

DAB1 – Datenbanken 1

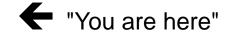
Dr. Daniel Aebi (aebd@zhaw.ch)

Lektion 4: Relationale Algebra, Relationale Bags

Wo stehen wir?



en	Einführung
Logische Grundlagen	Relationale Algebra Relationale Bags
	Entity-Relationship Design
	SQL



Zürcher Hochschule für Angewandte Wissenschaften

Rückblick



- Operationen:
 - Vereinigung ∪
 - Durchschnitt ∩
 - Differenz \ (oder -)
 - Kreuzprodukt (kartesisches Produkt) x
 - Verbundvarianten
 - Natural join M
 - Theta join Ν_θ
 - Outer joins (left, right, full) ⋈, ⋈, ⋈
 - •
- Umbenennung: $\rho_S(R)$ (Umbenennung von R in S) $\rho_{S(D,E)}(R(A,B))$ (Umbenennung R.A \rightarrow S.D, R.B \rightarrow S.E)

Lernziele Lektion 4



- Folgende (binäre) Operation der relationalen Algebra verstehen:
 - Division ÷
- Aggregatfunktionen verstehen
- Grundlagen der 'bag'-Theorie verstehen

Bemerkungen zur Notation



Bisher eingeführt:

- t: Tupelvariable <t(A),t(B),t(C)> oder einfach t(R)
- r, $ext(R) \subseteq dom(A) \times dom(B) \times dom(C)$ (Relation)
- Seien R(A,B,C,D) und S(B,D,E) dann werden die "join-Formate" wie folgt bezeichnet: R-S(A,B,C,D,E) und S-R(B,D,E,A,C)

Bemerkungen:

- Die Mengendifferenz wird oft auch mit dem Symbol (minus) bezeichnet (statt mit backslash), also z.B. S-R statt S\R
- Die Reihenfolge der Attribute spielt eine Rolle, ist aber nicht 'wichtig', es gilt:
 R(A,B,C) ≠ S(B,C,A) (aber S kann durch Projektion auf R erzeugt werden)
- Rangfolge: σ , π , ρ ⇒ x, \bowtie ⇒ \cap ⇒ \cup , \land

Zürcher Hochschule für Angewandte Wissenschaften





Übersicht der Operationen:

- σ Selektion
- π Projektion
- U Mengenvereinigung
- Mengendurchschnitt
- \ Mengendifferenz
- x Kreuzprodukt
- ⋈ Join (Verbund)
- ÷ Division
- ρ Umbenennung
- Linker äusserer Verbund (Symbol, oft weggelassen: ⋈)
- Rechter äusserer Verbund (Symbol, oft weggelassen: ⋈)
- Voller äusserer Verbund (Symbol, oft weggelassen: ⋈)



- Darstellung: R÷S Voraussetzung: S ⊆ R (bezogen auf Attribute)
- Resultat:
 - Ergebnis hat die Differenz der Attributmengen als Attributmenge und umfasst jene Tupel aus R die eingeschränkt auf die Differenz der Attribute (R \ S) für alle Tupel von S denselben Wert haben
- Die Division ist NICHT kommutativ: R÷S≠S÷R
- Der Divisionsoperator erlaubt die kompakte Formulierung von "für-alle"-Anfragen. Kann auch (muss) durch Projektion, Differenz und Kreuzprodukt gebildet werden:

$$R \div S = \pi_{R-S}(R) \setminus \pi_{R-S}((\pi_{R-S}(R) \times S) \setminus R)$$

Attribute von R ohne Attribute von S



Gegeben: 2 Relationenformate R(A, B, C) und S(A, B)

 $t = r \div s$ (t hat Relationenformat T(C))

Dabei gilt:

$$t \bowtie s \subseteq r$$

$$(r \bowtie s) \div s = r$$

 $t = r \div s$ ist die maximale Relation, für welche gilt: $t \bowtie s \subseteq r$

Vergleich ganzzahlige Division:

$$\begin{array}{c|cccc}
\hline
r_1 \\
A & B \\
\hline
a_1 & b_1 \\
a_1 & b_2 \\
a_1 & b_3 \\
a_2 & b_2 \\
a_2 & b_3
\end{array}
\quad \div \quad \begin{bmatrix}
r_2 \\
B \\
b_1 \\
b_2
\end{bmatrix}
\quad = \quad \begin{bmatrix}
r_1 \div r_2 \\
A \\
a_1
\end{bmatrix}$$



- Division in SQL leider nicht implementiert
- Beispiel:

Geg.: Relationenformat Trinken(Besucher, Biersorte, Restaurant) mit zugehöriger Relation r

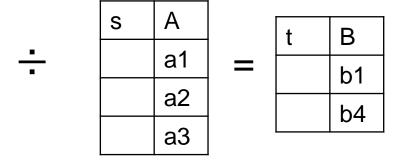
$$(r \div \pi_{Biersorte}(r)) \div \pi_{Restaurant}(r)$$

= Besucher, welche jede (alle) Biersorte in jedem Restaurant trinken



r	Α	В
	a1	b1
	a2	b1
	а3	b1
	a4	b1
	a1	b2
	аЗ	b2
	a2	b3
	а3	b3
	a4	b3
	a1	b4
	a2	b4
	a3	b4

Beispiel Division: Was ist r÷s?



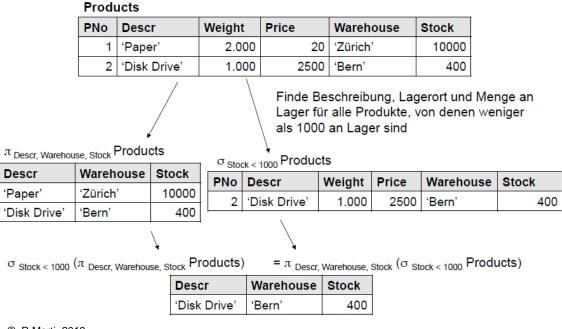
«Bei welchen B-Werten kommen die in s gegebenen A-Werte alle vor?»

Zwischenstand



- Was können wir bisher:
 - Komplexe Anfragen formulieren durch Kombination (Schachtelung) von Ausdrücken. Das Resultat eines (Teil)ausdruckes ist immer wieder eine Relation.
 - Darstellung: Als geschachtelter Ausdruck mittels Klammerung.





Zwischenstand



- Was fehlt?
 - Zusammenfassen von Attributwerten (aggregieren)
 - Gruppieren
 - Sortierung

- ..

Aggregat-Funktionen



- Alle bisherigen Operationen behandeln nur einzelne Tupel. Diese sind ungeordnet in der Menge (Relation) enthalten und «wissen» nichts voneinander. Viele Aufgaben in einer Datenbank müssen zueinander gehörende Daten zusammentragen (dem Lateinischen entlehnt: aggregieren). Wir brauchen also sogenannte Aggregatfunktionen.
- Beispiele:
 - Wie gross ist die Summe aller Verkäufe? Aggregation über alle Tupel
 - Wieviel hat welcher Verkäufer verkauft? Aggregation pro Verkäufer (Gruppierung)
- Typische Aggregatfunktionen: Summe, Durchschnitt, Minimum, Maximum, Anzahl Tupel

School of Engineering InIT Institut für angewandte Informationstechnologie

Gruppierung, Aggregat-Funktionen

- Sei R ein Relationenformat und S ein Teilformat von R (d.h. es gilt R·S = R)
- Sei X ein Attribut (keine Attributkombination!) von R aber nicht von S (d.h. R·X = R, aber S·X ≠ S). Zudem soll dom(X) Summenbildung zulassen.
- Beispiel: R(A,B,X) und S(A), dom(X) = INTEGER

Aggregat-Funktionen

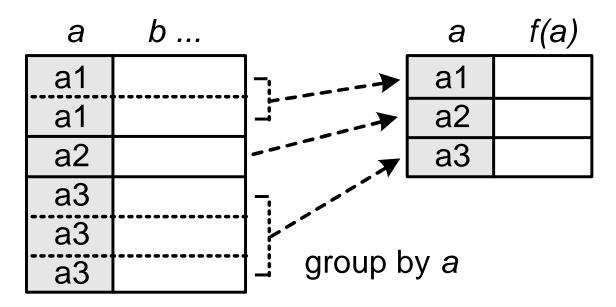


- Pro Gruppe mit gleichen A-Werten wird dann die Summe über Attribut X gebildet wie folgt:
- r = {<a1,b1,2>,<a1,b2,3>,<a2,b1,4>}, Teilformat ist S(A)
- $\Sigma_{S,X}(r) = \{ < a1,5 >, < a2,4 > \}$
- Wir haben also einen neuen Operator: die Summe (Σ)

Gruppierung



- Jede Gruppe ist eine Multimenge von Tupeln mit denselben Werten für das Gruppierungskriterium (Attributkombination)
- Zusammenfassung nach gleichen Werten für Attribut a:



Aggregat-Funktionen



- Analog dem zur Summe gesagten, führen wir einen Aggregats-Operator \mathcal{F} (ausgesprochen: «Skript-F») ein
- Es gibt 5 verschiedene Aggregat-Funktionen:
 - COUNT = Anzahl Tupel
 - MAX = Grösster Wert des betrachteten Attributs *
 - MIN = Kleinster Wert des betrachteten Attributs *
 - SUM = Summe des betrachteten Attributs *
 - AVG = Durchschnitt des betrachteten Attributs *(= Summe / Anzahl)
- * Das Attribut muss einen zählbaren Domain haben

Aggregat-Funktionen



 COUNT (nicht verwechseln mit SUM!) weicht von den anderen Operatoren ab, indem er nicht Attribute bzw. Attributswerte, sondern Tupel zählt

Anzahl =
$$\mathcal{F}_{COUNT(r)}$$

 Bemerkung: Das Resultat einer Aggregatfunktion ist wie bei den anderen relationalen Operatoren wiederum eine Relation (ggf. eine Relation mit einem einzigen Tupel und einem einzigen Attribut)

Aggregat-Funktionen mit Gruppierung



- Allgemein: geg. Relation R(A, B, C, D), C sei ein «zählbares» Attribut
- Resultat = \(\mathcal{F} \) Operator_{s.c}(r), mit Gruppierung beispielsweise auf S(A, B)
- Alternative Schreibweise: Resultat = <A, B> \$\mathcal{F}\$ Operator_c (r)
- r wird nach Attributskombination A, B gruppiert
- Für jede eindeutige Kombination von A, B enthält das Resultat ein Tupel
- Genaueres hierzu folgt im Kapitel SQL!

Bags: Theorie und Praxis



Projektion π

- Eine Konsequenz der Projektion ist die Notwendigkeit einer Duplikatelimination, falls nach der Projektion Tupel "zusammenfallen"
- Aufwand der Duplikatelimination: im wesentlichen gleich zur Sortierung
- Bemerkung: Aufwand Sortierung
 - simple Methoden (Bubblesort, Insertion Sort): O(n²)
 - 1000 Elemente: Grössenordnung 1'000'000 Vergleiche
 - 1'000'000 Elemente: Grössenordnung 1'000'000'000'000 Vergleiche
 - "beste" Methoden (Quicksort, Heapsort u.a.): O(n log_n)
 - 1000 Elemente: Grössenordnung 3000 Vergleiche
 - 1'000'000 Elemente: Grössenordnung 13,000,000 Vergleiche

Relationale Bags



Die Versuchung war gross, sich die Duplikatelimination zu sparen...

- Schon die ersten Datenbanksysteme, welche die relationale Idee aufgriffen, wichen vom relationalen Modell in dieser Hinsicht ab (70er-Jahre)
- Es gelten dann aber nicht mehr genau dieselben Gesetze wie bei relationalen Operationen
- Neuer Zweig der Theorie: "relationale Bags" (U. Dayal, N. Goodman, R.H. Katz: "An Extended Relational Algebra with Control over Duplicate Elimination", 1982)
- SQL ist «nicht ganz relational»

Relationale Bags



- Ein Bag ist ein "Sack", auch Multimenge genannt
- Man kann nicht unterscheiden zwischen identischen Elementen
- "Entnimmt" man ein Exemplar eines mehrfach vorhandenes Elements aus dem "Sack", so wählt man ein willkürliches Exemplar
- Es ist aber unterscheidbar, ob ein Bag ein Element keinmal, einmal, zweimal, dreimal, ... enthält

Beispiel



- Sei R(A,B) gegeben mit dom(A) und dom(B) = STRING
 Anzahl Vorkommen
- Bag r = {<a,b,1>, <a,bb,3>, <aa,b,2>, <aa,bb,1>}
 (wir listen nur jene t ∈ dom(R) mit Anzahl > 0)
- Darstellungen in Tabellenform, Beispiel:
- Reihenfolge nicht festgelegt!

Α	В
a	b
а	bb
a	bb
a	bb
aa	b
aa	b
aa	bb

Α	В
a	bb
aa	bb
a	bb
aa	b
aa	b
a	bb
a	b

Mathematische Definition



- Sei R(A₁, A₂, ... A_n) ein Relationenformat mit den Attributen A₁, A₂, ... A_n und ihren Domänen dom(A₁), dom(A₂), ... dom(A_n) und dom(R) = dom(A₁) x dom(A₂) x ... dom(A_n)
- Ein Bag zum Format R sei eine Abbildung:

$$r: dom(R) \rightarrow \mathbb{N}$$
, $t \mapsto r(t)$, wobei $\mathbb{N} = nat \ddot{u} r liche Zahlen$

 Es wird jedem Tupel t eine natürliche Zahl r(t) ≥ 0 zugeordnet, die Multiplizität (= «Anzahl gleiche Tupel») von t in r

Verhältnis Bag/Relation



- Ein Bag r kann genau dann mit einer Relation r identifiziert werden, wenn für die Tupel t von dom(R) nur die Multiplizitäten 0 und 1 vorkommen, d.h. wenn die Menge {t ∈ dom(R) | r(t) > 1} leer ist.
- Wir sagen dann "der Bag ist eine Relation", da er der Relation {t ∈ dom(R) | r(t) > 0} entspricht

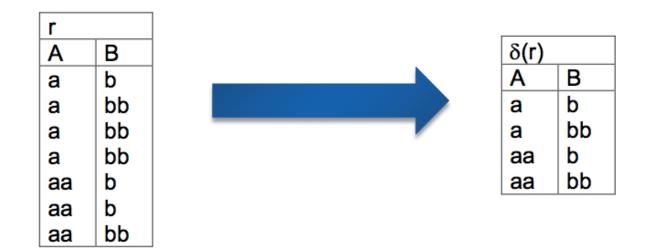
Dies ist die Trägerrelation des Bag r.

Operationen auf Bags



Neue Operation: Duplicate Elimination $\delta(r)$

$$\delta(r) = \{ \langle t, k \rangle \in \text{dom}(R) \ x \ | \ k = \min\{1, r(t)\} \}$$

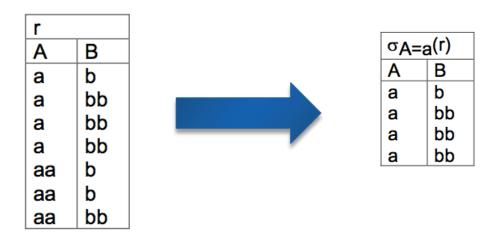


Selektion σ



- Die Selektion bei Bags ist einfach zu verstehen
- Es werden alle Tupel ausgewählt, die die Bedingung erfüllen, zusammen mit ihren Multiplizitäten, Beispiel:

$$\sigma_{A=a}(r) = \{ \langle t,k \rangle \in dom(R) \times \mathbb{N} \mid (t(A) = a \wedge k = r(t)) \vee (t(A) \neq a \wedge k = 0) \}$$



Beispiel: Selektion σ



- Gegeben: Relationenformat R(A,B,C), dom(A)=dom(B)=dom(C)=N
- Bag $r = \{<1,2,3,17>,<4,5,6,1>,<1,8,9,3>\}$ zum Format R
- Berechnen Sie: $\sigma_{A=1}(\sigma_{B=2}(r))$
- Lösung: {<1,2,3,17>}

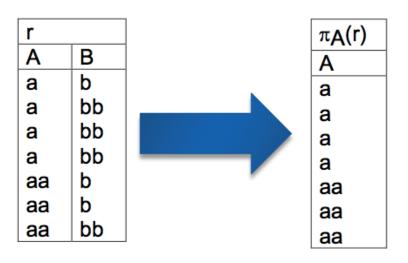
Projektion π



 Die Projektion ist etwas kniffliger. Es können Tupel «zusammenfallen», die vorher verschieden waren. Projektion von u auf das Teilformat R von U:

$$\pi_R(u) = \{ < t, k > \in dom(R) \ x \ | \ k = \Sigma_{x \in dom(U) \ \land \ x(R) = t} \ u(x) \ \}$$

 Beachten Sie: wenn u eine Relation ist, d.h. u(x) ∈ {0,1} für alle x ∈ dom(U), wird die Summe zur Anzahl der x, für welche x(R)=t.



Beispiel: Projektion π



- Sei U(A,B,C) und das Teilformat R(B,C) gegeben
- Bag u={<0,1,1,2>,<1,1,1,3>,<0,0,0,1>,<1,0,0,2>} zum Format U
- Berechnen Sie: $\pi_R(u)$
- Lösung: {<1,1,5>,<0,0,3>}

Join Operator ⋈



 Der (natural) join u = r ⋈ s der Bags r und s ist definiert als (es gilt U=R·S):

$$r\bowtie s = \{<\!t,k\!> \in dom(U) \; x\; \mathbb{N} \; | \; <\!t(R),r(t(R))\!> \in r \; \land \; <\!t(S),s(t(S))\!> \in s \\ \qquad \land \; k = r(t(R)) \; ^*s(t(S))\}$$

 Hier werden die Multiplizitäten von entsprechenden Tupeln der Bags r und s miteinander multipliziert, was im Grenzfall mit r, s als Relationen die bereits bekannte Definition liefert.

r	
Α	В
а	b
а	b
aa	bb
aa	bb

S	
В	С
b	С
bb	С
bb	С
bb	CC



r⋈s		
Α	В	С
а	b	С
а	b	С
aa	bb	cc
aa	bb	СС

Beispiel: Join Operator ⋈



- Bags r1 mit Format R1(A,B) und r2 mit Format R2(B,C)
- $r1 = \{<0,0,2>, <1,0,2>, <1,1,1>\}$
- $r2 = \{<0,0,2>,<0,1,2>,<1,0,1>\}$
- Berechnen Sie: r1 ⋈ r2
- Lösung: {<0,0,0,4>, <0,0,1,4>, <1,0,0,4>, <1,0,1,4>, <1,1,0,1>}

Vereinigung, Durchschnitt, Differenz



- Bei den bisherigen Verallgemeinerungen von Operationen von Bags handelte es sich um solche, die sich aufdrängen
- Bei den Mengenoperationen gibt es aber Konflikte
- Zum Beispiel galt bisher:
 - Gesetz 1: $\sigma_{A=a \vee B=b}(u) = \sigma_{A=a}(u) \cup \sigma_{B=b}(u)$
 - Gesetz 2: $\pi_R(u_1 \cup u_2) = \pi_R(u_1) \cup \pi_R(u_2)$
- Was müssen wir tun, um diese Eigenschaften zu erhalten?

Vereinigung



- Es sei allgemein: r ∪ s = {<t,k> ∈ dom(R) x ℕ | k=φ(r(t),s(t)) }
- Wir wählen nun φ ("phi") geeignet, um die Multiplizitäten zu bestimmen
- Behauptung: mit φ = max gilt Gesetz 1, aber nicht Gesetz 2
- Behauptung: mit φ = + gilt Gesetz 2, nicht aber Gesetz 1

Beispiel zu Gesetz 1



- Gilt: $\sigma_{A=a \vee B=b}(u) = \sigma_{A=a}(u) \cup \sigma_{B=b}(u)$ (Gesetz 1)?
- Beispiel: R(A,B,C) mit $r = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>,<1,0,0,4>\}$
- $\sigma_{A=0 \vee B=1}(r) = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>\}$
- $\sigma_{A=0}(r) = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>\}$
- $\sigma_{B=1}(r) = \{<0,1,0,5>,<0,1,1,3>\}$
- Bei Definition mit φ = max:
 - $\sigma_{\mathsf{A}=0}(\mathsf{r}) \cup \sigma_{\mathsf{B}=1}(\mathsf{r}) = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>\}$
- Bei Definition mit φ = +:
 - $\sigma_{A=0}(r) \cup \sigma_{B=1}(r) = \{<0,0,1,7>,<0,1,0,10>,<0,1,1,6>\}$
- Gesetz 1 gilt (nur) bei Definition mit $\varphi = \max!$

Beispiel zu Gesetz 2



- Gilt: $\pi_R(u_1 \cup u_2) = \pi_R(u_1) \cup \pi_R(u_2)$ (Gesetz 2)?
- Beispiel: R(A,B,C) und S(A,B,C) mit

$$r = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>,<1,0,0,4>\}$$

$$s = \{<0,0,1,3>,<0,1,1,4>,<1,0,1,2>,<1,0,0,3>\}$$

- Bei Definition mit φ = max:
 - $r \cup s = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,4>,<1,0,0,4>,<1,0,1,2>\}$
 - $\pi_{A,B}(r \cup s) = \{<0,0,7>,<0,1,9>,<1,0,6>\}$
- Bei Definition mit φ = +:
 - $r \cup s = \{<0,0,1,10>,<0,1,0,5>,<0,1,1,7>,<1,0,0,7>,<1,0,1,2>\}$
 - $\pi_{AB}(r \cup s) = \{<0,0,10>,<0,1,12>,<1,0,9>\}$

Beispiel zu Gesetz 2



- R(A,B,C) und S(A,B,C)
- $r = \{<0,0,1,7>,<0,1,0,5>,<0,1,1,3>,<1,0,0,4>\}$
- $s = \{<0,0,1,3>,<0,1,1,4>,<1,0,1,2>,<1,0,0,3>\}$
- $\pi_{A,B}(r) = \{<0,0,7>,<0,1,8>,<1,0,4>\}$
- $\pi_{A,B}(s) = \{<0,0,3>,<0,1,4>,<1,0,5>\}$
- Bei Definition mit φ = max:
 - $\quad \pi_{A,B}(r) \cup \pi_{A,B}(s) = \{<0,0,7>,<0,1,8>,<1,0,5>\}$
- Bei Definition mit φ = +:
 - $\quad \pi_{A,B}(r) \cup \pi_{A,B}(s) = \{<0,0,10>,<0,1,12>,<1,0,9>\}$
- Gesetz 2 gilt (nur) bei Definition mit $\varphi = +!$

Vereinigung



- Wir wählen:
- r ∪ s = {<t,k> ∈ dom(R) x N | k=max(r(t),s(t))} "bag union"
- r ⊔ s = {<t,k> ∈ dom(R) x N | k=r(t) + s(t)} "bag concatenation"
- SQL bietet folgendes:

 $\delta(r \cup s)$ als UNION DISTINCT (bzw. als UNION, d.h. default ist DISTINCT) $r \sqcup s$ als UNION ALL

• Es gilt: $r \cup s = (r \setminus s) \sqcup s$

ürcher Hochschule ür Angewandte Wissenschafte





- Zwei ergänzende Definitionen:
- r ∩ s = {<t,k> ∈ dom(R) x N | k = min{r(t),s(t)} }
- $r \ s = \{ < t, k > \in dom(R) \ x \ N \mid k = max\{0, r(t) s(t) \} \}$
- Die Definitionen stimmen im Grenzfall der Relationen mit den entsprechenden Mengenoperationen überein

Zeit für ein Bierchen!



- Gegeben sind die Relationenformate:
 - Gast(Besucher, Restaurant)
 - Sortiment(Restaurant, Biersorte)
 - Vorzug(Besucher, Biersorte)
 - sowie je ein Bag
 - g zum Format Gast
 - s zum Format Sortiment
 - v zum Format Vorzug



- Man weiss nicht, ob die Bags Relationen sind oder nicht.
- Gesucht: Alle Besucher des Restaurant Ochsen, die keine Biere bevorzugen

«The attack of the Clones»



Was ist nun korrekt?

$$\pi_{\text{Besucher}}(\sigma_{\text{Restaurant='Ochsen'}}(g)) \setminus \pi_{\text{Besucher}}(v)$$

ODER

$$\delta(\pi_{\text{Besucher}}(\sigma_{\text{Restaurant='Ochsen'}}(g))) \setminus \pi_{\text{Besucher}}(v)$$

Wie oft gibt es jede Person?
Wie oft verzeichnen wir ihren Besuch in einem Restaurant?

Und weiter...



Das nächste Mal: Schlüssel, Datenbankentwurf

