**Database Systems: The Complete Book**

# Chapter 22

## Section 1

### 1

- a 25%
- b 0%
- c 33%
- d 40%
- e 50%
- f {milk, beer}
  {milk, pepsi}
  {beer, pepsi}
- g {milk, beer}
- h {pepsi} => milk
  has confidence 75%

## Section 2

- 1 $F_1$ = {milk, coke, beer, pepsi, juice}
  $F_2$ = {milk, beer}, {milk, pepsi}, {beer, pepsi}
  $F_3$ = { }

### 2

- a a[n], where n = $i_1 + n*i_2 + 2*n*i_3 + 3*n*i_4 + ...$
- b $n^2$
- c Hash the tuple $(i_1, i_2, ... , i_n)$ to locate the entry to store the count.
- d pn

### 3

- a $(1/10)^{(sb)} + (1/100)^{(sb)}$
- b $(1/10)^{2(sb)} + (1/1000)^{(sb)} + (1/10000)^{(sb)}$

### 4

- a $(1/10)^{(sb)} + (1/100)^{(sb)}$
- b $(1/10)^{2(sb)} + (1/1000)^{(sb)} + (1/10000)^{(sb)}$

### 5

- a $(1/10)^{2(sb)} + (1/1000)^{(sb)} + (1/10000)^{(sb)}$
- b $(1/1000)^{(sb)} + (1/10000)^{(sb)} + (1/10000000)^{(sb)} + (1/1000000)^{(sb)}$

### 6

- a Map: Each processor counts all local items.
  Reduce: Each processor is assigned to count a single item.
- b Map: Each processor counts all local items.

Reduce: Each processor is assigned to count pairs for a bucket.

▼❑**Section 3**

- ❑1 {1, 2, 3, 4, 5} {1, 6, 7} => 2/7
  {1, 6, 7} {2, 4, 6, 8} => 1/6
  {1, 2, 3, 4, 5} {2, 4, 6, 8} => 1/7
- ❑2 ‘abc ‘
  ‘bc d’
  ‘c de’
  ‘ def’
  ‘def ‘
  ‘ef g’
  ‘f gh’
  ‘ ghi’

▼❑3

- ❑a {3, 3, 9}
- ❑b {2, 5, 0}
- ❑c {2, 7, 4}

| | Estimated | Jaccard similarity |
|---|---|---|
| {a,b} | 1/5 | 1/6 |
| {b,c} | 1/5 | 2/5 |
| {a,c} | 0 | 1/5 |

▼❑4

- ❑a {3, 6, 3}
- ❑b {7, 5, 5}
- ❑c {7, 5, 5}

| | Similarity | Jaccard similarity |
|---|---|---|
| {a, b} | 0 | 1/5 |
| {b, c} | 1 | 2/5 |
| {a, c} | 0 | 1/5 |

- ❑

▼❑5

- ❑a Map: Compute minhash for each row.
  Reduce: Process results locally.
- ❑b Map: Each processor computer minhash-so-far. Do until all data has been processed.

▼❑**Section 4**

▼❑1

- ❑a 890,000
- ❑b 70%
- ❑c 30%

▼❑2

| b | r | s |
|---|---|---|
| 24 | 1 | 1 |
| 12 | 2 | .97 |
| 8 | 3 | .67 |
| 6 | 4 | .32 |
| 4 | 6 | .06 |
| 3 | 8 | .01 |
| 2 | 12 | 0 |
| 1 | 24 | 0 |

- ❑a

| b | r | s |
|---|---|---|
| 24 | 1 | .03 |
| 12 | 2 | .24 |
| 8 | 3 | .44 |
| 6 | 4 | .57 |
| 4 | 6 | .73 |
| 3 | 8 | .82 |
| 2 | 12 | .90 |
| 1 | 24 | .97 |

- ❑b

▼❑**Section 5**

▼❑1

| | A | B | C | D |
|---|---|---|---|---|
| B | 4 | 3 | 5 | 4 |
| C | 4 | 8 | 3 | 2 |
| D | 5 | 6 | 5 | |
| E | 7 | 8 | | |
| F | 5 | | | |

- ❑a

| | A | B | C | D |
|---|---|---|---|---|
| B | 3 | 2 | 3 | 3 |
| C | 2 | 4 | 3 | 1 |
| D | 4 | 5 | 3 | |
| E | 5 | 5 | | |
| F | 5 | | | |

- ❑b

- ❑2 1. d(x,y) >= 0 for all x,y
  summation of positive numbers are always postive

  2. d(x,y) = 0 if x=y
  if (x,y) = 0, then every xi - yi = 0 => xi = yi

  3. d(x,y) = d(y,x)
  |x-y| = |y-x|

  4. d(x,y) <= d(x,z) + d(z,y)

  $(|x_1 - y_1|^r + ... + |x_n - y_n|^r)^{1/r} <= (|x_1 - z_1|^r + ... + |x_n - z_n|^r)^{1/r} + (|z_1 - y_1|^r + ... + |z_n - y_n|^r)^{1/r}$
  Let $z_n = y_n + a_n$.
  $(|x_1 - y_1|^r + ... + |x_n - y_n|^r)^{1/r} <= (|x_1 - y_1 - a_1|^r + ... + |x_n - y_n - a_n|^r)^{1/r} + ...$
  Notice that $|x_i - y_i|^r < |x_i - y_i - a_i|^r$
  Therefore, 4 is true.

▼❑3

| | A | B | C | DF | E |
|---|---|---|---|---|---|
| E | 5.39 | 5.10 | 3.00 | 2.55 | |
| DF | 4.53 | 5.70 | 3.54 | | |
| C | 2.83 | 2.24 | | | |
| B | 3.00 | | | | |
| A | | | | | |

| | A | BC | DF | E |
|---|---|---|---|---|
| E | 5.39 | 4.03 | 2.55 | |
| DF | 4.53 | 4.61 | | |
| BC | 2.69 | | | |
| A | | | | |

| | A | BC | DFE |
|---|---|---|---|
| DFE | 4.71 | 4.30 | |
| BC | 2.69 | | |
| A | | | |

- ❑a

  Clusters: {ABC, DFE}

| | A | BF | C | D | E |
|---|---|---|---|---|---|
| E | 5.39 | 5.10 | 3.00 | 3.16 | |
| D | 4.12 | 5.66 | 3.61 | | |
| C | 2.83 | 3.61 | | | |
| BF | 4.00 | | | | |
| A | | | | | |

| | A | BFD | C | E |
|---|---|---|---|---|
| E | 5.39 | 5.10 | 3.00 | |
| C | 2.83 | 3.61 | | |
| BFD | 4.12 | | | |
| A | | | | |

| | AE | BFD | C |
|---|---|---|---|
| C | 3.61 | 3.61 | |
| BFD | 5.10 | | |
| AE | | | |

- ❑b

  Clusters: {BFDAE, C}

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| F | 5 | 5 | 3 | 1 | 2 | |
| E | 5 | 5 | 3 | 3 | | |
| D | 4 | 4 | 3 | | | |
| C | 2 | 2 | | | | |
| B | 3 | | | | | |
| A | | | | | | |

| | AE | B | C | D | F |
|---|---|---|---|---|---|
| F | 5 | 5 | 3 | 1 | |
| D | 4 | 4 | 3 | | |
| C | 3 | 2 | | | |
| B | 5 | | | | |
| AE | | | | | |

| | AEB | C | D | F |
|---|---|---|---|---|
| F | 5 | 3 | 1 | |
| D | 4 | 3 | | |
| C | 3 | | | |
| AEB | | | | |

| | AEBF | C | D |
|---|---|---|---|
| D | 4 | 3 | |
| C | 3 | | |
| AEBF | | | |

- ❑4
  Clusters: {AEBFC, D}

▼❑5

- ❑a B, F, A
- ❑b C, F, A

- ❑6 N = 4
  $SUM_i$ = (6, 12)
  $SUMSQ_i$ = (14, 46)

▼❑7

- ❑a 2.12
- ❑b 3.20