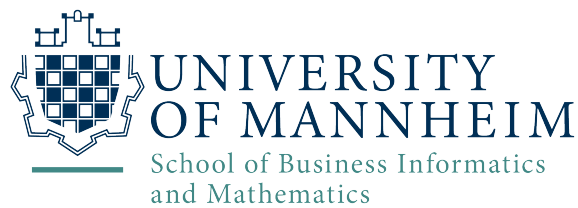
Database Systems II – Exercise #10

Exam Q&A

Daniel Flachs

Chair of Practical Computer Science III:
Database Management Systems

05/06/2019



Created June 5, 2019

Contents

Contents

Problem Statement

Consider two data streams F and G containing pairs (key, frequency), key $\in \{1, 2, 3, 4, 5\}$:

$$F = \big((1, 5), (4, -2), (1, 2), (2, 3), (3, 1), (1, -3), (3, 2), (5, 2), (4, 3)\big)$$
$$G = \big((2, 1), (4, 3), (3, 2), (1, 3), (3, -2), (1, 2), (5, -1), (1, 2), (4, -1)\big)$$

We would like to estimate the size of the join $F \bowtie G$ of the two streams using a sketch.

AGMS and FastAGMS

- In general, a sketch is given a data stream. It sees each item once and only once, i. e., it cannot go back in the stream to look something up.
- The two sketches, AGMS and FastAGMS, work by using a vector of counters that are updated in a certain way by evaluating hash functions.
- In our example, AGMS uses 3 counters, and FastAGMS $3 \cdot 3 = 9$ counters.
- All streams and vectors are 0-indexed, i. e., the first element of a vector v is $v[0]$.

AGMS Example with 3 Counters

Let x be the counter vector of length 3. The i^{th} counter ($0 \leq i \leq 2$) is denoted by $x[i]$ and initialized to 0.

When a new data stream item (e, w) arrives, $x[i]$ is updated by multiplying the frequency w with either 1 or -1, depending on the evaluation of a hash function h on e:

$$x[i] = x[i] + w \cdot h(e)[i]$$

Hash function $h : \{1, 2, 3, 4, 5\} \rightarrow \{-1, +1\}^3$

Counter	**Key Domain**				
	1	2	3	4	5
0	+1	+1	+1	−1	−1
1	+1	−1	−1	+1	+1
2	−1	+1	−1	+1	−1

$$x = \begin{pmatrix} x[0] \\ x[1] \\ x[2] \end{pmatrix} \overset{\text{Init.}}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

AGMS Example with 3 Counters

Calculations for Stream F

$$F = \big((1,5),(4,-2),(1,2),(2,3),(3,1),(1,-3),(3,2),(5,2),(4,3)\big)$$

Iter.	(e,w)	$h(e)$	$x[0]$	$x[1]$	$x[2]$
0	–	–	0	0	0
1	$(1,5)$	$(+1,+1,-1)$	5	5	-5
2	$(4,-2)$	$(-1,+1,+1)$	7	3	-7
3	$(1,2)$	$(+1,+1,-1)$	9	5	-9
4	$(2,3)$	$(+1,-1,+1)$	12	2	-6
5	$(3,1)$	$(+1,-1,-1)$	13	1	-7
6	$(1,-3)$	$(+1,+1,-1)$	10	-2	-4
7	$(3,2)$	$(+1,-1,-1)$	12	-4	-6
8	$(5,2)$	$(-1,+1,-1)$	10	-2	-8
9	$(4,3)$	$(-1,+1,+1)$	7	1	-5

Ctr.	Key Domain				
	1	2	3	4	5
0	+1	+1	+1	−1	−1
1	+1	−1	−1	+1	+1
2	−1	+1	−1	+1	−1

$$\Rightarrow x_f = \begin{pmatrix} 7 \\ 1 \\ -5 \end{pmatrix}$$

Calculation of vector x_g for stream G analogously.

AGMS Example with 3 Counters

Result Vectors

$$x_f = \begin{pmatrix} 7 \\ 1 \\ -5 \end{pmatrix} \qquad x_g = \begin{pmatrix} 7 \\ 7 \\ -3 \end{pmatrix}$$

Join Size Estimate: component-wise multiplication of x_f and x_g, then averaging the resulting vector's components

$$X \quad = x_f \cdot x_g = \begin{pmatrix} 7 \\ 1 \\ -5 \end{pmatrix} \cdot \begin{pmatrix} 7 \\ 7 \\ -3 \end{pmatrix} = \begin{pmatrix} 49 \\ 7 \\ 15 \end{pmatrix}$$

$$|F \bowtie G| \approx \frac{1}{3} \sum_{i=0}^{2} X[i] = \frac{49 + 7 + 15}{3} = \frac{71}{3} \approx 23.67$$

True Join Size: $|F \bowtie G| = 31$.

FastAGMS Example with 9 Counters

$$F = \big((1,5),(4,-2),(1,2),(2,3),(3,1),(1,-3),(3,2),(5,2),(4,3)\big)$$
$$G = \big((2,1),(4,3),(3,2),(1,3),(3,-2),(1,2),(5,-1),(1,2),(4,-1)\big)$$

Hash Functions

(Value, Row) → Ctr.

Row	Key Domain				
	1	2	3	4	5
0	1	1	3	2	3
1	1	3	2	1	2
2	3	2	3	3	1

(Value, Ctr.) → $\{+1,-1\}^3$

Ctr.	Key Domain				
	1	2	3	4	5
0	+1	+1	+1	−1	−1
1	+1	−1	−1	+1	+1
2	−1	+1	−1	+1	−1

Result

$$x_f = \begin{pmatrix} 7 & -1 & 1 \\ 5 & -1 & -3 \\ -2 & 3 & -6 \end{pmatrix} \begin{matrix} \text{Counters} \\ \\ \\ \text{Rows} \end{matrix}$$

$$x_g = \begin{pmatrix} 8 & -2 & 1 \\ 9 & -1 & -1 \\ 1 & 1 & -5 \end{pmatrix}$$

FastAGMS Example with 9 Counters

We now multiply x_f and x_g component-wise, sum up each row and compute the median of the resulting vector.

$$x_f \cdot x_g = \begin{pmatrix} 7 & -1 & 1 \\ 5 & -1 & -3 \\ -2 & 3 & -6 \end{pmatrix} \cdot \begin{pmatrix} 8 & -2 & 1 \\ 9 & -1 & -1 \\ 1 & 1 & -5 \end{pmatrix} = \begin{pmatrix} 56 & 2 & 1 \\ 45 & 1 & 3 \\ -2 & 3 & 30 \end{pmatrix} \xrightarrow[\text{sums}]{\text{row}} \begin{pmatrix} 59 \\ 49 \\ 31 \end{pmatrix}$$

Join Size Estimate: $|F \bowtie G| \approx \text{median}\{59, 49, 31\} = 49$.

Contents

Cache Misses

Q: Consider an array that is filled with random integer values. When do we have more cache misses? When summing up the array sequentially or when splitting the array and summing it up in parallel?

A: The number of cache misses is exactly the same for both variants because no matter how we iterate over the array, we need to read each value exactly once and use 100 % of each cache line we load. This can be confirmed experimentally by running a tool like `cachegrind`, cf. Exercise 2.

Calculator

Q: Will we need a calculator in the exam?

A: No, all calculations should be easily possible to do manually.

Notes from the front page of the exam

1. Check your exam for completeness.
2. Put your name and student ID number at the top of every sheet that is not connected to the personalized cover page.
3. Use permanent pens only, e. g. no pencils. Do not use red pens.
4. Solve each task on the respective page. If the space does not suffice, continue on the back of the page.
5. No auxiliary means other than writing utensils are allowed. A violation will be considered an attempt to deceive.
6. Please sign your exam below.

Exam – General Information

- **Exam Date**: 14/06/2019, 14:00–15:30, room O 135
- **Seat Numbers**: Assigned by the student services, visible one week before the exam in Portal[2].
- **Duration**: 90 min
- **Points**: 90 points to achieve at maximum
- $\frac{90\text{points}}{90\,\text{min}} = 1$ point per min $\Rightarrow$ This gives you a rough idea on how much time to spend on each task.
- The exam is passed if 45 points are achieved. This number might be lowered during the grading process, but not increased.

Important Note!

Please refer to the web page of the student services for late registration, withdrawal and especially in case of illness!

▶ https://www2.uni-mannheim.de/studienbueros/english/exams/