

UNIDAD 4: ALGORITMOS PARA FLUJOS DE DATOS

Gibran Fuentes-Pineda

Diapositivas basadas en las de la M. en C. Blanca Vázquez

Mayo 2021

- Generalización del problema del conteo de elementos distintos.
- Objetivo: estimar los momentos en un flujo de datos, lo cual se obtiene mediante la distribución de frecuencias de los diferentes elementos.

- Sea m_e es el número de ocurrencias del elemento e en el flujo, el momento i -ésimo está definido por

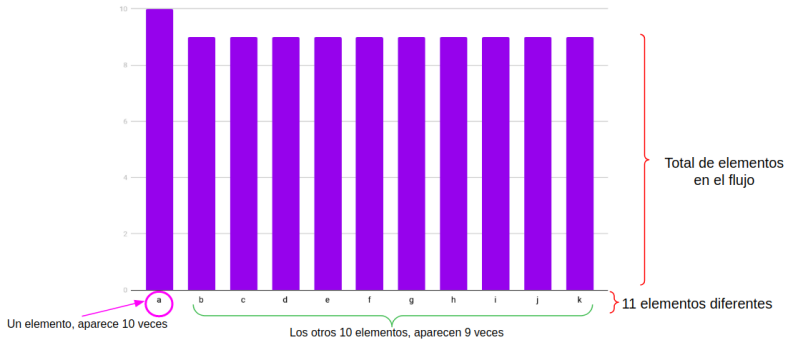
$$\sum_{e \in \mathbb{U}} (m_e)^i$$

donde \mathbb{U} es el conjunto universal.

- El momento 0 es el número de elementos distintos
- El momento 1 es la suma de m_i , es decir, el tamaño del flujo de datos
- El momento 2 es la suma de los cuadrados de m_i , también conocido como número sorpresa

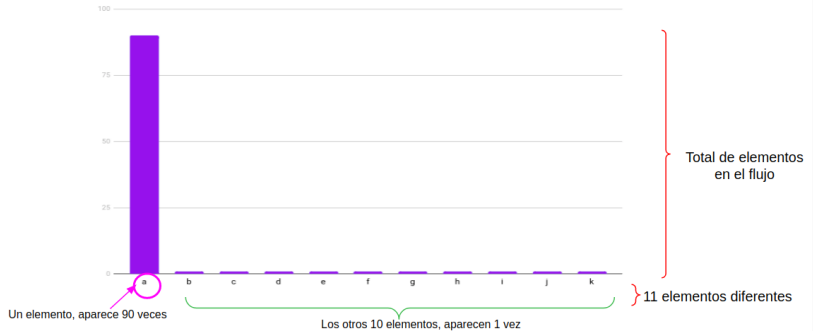
EJEMPLO (1)

Distribución uniforme de los elementos en el flujo de datos



EJEMPLO (2)

Distribución uniforme de los elementos en el flujo de datos



ALGORITMO DE ALON-MATIAS-SZEGEDY (AMS)

- Algoritmo para el cálculo de momentos en flujos de datos, definido por Noga Alon, Yossi Matias y Mario Szegedy.
- Se enfoca en aproximar la suma de las entradas al cuadrado de un vector definido por un flujo de datos.
- Permite calcular cualquier momento aún si no es posible almacenar todas las cuentas m_i de todos los elementos.

ALGORITMO DE AMS PARA ESTIMAR EL SEGUNDO MOMENTO

- Dado un flujo de tamaño n constante, se toman K variables X_1, X_2, \dots, X_K seleccionando K posiciones en el flujo de forma aleatoria y uniforme.
- Las variables almacenan el elemento de la posición correspondiente en $X_k.elemento$ y un valor entero $X_k.valor$, el cual se inicializa con 1 y se incrementa en 1 cada vez que se encuentra una ocurrencia de $X_k.elemento$
- El segundo momento de cualquier variable X_k se estima con $n \cdot (2 \cdot X_k.valor - 1)$.

EJEMPLO: CÁLCULO DEL 2DO MOMENTO

- Considera el flujo $a, b, c, b, d, a, c, d, a, b, d, c, a, a, b$

Elemento (e)	Número de ocurrencias (m_e)
a	5
b	4
c	3
d	3

- El primero momento es $5 + 4 + 3 + 3 = 15$
- El segundo momento es $5^2 + 4^2 + 3^2 + 3^2 = 59$

EJEMPLO: SEGUNDO MOMENTO CON EL ALGORITMO AMS (1)

1. Seleccionamos 3 variables aleatorias: X_1, X_2 y X_3
2. Supongamos que las posiciones para las 3 variables son 3, 8 y 13.
 - En la posición 3, encontramos el elemento c , al cual llamamos: $X_1.elemento = c$
 - En la posición 8, encontramos el elemento d , al cual llamamos: $X_2.elemento = d$
 - En la posición 13, encontramos el elemento a , al cual llamamos: $X_3.elemento = a$

$a, b, \overset{3}{\underbrace{c}}, b, d, a, c, \overset{8}{\underbrace{d}}, a, b, d, c, \overset{13}{\underbrace{a}}, a, b$

EJEMPLO: SEGUNDO MOMENTO CON EL ALGORITMO AMS (2)

$a, b, \overset{3}{\underbrace{c}}, b, d, a, c, \overset{8}{\underbrace{d}}, a, b, d, c, \overset{13}{\underbrace{a}}, a, b$

$X_1.elemento = c$	$X_1.valor = 1$	$X_1.valor = 2$	$X_1.valor = 3$
$X_2.elemento = d$	$X_2.valor = 1$	$X_2.valor = 2$	
$X_3.elemento = a$	$X_3.valor = 1$	$X_3.valor = 2$	

El valor final es $X_1.valor = 3$, $X_2.valor = 2$ y $X_3.valor = 2$

EJEMPLO: SEGUNDO MOMENTO CON EL ALGORITMO AMS (3)

$a, b, \overbrace{c}^3, b, d, a, c, \overbrace{d}^8, a, b, d, c, \overbrace{a}^{13}, a, b$

- Estimamos el segundo momento con $n \cdot (2 \cdot X_k.\text{valor} - 1)$
 - Para X_1 : $15 \cdot (2 \cdot 3 - 1) = 75$
 - Para X_2 : $15 \cdot (2 \cdot 2 - 1) = 45$
 - Para X_3 : $15 \cdot (2 \cdot 2 - 1) = 45$
- Promediando las estimaciones de cada variable tenemos $(75 + 45 + 45)/3 = 55$

- El valor esperado de cualquier variable es el segundo momento del flujo del que fue generada
- Sea $e(i)$ el elemento que aparece en la posición i en el flujo y $c(i)$ el número de veces que aparece este elemento de la posición i a la n , el valor esperado del estimador del segundo momento es

$$E[n \cdot (2 \cdot X_k \cdot \text{valor} - 1)] = \frac{1}{n} \sum_{i=1}^n n \cdot (2 \cdot c(i) - 1) = \sum_{i=1}^n (2 \cdot c(i) - 1)$$

ANÁLISIS DEL ALGORITMO AMS (2)

- Podemos el valor esperado de otra forma si agrupamos los términos de todas las posiciones que tienen el mismo elemento
- Para el ejemplo anterior, la letra a aparece m_a veces. Si tomamos los términos de la última posición hacia la primera tendríamos

$$2 \cdot 1 - 1 = 1$$

$$2 \cdot 2 - 1 = 3$$

$$2 \cdot 3 - 1 = 5$$

$$\vdots$$

$$2 \cdot m_a - 1$$

- Por lo tanto, podemos reescribir el valor esperado para cada elemento $e \in \mathbb{U}$ de la siguiente manera

$$\mathbb{E}[n \cdot (2 \cdot X_k.\text{valor} - 1)] = \sum_{e \in \mathbb{U}} [1 + 3 + 5 + \dots + (2 \cdot m_e - 1)]$$

- Dado que $[1 + 3 + 5 + \dots + (2 \cdot m_e - 1)] = (m_e)^2$

$$\mathbb{E}[n \cdot (2 \cdot X_k.\text{valor} - 1)] = \sum_e (m_e)^2$$

- Para el estimador del segundo momento tenemos

$$n \cdot (2 \cdot X_k.\text{valor} - 1) = n \cdot (X_k.\text{valor}^2 - (X_k.\text{valor} - 1)^2)$$

- El estimador del tercer momento es

$$n \cdot (X_k.\text{valor}^3 - (X_k.\text{valor} - 1)^3) = n \cdot (3 \cdot X_k.\text{valor}^2 - 3 \cdot X_k.\text{valor} + 1)$$

- En general, el estimador del i -ésimo momento es

$$n \cdot (X_k.\text{valor}^i - (X_k.\text{valor} - 1)^i)$$

ESTIMACIÓN PARA FLUJOS INFINITOS (1)

- Hasta ahora hemos considerado que n es constante, sin embargo, en la práctica no lo es
- ¿Cómo seleccionamos las posiciones para la variables?
 - Si se hace al inicio y no se actualiza, conforme el flujo crezca habría un sesgo hacia las primeras posiciones y la estimación sería demasiado grande
 - Si se espera demasiado y no mantenemos suficientes variables en las posiciones al principio del flujo, podemos tener una estimación poco confiable

- Estrategia de selección de posiciones
 1. Se toman las primeras s posiciones del flujo como variables.
 2. Se elige la posición $n > s$ con probabilidad $\frac{s}{n}$
 - Si es elegida, se selecciona de forma aleatoria y uniforme una de las s variables y se reemplaza por la de la posición n
 - En caso contrario se mantienen las posiciones de las s variables