

UNIDAD 2: MINERÍA DE ELEMENTOS FRECUENTES

ALGORITMOS DE MEMORIA PRINCIPAL

Blanca Vázquez y Gibran Fuentes-Pineda

6 de septiembre de 2021

- El algoritmo apriori es práctico cuando las estructuras de datos para contar caben en memoria principal
- Si es necesario leer constantemente de disco se vuelve muy lento
- Estrategias para reducir el número de candidatos
 - Algoritmo de Park, Chen y Yu (PCY)
 - Algoritmo multietapa
 - Algoritmo multihash

ALGORITMO DE PCY (1)

- Usa una tabla de dispersión adicional para excluir pares poco frecuentes, similar a un filtro de Bloom
- Pares de elementos se mapean a índices de cubetas que almacenan contadores
- Un par candidato $\{i, j\}$ es aquel cuyos elementos i y j son frecuentes y cuyo entero asociado en la tabla de dispersión es mayor a un umbral (es decir, es frecuente)
- Reduce el número de pares candidatos a considerar

ALGORITMO DE PCY (2)

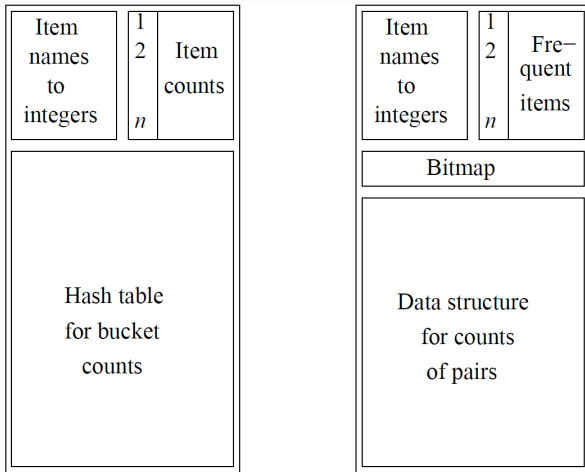
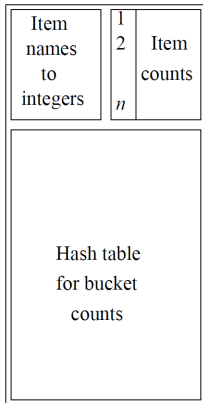


Imagen tomada de Leskovec et al. Mining of Massive Datasets, 2nd edition, 2014

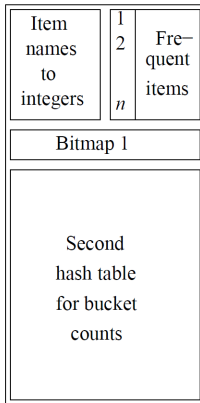
ALGORITMO MULTIETAPA (1)

- Usa varias tablas de dispersión sucesivas para reducir aún más los pares candidatos
- Tablas subsecuentes solo consideran los pares que ocurrieron en una cubeta frecuente en las tablas precedentes
- Condiciones para que un par sea candidato
 1. i y j son elementos frecuentes en \mathcal{F}_1
 2. $\{i, j\}$ tienen una cubeta frecuente asociada en la primera tabla de dispersión
 3. $\{i, j\}$ tienen una cubeta frecuente asociada en la segunda tabla de dispersión

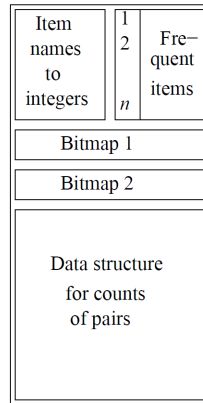
ALGORITMO MULTITETAPA (2)



Pass 1



Pass 2



Pass 3

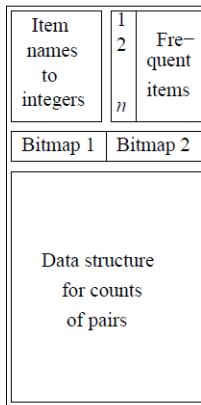
Imagen tomada de Leskovec et al. Mining of Massive Datasets, 2nd edition, 2014

- Realiza el filtrado sin necesidad de realizar más pasadas
- Usa múltiples tablas de dispersión con memoria compartida en una sola pasada
- Si usamos demasiadas tablas es posible que muchas cubetas sean frecuentes

ALGORITMO MULTIHASH (2)



Pass 1



Pass 2

Imagem tomada de Leskovec et al. Mining of Massive Datasets, 2nd edition, 2014

EJEMPLO PCY

- Búsqueda de conjuntos con soporte mínimo de 0.3
- Con una tabla de dispersión de 11 cubetas, donde cada par $\{i, j\}$ se almacena en la cubeta $i \times j \bmod 11$

ID	Transacción
1	{1, 2, 3}
2	{1, 3, 5}
3	{3, 5, 6}
4	{2, 3, 4}
5	{2, 4, 6}
6	{1, 2, 4}
7	{3, 4, 5}
8	{1, 3, 4}
9	{2, 3, 5}
10	{4, 5, 6}
11	{2, 4, 5}
12	{3, 4, 6}

EJEMPLO MULTIETAPA

- Búsqueda de conjuntos con soporte mínimo de 0.3
- Primera tabla con 11 cubetas: $i \times j \text{ mód } 11$
- Segunda tabla con 9 cubetas: $i + j \text{ mód } 9$

ID	Transacción
1	{1, 2, 3}
2	{1, 3, 5}
3	{3, 5, 6}
4	{2, 3, 4}
5	{2, 4, 6}
6	{1, 2, 4}
7	{3, 4, 5}
8	{1, 3, 4}
9	{2, 3, 5}
10	{4, 5, 6}
11	{2, 4, 5}
12	{3, 4, 6}

EJEMPLO MULTIHASH

- Búsqueda de conjuntos con soporte mínimo de 0.3
- Primera tabla con 5 cubetas: $2i + 3j + 4 \pmod{5}$
- Segunda tabla con 5 cubetas: $i + 4j \pmod{5}$

ID	Transacción
1	{1, 2, 3}
2	{1, 3, 5}
3	{3, 5, 6}
4	{2, 3, 4}
5	{2, 4, 6}
6	{1, 2, 4}
7	{3, 4, 5}
8	{1, 3, 4}
9	{2, 3, 5}
10	{4, 5, 6}
11	{2, 4, 5}
12	{3, 4, 6}