DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 2

13 de Junio de 2014 — Base de Datos

Bobby Tables

Integrante	LU	Correo electrónico
Mancuso Emiliano	597/07	emiliano.mancuso@gmail.com
Mataloni Alejandro	706/07	amataloni@gmail.com
Gauder María Lara	027/10	marialaraa@gmail.com
Reartes Marisol	422/10	marisol.r5@hotmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

0. Classic Histogram and Distribution Steps

0.1. Introducción

Tanto Classic Histogram como Distrbution Steps son estimadores de tuplas que satisfacen una condicion. Se puede realizar una búsqueda por aquellas tuplas que cumplan una condición de igualdad, por menor o menor igual y por mayor o mayor igual a cierto valor. Los estimadores son utilizados en las bases de datos en el proceso de seleccionar un plan de ejecucion óptimo al momento de realizar una consulta.

Son varios los factores que se involucran en el cálculo de estimadores. Alguno de ellos son la precisión, la cantidad de información que contenga la base de datos, los errores en la estimación, el espacio que consume el estimador y la estructura para consultar la información requerida. Cada uno de estos factores se ajustan al modificar algunas variables que son pasadas por parámetros a los algoritmos de los estimadores.

Las variables en cuestión son:

- La tabla a la que se desea realizar las consultas,
- Una columna
- Una variable denominada PARAM que representan la cantidad en la que se va a agrupar los datos de la base, es decir, la cantidad de steps.

Los steps denotan la cantidad en la que se dividen los conjuntos de datos. Por lo tanto, es claro que al aumentar el valor de steps se aumenta la presición y se disminuye el error en la estimación.

El beneficio que aporta el estimador Distrbution Steps podrá ser observado al compararlo con Classic Histogram y evaluando diferentes factores. Principalmente se tiene en cuenta el costo temporal y espacial de la construcción de las estructuras requeridas por el estimador, el costo temporal de la consulta y el error que comete en la misma.

0.2. Análisis

Para entender el funcionamiento de ambos estimadores, se creó el siguiente caso simple de prueba.

[[EXPLICAR EL EJEMPLO CON PALABRAS Y QUE REPRESENTA]]

![Imagen Distribution Steps](url)

![Imagen Classic Histogram](url)

Basándose en el ejemplo planteado, se observa que el comportamiento de los distintos estimadores depende de la distribición de los datos. Por ejemplo, cuando los datos están agrupados, Distribucion Steps brinda un menor error en la estimación pero, cuando los datos estan dispersos el error es del estimador es mucho mayor que el que presenta Classic Histogram.

A continuación se experimentará con datos que siguen una distribución uniforme o normal.

0.2.1. Generación casos de prueba

Los casos de prueba que se van a utilizar para el análisis del comportamiento de ambos estimadores se generarán de manera aleatoria. Se define una función que genera números aleatorios que siguen una

distribucion normal o uniforme.

Los dos estimadores serán comparados utilizando la misma cantidad de bins, ya que en ambos representan lo mismo. Uno va a tener bins más altos y el otro, más anchos. Esto se debe a la particularidad de como se construyen los histogramas.

0.2.2. Factores

Debido a la estructura que se utiliza para construir los estimadores y el algoritmo de construcción de los mismos, se podrá afirmar que:

Histogram

- Costo de creación de estimador:

$$O(b * n + n) \tag{1}$$

$$O(b * n) \tag{2}$$

  Siendo n la cantidad total de tuplas existentes en la base de datos. Por el otro lado, b representa el valor pasado en PARAM. Como se recorre toda la tabla por cada dato que se requiere para la tabla del estimador, se deberá considerar un costo de $b * n$. Por otro lado, también se requiere almacenar en la estructura del estimador, el máximo, mínimo y la cantidad total de tuplas, lo cual es una búsqueda que realiza el motor de bases de datos en tiempo lineal.

- Costo espacial del estimador:

$$O(2 * b + 3) \tag{3}$$

$$O(2 * b) \tag{4}$$

  Se utiliza un diccionario que presenta un costo espacial del valor pasado en PARAM por dos, ya que por cada bin se almacen dos valores más (la cantidad y el acumulado hasta ese valor). Por otro lado, se requere almacenar el máximo, mínimo y la cantidad total.

- Costo de consulta:

$$O(1) \tag{5}$$

  Lo único se se requere es el acceso al diccionario, lo cual presenta un costo constante.

Distribution Steps

- Costo de creación del estimador:

$$O(n * log(n) + n + n) \tag{6}$$

$$O(n * log(n)) \tag{7}$$

  El algoritmo consiste en el ordenamiento de las tuplas de acuerdo a la columna pasada por parámetro. El ordanamiento lo realiza el motor de la base de datos en costo $O(n * log(n))$. Por otro lado se deberá calcular el máximo, mínimo y total de tuplas existentes en la base.

- Costo espacial del estimador:

$$O(b + 1) \tag{8}$$

$$O(b) \tag{9}$$

  Se requiere unicamente almacenar un array con las tuplas que se encuentran en las distintas posiciones de los steps. Además, se deberá almacenar la cantidad de tuplas totales existentes en la tabla pasada por parámetro.

- Costo de consulta:

$$O(n) \tag{10}$$

  Tanto para una búsqueda por igualdad o por menor, se deberá recorrer el array buscando el valor de tupla deseado.

0.2.3. Distribución Uniforme

Partimos con una base de datos cuya información cumple con una distribución uniforme.

![Grafico Distribution Steps](url)

![Grafico Classic Histogram](url)

Con respecto al error de factor en la estimacion, ambos estimadores no difieren significativamente en sus resultados. Dado que al tener los datos distribuidos de forma uniforme, la probabilidad de cada tupla es $1/n$, siendo n la cantidad total de tuplas de la tabla.

(Para rango —a-b—/(max - min))

Debido a que se los compara con igual cantidad de steps, entonces la altura, en el caso de Classic Histogram, y el ancho, en el Distribution Steps, son similares.

Sin embargo, teniendo en cuenta el factor de costos de creación, el estimador "Distribution Steps."es más costoso. El motivo principal es debido al ordenamiento de los datos antes de construir el histograma requerido.

En conclusión, para datos que cumplen una distribución uniforme, sugerimos implementar Classic Histogram, debido al menor costo de construcción del histograma.

0.2.4. Distribución Normal

Partimos con una base de datos que la información sigue una distribución normal.

![Grafico Distribution Steps](url)

![Grafico Classic Histogram](url)