

Методы и системы обработки больших данных

7. Статистические характеристики данных

Виды статистик

- **Минимальные статистики:**
 $T(R)$ и все $V(R, A_i)$
- **Гистограммы** - позволяют более точно оценивать размеры промежуточных отношений по сравнению с минимальными статистиками $V(R, A_i)$

Статистические данные хранятся в словаре (справочнике) базы данных

Пример описания статистик для R в
словаре базы данных

$R(A^*, B, C)$	
$T(R)$	5
$V(R, B)$	2
$V(R, C)$	3

Словарь БД

R		
A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2

Гистограммы

- Полные
- Частичные

Полные гистограммы

Если $V(R, A)$ не велико, то для каждого значения хранится количество его вхождений в $R.A$

Пример полной гистограммы для R

$R(A^*, B, C)$	Знач-е	Кол-во вхожд-й
$T(R)$	any	10
$R.B$	3	4
$R.B$	7	6
$R.C$	2	5
$R.C$	4	3
$R.C$	9	2

Атрибут A не включен в гистограмму так как он является первичным ключом

R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9
7	3	2
8	7	4
9	7	2
10	7	2

*) Первичный ключ

Гистограммы с наиболее часто встречающимися значениями

$$R(A, B) \propto S(A, C)$$

Гистограмма для $R.A$:

Значение	Количество в $R.A$
1	200
0	150
5	100
остальные	550

Гистограмма для $S.A$:

Значение	Количество в $S.A$
0	100
1	80
2	70
остальные	250

Из тотальных статистик

$$V(R.A) = 14 \Rightarrow$$

550 кортежей R с неизвестным значением атрибута A распределены по $14 - 3 = ??$ группам (в среднем) по ?? кортежей.

$$V(S.A) = 13 \Rightarrow$$

250 кортежей S с неизвестным значением атрибута A распределены по $13 - 3 = ??$ группам (в среднем) по ?? кортежей.

Размер соединения $R \bowtie S$

Значение атрибута соединения	Количество вхождений в R	Количество вхождений в S	Количество пар в $R \bowtie S$
0	150	100	$150 \times 100 = 15\,000$
1	200	80	$200 \times 80 = 16\,000$
2	$\frac{550}{14 - 3} = 50$	70	$50 \times 70 = 3\,500$
5	100	$\frac{250}{13 - 3} = 25$	$100 \times 25 = 2\,500$
Остальные	$550 - 50 = 500$	$250 - 25 = 225$	$\frac{500 \times 225}{\max(14 - 4; 13 - 4)} = 11\,250$
Итого			48 250

В соответствии с методом из лекции 5 получим:
 $1000 \times 500 / 14 = 35\,714$ Разница: **26%**

Пересмотр статистик

1. СУБД выполняет процесс сбора статистик автоматически
 - По истечении определенного промежутка времени
 - После выполнения определенного количества обновлений базы данных
2. Пересмотр статистик производится принудительно по команде администратора базы данных

Тактика сбора статистик

- Собирать не часто (статистическая картина БД меняется медленно)
- Ограничиваться малыми выборками

Пусть отношение $\acute{R}$ составляет 1% от R . Тогда:

если $V(\acute{R}, A) \approx T(\acute{R})$, то $V(R, A) \approx T(R)$

если $V(\acute{R}, A) \ll T(\acute{R})$, то $V(R, A) \approx V(\acute{R}, A)$

- Инкрементный подход
 - При добавлении кортежа $T(R) = T(R) + 1$
 - При удалении кортежа $T(R) = T(R) - 1$
 - $V(R, A)$ пересчитывается по индексу A при добавлении и удалении кортежей
 - Если A – ключ, то $V(R, A) = T(A)$
 - Рудиментарные индексы (не содержат указателей)

Конец лекции 7