Big Data Übungsblatt 08

Anton Bulat, Josephine Geiger, Julia Siekiera

December 8, 2017

Aufgabe 1: Kommunikationskosten

Bestimmen Sie die Kommunikationskosten der folgenden Probleme in Abhängigkeit der genannten Eingabegrößen:

a)

Schnitt zweier Relationen $R \cap S$ mit r bzw. s Tupeln (Folie 38, Vorlesung 5):
Die Eingabe in den Mapper besteht aus allen Tupeln aus beiden Relationen, also sind die Kommunikationskosten hier in $\mathcal{O}(r + s)$. Da der Mapper nur die Identität ist, ist die Eingabe in den Reducer genauso groß. Also sind die Gesamtkommunikationskosten dieses Problems in $\mathcal{O}(2r + 2s) = \mathcal{O}(r + s)$.

b)

Gruppenbasierter Similarity-Join mit n Bildern und Gruppengröße h (Folie 39, Vorlesung 7):
Die Eingabe in den Mapper besteht aus Tupeln (i, P_i) mit dem Bildindex und dem Bild. Bei n Bildern sind diese Kommunikationskosten also in $\mathcal{O}(n)$. Die Eingabe in den Reducer ist höher. Es werden die n Bilder an jeweils $g - 1$ Reducer geschickt, also liegen diese Kommunikationskosten in $\mathcal{O}(n \times \frac{n}{h})$. Die Gesamtkommunikationskosten dieses Problems liegen somit in $\mathcal{O}(\frac{n^2}{h} + n)$.

Aufgabe 2: Graphische Modelle

a)

Grundsätzlich lassen sich anhand eines graphischen Modells Aussagen treffen über

- die minimal mögliche Reducergröße q:

$$q \geq \max_{n \in B} deg(n).$$

  Das heißt, man braucht mindestens so viele Eingaben in einem Reducer wie die höchste Anzahl an Eingangskanten in einen Ausgabeknoten.

- die maximal benötigte Replikationsrate r:

$$r \leq \sum_{n \in A} deg(n) = \sum_{n \in B} deg(n).$$

  Das heißt, man braucht höchstens so viele Replikationen einer einzelnen Eingabe wie Kanten insgesamt. (Noch öfter braucht eine Eingabe nicht geschickt zu werden.)

b)

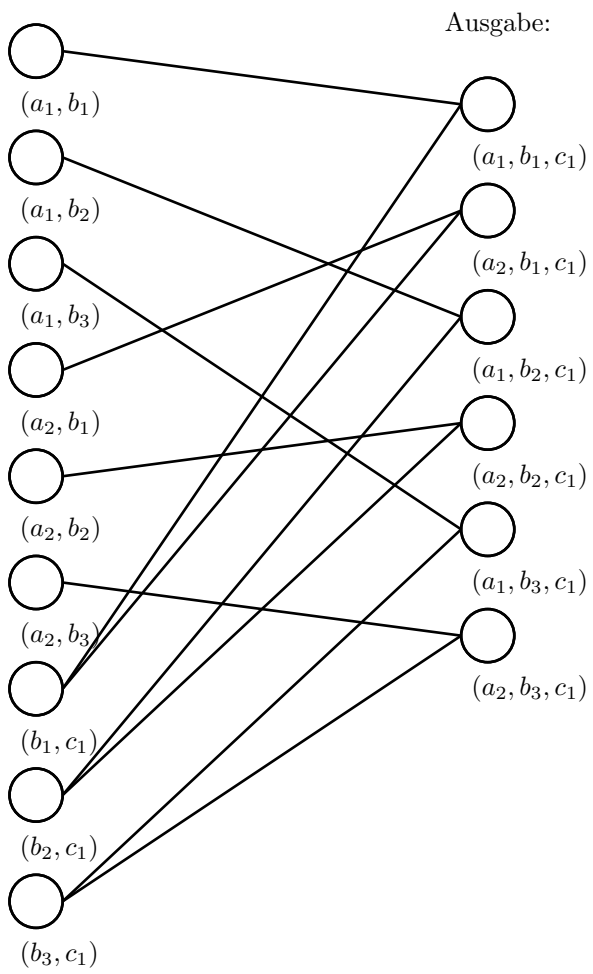
Natürlicher Join $R(A, B) \bowtie S(B, C)$ mit a möglichen Werten für A, b möglichen Werten für B und c möglichen Werten für C:

i)

Das zugehörige graphische Modell besitzt $a * b + b * c$ Eingabeknoten und $a * b * c$ Ausgabeknoten.

ii)

Modell für $a = 2$, $b = 3$ und $c = 1$:



iii)

Schranken für q und r in Abhängigkeit von a, b und c:

$$q \geq \max_{n \in B} deg(n) \Rightarrow q \geq 2.$$

Die untere Schranke für q hängt hier nicht von a, b und c ab, sondern ist hier fest ≥ 2, weil zwei Relationen gejoint werden und deshalb zwei Eingabetupel für ein mögliches Ausgabetupel benötigt werden.

$$r \leq \sum_{n \in B} deg(n) \Rightarrow r \leq 2 * a * b * c.$$

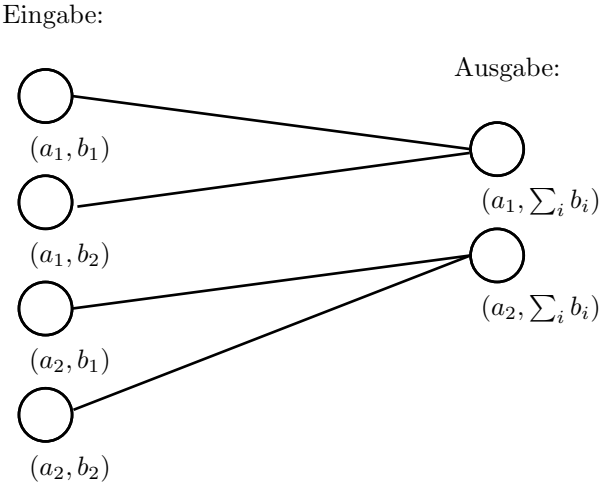
c)

Gruppierung $\gamma_{A,SUM(B)}R(A,B)$ mit a möglichen Werten für A und b möglichen Werten für B:

i)

Das zugehörige graphische Modell besitzt $a*b$ Eingabeknoten (gleich der Anzahl an Tupeln in der Relation R) und a Ausgabeknoten, da für jeden Wert von A eine Summe berechnet wird.

ii)



iii)

Schranken für q und r in Abhängigkeit von a und b:

$$q \geq \max_{n \in B} deg(n) \Rightarrow q \geq b.$$

Für jeden Wert von A gibt es bis zu b Summanden, wofür jeweils ein Tupel an den Reducer geschickt werden muss.

$$r \leq \sum_{n \in B} deg(n) \Rightarrow r \leq a * b.$$