Duplikate)
- skaliert nicht kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 4: Cluster-Tupel Ratio

- $Cluster$ = Menge aller Duplikatcluster
- $Tupel$ = Menge aller Tupel

$$QM_4 = \frac{|Cluster|}{|Tupel|}$$

- Normiert auf $[0, 1]$ (da $|Cluster| \leq |Tupel|$) und 1 wenn $|Tupel| = |Cluster|$ was genau dann eintritt, wenn kein Cluster mehr als ein Tupel beinhaltet (d.h. keine Duplikate)
- skaliert nicht kardinal
- Entspricht der inversen durchschnittlichen Clustergröße da $|Tupel| = \sum_{C \in Cluster} |C|$

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 5: Normierte inverse durchschnittlichen Clustergröße

- $Cluster$ = Menge aller Duplikatcluster
- $|C|$ = Größe von Duplikatcluster $C \in Cluster$
- $maxSize$ = maximal möglicher Wert für die durchschnittlichen Clustergröße (also alle Tupel in einem einzigen Cluster, d.h. $maxSize = |Tupel|$)

$$QM_5 = \frac{maxSize - \frac{|Cluster|}{|Tupel|} + 1}{maxSize}$$

- Normiert auf $[0, 1]$ (da $|Cluster| \leq |Tupel|$) und 1 wenn $|Tupel| = |Cluster|$ was genau dann eintritt, wenn kein Cluster mehr als ein Tupel beinhaltet (d.h. keine Duplikate)
- skaliert nicht kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 6: Normierte Cluster-Tupel Differenz

- $Cluster$ = Menge aller Duplikatcluster
- $Tupel$ = Menge aller Tupel
- Beobachtung 1: $|Tupel| - |Cluster| >> 0 \Rightarrow$ viele Duplikate
- Beobachtung 2: $|Tupel| - |Cluster| = 0 \Rightarrow$ keine Duplikate

$$QM_6 = 1 - \frac{|Tupel| - |Cluster|}{|Tupel| - 1}$$

- Normiert auf $[0, 1]$ und 1 wenn $|Tupel| = |Cluster|$ was genau dann eintritt, wenn kein Cluster mehr als ein Tupel beinhaltet (d.h. keine Duplikate)
- skaliert nicht kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 7: Relativer Anteil Duplikate

- $Cluster$ = Menge aller Duplikatcluster
- $Tupel$ = Menge aller Tupel

$$QM_7 = \frac{|Dups|}{|Tupel|}$$

- Normiert auf $[0, 1[$, aber 0 wenn $|Dups| = 0$ (d.h. keine Duplikate)
- skaliert nicht kardinal, kann nicht 1 werden
- Identisch zu Normierte Cluster-Tupel Differenz (da $|Tupel| - |Cluster| = |Dups|$)

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 8: Normierter Anteil Duplikate

- $|Dups|$ = Anzahl Duplikate
- $Tupel$ = Menge aller Tupel

$$QM_8 = \frac{max - |Dups|}{max} \quad \text{mit} \quad max = |Tupel| - 1$$
$$= 1 - \frac{|Dups|}{|Tupel| - 1}$$

- Normiert auf $[0, 1]$ und 1 wenn $|Dups| = 0$ (d.h. keine Duplikate)
- skaliert nicht kardinal
- Identisch zu Normierte Cluster-Tupel Differenz (da $|Tupel| - |Cluster| = |Dups|$)

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 9: Duplikatrate

- $|Dups|$ = Anzahl Duplikate

$$QM_9 = e^{-|Dups|}$$

- Normiert auf $[0, 1]$ und 1 wenn $|Dups| = 0$ (d.h. keine Duplikate)
- skaliert nicht kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Methode 10: Duplikat-Nichtduplikat Ratio

- $|Dups|$ = Anzahl Duplikate
- $Tupel$ = Menge aller Tupel
- $Cluster$ = Menge aller Cluster

$$QM_{10} = \frac{|Dups|}{|Tupel| - |Dups|} = \frac{|Dups|}{|Cluster|}$$

- Nicht normiert und steigt mit wachsender Anzahl Duplikate
- skaliert kardinal

Aufgabe 1-2: DQ-Maß für Duplikate

Selben Maße mit paarweiser Betrachtung:

- $|M|$ statt $|Dups|$
- $|Pairs|$ statt $|Tupel|$
- Schlimmste Fall: $|M| = |Pairs|$
- Beste Fall: $|M| = 0$