Aggregatfunktionen in der Relationenalgebra ?

Dieter Sosna

Gliederung

- Motivation
- Begriffe
- Definitionen
- Anwendungen
- Zusammenfassung

Motivation 1

- Aggregatfunktionen in SQL ermöglichen effiziente Anfragen:
  buch(author, title, isbn, verlag, jahr)
  exemplar (isbn, expl-nr)

  Von welchen Büchern(author, title, isbn) gibt es mindestens zwei Exemplare ?

  SELECT author, title, isbn
  FROM (**buch** JOIN **exemplar**)
  GROUP BY isbn
  HAVING count(isbn) >= 2

Motivation 2

- Lösung Relationenalgebra

  $\mathbf{e1} \Leftarrow exemplar$, $\mathbf{e2} \Leftarrow exemplar$

  $\mathbf{e} \Leftarrow \pi_{isbn}$

  $(\sigma_{e1.isbn=e2.isbn \quad \wedge \quad e1.expl-nr \neq e2.expl-nr}(\mathbf{e1} \times \mathbf{e2}))$

  $\pi_{author,title,isbn}(buch \bowtie \mathbf{e})$

- Lösung nicht generisch bezüglich der Anzahl

Attribute

- Definition:
  Ein Attribut wird gegeben durch
  - den Namen A
  - den Wertebereich $W(A)$
  - (seine semantische Beschreibung)
- Die Referenz auf ein Attribut geschieht durch Benutzen des Namens.

Relationsschema

- Definition:
  Ein Relationsschema $\mathbf{R}(A, B, C, ...)$wird gegeben durch
  - den Namen $\mathbf{R}$ (dieser ist jedoch optional)
  - eine Menge $\{A, B, C, ...\}$ von Attributen $A, B, C, ...$
  - (seine semantische Beschreibung)
- Beispiel: **buch**(author, title, isbn) ,
  dabei müssen die Attribute vorher definiert sein.
- Ein Relationsschema ohne Namen heißt *literal*:

$$\pi_{author,title,isbn}(\mathbf{buch} \bowtie \mathbf{e})$$

Relationsschema

- Definition:
  Ein Relationsschema $\mathbf{R}(A, B, C, ...)$wird gegeben durch
  - den Namen $\mathbf{R}$ (dieser ist jedoch optional)
  - eine Menge $\{A, B, C, ...\}$ von Attributen $A, B, C, ...$
  - (seine semantische Beschreibung)
- Beispiel: **buch**(author, title, isbn) ,
  dabei müssen die Attribute vorher definiert sein.
- Ein Relationsschema ohne Namen heißt *literal*:

$$\pi_{author,title,isbn}(\mathbf{buch} \bowtie \mathbf{e})$$

$$\mathbf{E} \Leftarrow \quad \pi_{author,title,isbn}(\mathbf{buch} \bowtie \mathbf{e})$$

Relation

- Definition:
  Eine Relation $\mathbf{r}(\mathbf{R})$ wird gegeben durch
  - ein Relationsschema $\mathbf{R}$
  - eine Menge von Tupeln
  $t = (t_A, t_B, t_C, ...) \in (W(A) \times W(B) \times W(C) \times ...)$
- Wenn keine Verwechslungen entstehen, schreibt man auch $\mathbf{R}$ statt $\mathbf{r}(\mathbf{R})$.

Relationenalgebra

- Die Menge aller Relationen bildet mit den Verknüpfungen *Vereinigung, Differenz, (mod.) Kreuzprodukt* eine Algebra.
- Wegen *Selektion* σ und *Projektion* π ist es eine Operatoralgebra.
- In der Datenbanktheorie spricht man von *zweistelligen* und *einstelligen* Operatoren.
- Lösung einer Aufgabe mit den Mitteln der R.A. heißt:
  Angabe einer Konstruktionsvorschrift, mit deren Hilfe aus gegeben Relationen unter Benutzung der Operatoren eine neue Relation zu bilden ist, die genau die Tupel enthält, die die in der Aufgabe genannten Eigenschaften besitzen.

Aggregatfunktionen

- Maximum (MAX):
  Sei $\mathbf{r}(\mathbf{R})$ eine Relation zum Schema $\mathbf{R}$,
  A ein Attribut von $\mathbf{R}$, auf dessen Wertevorrat eine Vergleichsoperation definiert ist.
  Somit kann das Maximum einer Menge solcher Werte ermittelt werden.
  Definition:
  $MAX\,[A]\,(\mathbf{R}) := MAX(\pi_A(\mathbf{r}(\mathbf{R}))) := \max_{t \in r}(t.A)$
- $MAX\,[A]\,(\mathbf{R})$ ist eine Abbildung aus der Relationenalgebra in den Wertevorrat $W(A)$.

Aggregatfunktionen (2)

- Analog werden definiert MIN, AVG, $\sum$.
  Bei diesen Definitionen wie auch bei MAX wird vorausgesetzt, dass im Wertebereich des in der Definition auftretenden Attributs eine Ordnung (bei MAX und MIN) bzw. Rechenoperationen (bei AVG und $\sum$) definiert sind),
- $\sharp\,[A]\,(\mathbf{R})$ liefert als Ergebnis die Anzahl der Tupel aus $\mathbf{r}(\mathbf{R})$, für die das Attribut A **keinen unbestimmten Wert** hat.
  $\sharp(\mathbf{R})$, in alternativer Schreibweise $\sharp\,[*]\,(\mathbf{R})$, hat als Ergebnis die Kardinalität von $\mathbf{r}(\mathbf{R})$.

Problem

- Abbildung $MAX(\pi_A(\mathbf{R})) : \mathbf{R} \mapsto W(A)$.

  Das Bild ist ein Wert (aus dem Wertevorrat oder eine natürliche Zahl, also kein Element der Algebra!
- Dies trifft für jeder Aggregatfunktion zu.

Ein Blick zu SQL

- **buch**(author, title, isbn, verlag)

  Wieviele Verlage gibt es?

  SELECT (COUNT(DISTICT verlag) AS Verlagszahl)
  FROM **buch**;

Ein Blick zu SQL

- **buch**(author, title, isbn, verlag)

  Wieviele Verlage gibt es?

  SELECT (COUNT(DISTINCT verlag) AS Verlagszahl)
  FROM **buch**;

- Die SELECT-Anweisung konstruiert eine Relation:

Verlagszahl
43

Einbettungsoperatoren

- Sei A ein Attribut mit dem Wertebereich $W(A)$,
  $x \in W(A)$ ein Wert,
  B ein Attributname.
- Definition: Einbettungsoperator η
  Unter $\eta_B(x)$ soll die Abbildung verstanden werden, die dem Wert x die (literale) Relation vom Grade 1 zuordnet, deren Attribut B heißt und deren einziges Tupel t die Gleichung $t.B = x$ erfüllt.

Anwendung 1

- $\mathbf{buch}$(author, title, isbn, verlag, jahr)

  Wieviele Verlage gibt es?

  $\eta_{Verlagszahl}(\sharp\,[verlag]\,(\mathbf{buch}))$

Anwendung 2

- $\mathbf{buch}$(author, title, isbn, verlag, jahr)

  Welches sind die ältesten Bücher (author, title)?

  $\mathbf{A} \Leftarrow \eta_{alt}(MIN\,[jahr]\,(\mathbf{buch}))$

  $\pi_{author,title}(\sigma_{\mathbf{buch}.jahr=\mathbf{A}.alt}(\mathbf{buch} \times \mathbf{A}))$

Einbettungsoperator mit Gruppierung

Definition:

Sei $\mathbf{r}(\mathbf{R})$ eine Relation zum Relationsschema $\mathbf{R}(A_1, A_2, ..., A_n)$.

Sei $\{B_1, B_2, ..., B_l\}$ eine Teilmenge der Attribute von $\mathbf{R}$,

A ein Attribut mit

$A \in \{A_1, A_2, ..., A_n\}$ und

$A \notin \{B_1, B_2, ..., B_l\}$.

ϕ sei eine der oben definierten Aggregatfunktionen.

(Bei $\sharp\,[*]\,(.)$ ist als Parameter auch das Symbol $*$ anstelle eines Attributs zugelassen.)

X sei ein Attributname.

Einbettungsoperator mit Gruppierung (Forts.)

Dann ist

$$\eta_{B_1,B_2,...,B_l;X}(\phi\,[A)]\,(\mathbf{R})) := \{t|t = (t.B_1, ..., t.B_l, t_X)\} \qquad (1)$$

eine literale Relation $\{t\}$ mit dem Schema $(B_1, B_2, ..., B_l, X)$ und den Tupeln $t = (t.B_1, ..., t.B_l, t.X)$ mit

(1) $(t.B_1, ..., t.B_l) \in \pi_{B_1,...,B_l}(\mathbf{R})$ und

(2) $t.X = \phi\,[A]\,(\sigma_{B_1=t.B_1 \wedge ... \wedge B_l=t.B_l}(\mathbf{R}))$,

(3) dabei erfüllt $\{t\}$ die Bedingung

$$\pi_{B_1,...,B_l}(\{t\}) = \pi_{B_1,...,B_l}(\mathbf{R}).$$

Kommentar

- Eigenschaft (3) bewirkt, dass jede Wertekombination $(t.B_1, ..., t.B_l)$, die in $\mathbf{R}$ vorkommt, auch zu einem Tupel im Ergebnis führt.
- Die Funktion ϕ wird wegen (2) für jede Wertekombination $(t.B_1, ..., t.B_l) \in \mathbf{r}(\mathbf{R})$ neu berechnet und zwar über genau den Tupeln, die bei $\pi_{B_1,...,B_l}(\mathbf{R})$ auf die Wertekombination $(t.B_1, ..., t.B_l)$ abgebildet werden.
- Prozedurale Beschreibung ($\mathbf{E}$ Ergebnismenge):

  $\mathbf{E} = \emptyset$;

  FOR $x \in \pi_{B_1,...,B_l}(\mathbf{R})$ **DO**

  $\mathbf{E} = \mathbf{E} \bigcup (x.B_1, ..., x.B_l, \phi\,[A]\,(\sigma_{B_1=x.B_1 \wedge ... \wedge B_l=x.B_l}(\mathbf{R}))$

  DONE .

Anwendung 3

- Von welchem Buch (isbn) gibt es mehr als drei Exemplare?

  $\mathbf{buch}$(author, title, isbn, verlag, jahr)

  $\mathbf{exemplar}$(isbn, expl-nr)

  (es wird unterstellt, das jedes Buch eine Exemplarnummer $expl - nr$ hat.)

$$\pi_{isbn}(\sigma_{anzahl>3}(\eta_{isbn;anzahl}(\sharp\,[expl - nr]\,(\mathbf{exemplar}))))$$

Zusammenfassung

- Die Aggregatfunktionen sind nach ihrer Definition keine Operatoren der Relationenalgebra. Durch einen Einbettungsoperator können die Funktionswerte auf eine einstellige Relation mit der Kardinalität 1 abgebildet werden.
- Auch die GROUP-BY-Klausel der SQL-Sprache läßt sich mit einem geeigneten Einbettungsoperator in die Algebra integrieren.
- Mit Hilfe der Konstruktion von Einbettungsopratoren lassen sich weitere Funktionen in die Algebra einbringen.

Danke.