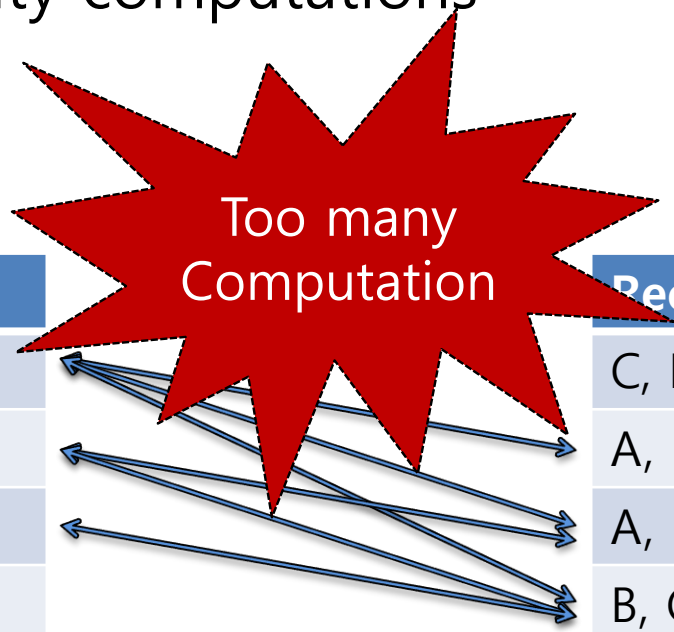
Similarity Join

Younghoon Kim
(nongaussian@hanyang.ac.kr)

SET SIMILARITY JOIN USING MAPREDUCE

A Traditional Brute-force Algorithm

- Enumerate every pair of records and compute their similarities
- Expensive for large datasets
  - $O(n^2)$ similarity computations



Record
C, D, F
A, B, E, F, G
A, B, C, D, E
B, C, D, E, F

Record
C, D, F
A, B, E, F, G
A, B, C, D, E
B, C, D, E, F

Similarity Joins using Inverted Lists

- Make an inverted lists for all items in set data
- Generate candidates by considering every pair of record IDs in each inverted list
- Find similar pairs by verifying each candidate
  - Relationship between Jaccard and Overlap similarity measures

$$\text{Jaccard}(x, y) \geq \sigma \Leftrightarrow \text{Overlap}(x, y) \geq \sigma /(1+\sigma) \cdot (|x| + |y|) = \alpha$$

    - We call α the overlap threshold
  - Check overlap(x,y) ≥ α instead of Jaccard(x,y) ≥ σ

Building Inverted Lists

- With each record in a map function
  - Output the identifier of the record (RID) to generate the inverted lists in reduce functions

RID	Items
R_1	C, D, F
R_2	A, B, E, F, G
R_3	A, B, C, D, E
R_4	B, C, D, E, F
R_5	A, E, G

Item	RIDs
A	R_2
B	R_2
C	R_1
D	R_1
E	R_2
F	R_1, R_2
G	R_2

Building Inverted Lists

- With each record in a map function
  - Output the identifier of the record (RID) to generate the inverted lists in reduce functions

RID	Items
R_1	C, D, F
R_2	A, B, E, F, G
R_3	A, B, C, D, E
R_4	B, C, D, E, F
R_5	A, E, G

Item	RIDs
A	R_2, R_3, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_3, R_4
E	R_2, R_3, R_4, R_5
F	R_1, R_2, R_4
G	R_2, R_5

Generating Candidates

- Generate candidates by making every RID pair in the each inverted list entry
  - Increase the overlap of the candidate pair

Item	RIDs
A	R_2, R_3, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_3, R_4
E	R_2, R_3, R_4, R_5
F	R_1, R_2, R_4
G	R_2, R_5

Candidate pair	Overlap
(R_2, R_3)	1
(R_3, R_5)	1
(R_2, R_5)	1

Generating Candidates

- Generate candidates by making every RID pair in the each inverted list entry
  - Increase the overlap of the candidate pair

Item	RIDs
A	R_2, R_3, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_3, R_4
E	R_2, R_3, R_4, R_5
F	R_1, R_2, R_4
G	R_2, R_5

Candidate pair	Overlap
(R_2, R_3)	2
(R_3, R_5)	1
(R_2, R_5)	1
(R_3, R_4)	1
(R_2, R_4)	1

Generating Candidates

- Generate candidates by making every RID pair in the each inverted list entry
  - Increase the overlap of the candidate pair

Item	RIDs
A	R_2, R_3, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_3, R_4
E	R_2, R_3, R_4, R_5
F	R_1, R_2, R_4
G	R_2, R_5

Candidate pair	Overlap
(R_2, R_3)	3
(R_3, R_5)	2
(R_2, R_5)	3
(R_3, R_4)	4
(R_2, R_4)	3
(R_1, R_3)	2
(R_1, R_4)	3

Candidate pair	Overlap
(R_4, R_5)	1
(R_1, R_2)	1

Finding Similar Pairs

Jaccard coefficient threshold $\sigma = 0.6$

Recall $\text{Jaccard}(x, y) \geq \sigma \Leftrightarrow O(x, y) \geq \alpha = \sigma/(1+\sigma)(|x| + |y|)$

Substitute σ values

We need the size of each record

Candidate pair	Overlap	Overlap threshold α
(R_2, R_3)	3	3.75
(R_3, R_5)	2	
(R_2, R_5)	3	
(R_3, R_4)	4	
(R_2, R_4)	3	
(R_1, R_3)	2	
(R_1, R_4)	3	
(R_4, R_5)	1	
(R_1, R_2)	1	

RID	Size
R_1	3
R_2	5
R_3	5
R_4	5
R_5	3

Calculate each record size

Verifying Candidates

Recall Jaccard(x, y) ≥ σ ⇔ O(x, y) ≥ $\alpha = \sigma/(1+\sigma)(|x| + |y|)$

Candidate pair	Overlap	Overlap threshold α
(R_2, R_3)	3	3.75
(R_3, R_5)	2	3
(R_2, R_5)	3	3
(R_3, R_4)	4	3.75
(R_2, R_4)	3	3.75
(R_1, R_3)	2	3
(R_1, R_4)	3	3
(R_4, R_5)	1	3
(R_1, R_2)	1	3.75

Overlap is smaller than the overlap threshold α
⇒ Not a similar pair

Similar pair
(R_2, R_5)
(R_3, R_4)
(R_1, R_4)

Pseudocode of Map/Reduce Functions

- Spark is more suitable for implementing this algorithm!
- RDD
  - .flatMap(lambda s: /* list of (item, set id) */)
  - .groupByKey().mapValues(lambda x: list(x))
  - .flatMap(lambda inv: /* list of (set id pair, 1) */)
  - .reduceByKey(lambda x,y: x+y)
  - .filter(lambda t: /* if count >= alpha(x,y) */

Trouble

- If there exists a long inverted list,
  - It generates too many candidates
  - ➔ Any trick to reduce the size of inverted list?

Item	RIDs
A	R_1, R_2, R_3, R_4, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_3, R_4
E	R_2, R_3, R_4, R_5
F	R_1, R_2, R_4
G	R_2, R_5

Candidate pair	Overlap
(R_1, R_2)	1
(R_1, R_3)	1
(R_1, R_4)	1
(R_1, R_5)	1
(R_2, R_3)	1
(R_2, R_4)	1
(R_2, R_5)	1

⋮

Problem 1.

- Given
  - facebook_combinied.txt
    - ID1, ID2: ID1과 ID2는 친구
    - Undirected graph
  - Similarity threshold: τ
    - 0.9, 0.8, 0.7, 0.6
- Goal
  - 아직 친구가 아닌 모든 사용자 쌍 중에서
  - 공통되는 친구의 비율이 τ 이상인 아이디의 쌍을 모두 찾으시오.
  - 즉,
    - $F(u)$: u의 친구 집합
    - $\forall(u,v) \in U \times U, s.t. u \notin F(v) \wedge v \notin F(u) \wedge \frac{|F(u) \cap f(v)|}{|F(u) \cup F(v)|} \geq \tau$

[심화학습] SET SIMILARITY JOIN: CANDIDATE REDUCTION

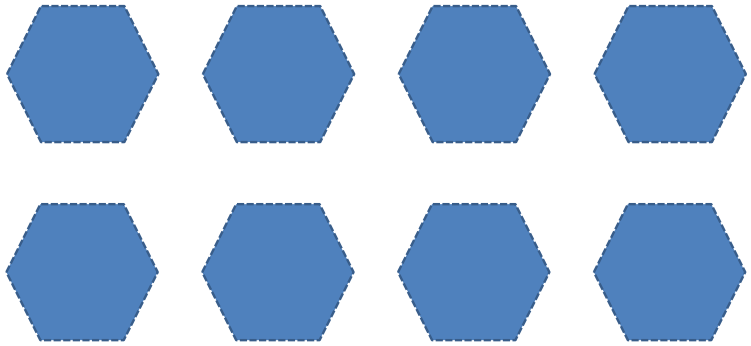
Reference

- A primitive operator for similarity joins in data cleaning
  - S. Chaudhuri, V. Ganti, and R. Kaushik
  - In ICDE, 2006.
- Efficient Similarity Joins for Near Duplicate Detection
  - Chuan Xiao, et al.
  - In WWW, 2008

Set Similarity Candidate Reduction

- A sorted vocabulary

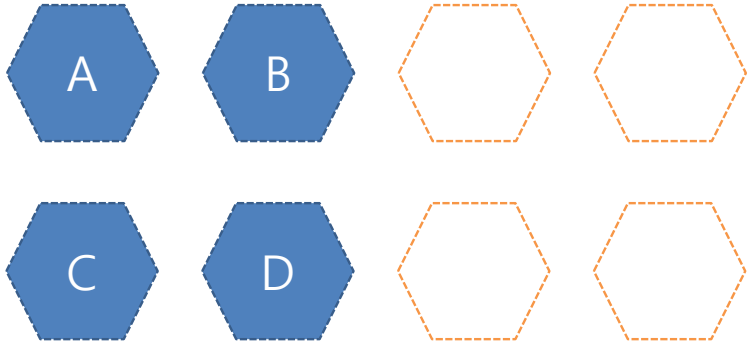
- Examine if two sets of size 4 are similar than a given threshold
  - X
  - Y



Set Similarity Candidate Reduction

- A sorted vocabulary

- Examine if two sets of size 4 are similar than a given threshold
  - X
  - Y

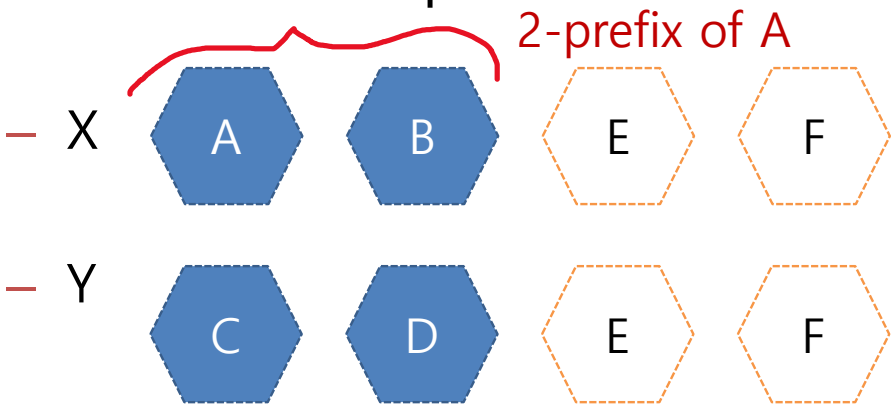


What if the first two elements do not overlap at all?

Set Similarity Candidate Reduction

- A sorted vocabulary

- Examine if two sets of size 4 are similar than 2 in terms of overlap size α



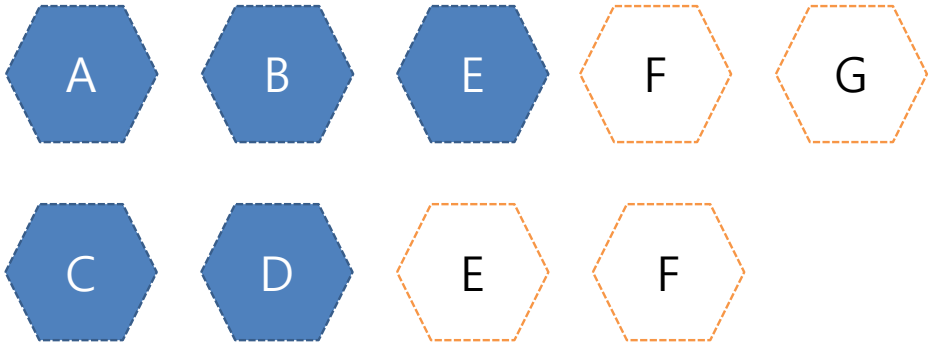
Even if the following two elements are all identical, they cannot share more than 2 common elements

$$|A \cap B| \leq 2$$

Set Similarity Candidate Reduction

- A sorted vocabulary

- Examine if two sets of size 4 are similar than 0.6 in terms of Jaccard coef.
  - X
  - Y



Even if the following two elements are all identical, their similarity cannot larger than 0.5

$$\frac{|X \cap Y|}{|X \cup Y|} \leq \frac{2}{|X| + |Y| - 2} \leq \frac{2}{|X| + 1 - 2} = \frac{2}{4} = 0.5 < 0.6$$

Prefix Filtering

Jaccard(x, y) ≥ σ ⇔ O(x, y) ≥ α=σ/(1+σ)(|x| + |y|)

- **Lemma.** If O(x, y) ≥ α, then the (|x| − α + 1)-prefix of x and the (|y| − α + 1)-prefix of y must share at least one token
  - **Contraposition**. if the (|x| − α + 1)-prefix of x and the (|y| − α + 1)-prefix of y do not have any common token, O(x, y) < α
  - **Proof.**
    - 1) $O(x, y) \leq |x| - (|x| - \alpha + 1) = \alpha - 1 < \alpha$
    - 2) $O(x, y) \leq |y| - (|y| - \alpha + 1) = \alpha - 1 < \alpha$

Building Inverted Lists

- With each record,
  - Insert the identifier of the record (RID) into the inverted list entries of its items in prefix only

$$\text{Jaccard}(x, y) \geq \sigma \Leftrightarrow O(x, y) \geq \alpha = \sigma/(1+\sigma)(|x| + |y|)$$

Q: How can we determine the prefix size?

RID	Items
R_1	C, D, F
R_2	A, B, E, F, G
R_3	A, B, C, D, E
R_4	B, C, D, E, F
R_5	A, E, G

Item	RIDs
A	
B	
C	R_1
D	R_1
E	
F	
G	

How to Determine Prefix Size

- Use the longest prefix for each record x, $|x| - \lceil \sigma \cdot |x| \rceil + 1$
- Proof.
  - $O(x, y) \leq |x| - (|x| - \lceil \sigma \cdot |x| \rceil + 1) = \lceil \sigma \cdot |x| \rceil - 1 < \sigma \cdot |x|$
  - $O(x, y) \leq |y| - (|y| - \lceil \sigma \cdot |y| \rceil + 1) = \lceil \sigma \cdot |y| \rceil - 1 < \sigma \cdot |y|$
  - $O(x, y) < \frac{\sigma}{2}(|x| + |y|) \leq \frac{\sigma}{1+\sigma}(|x| + |y|) = \alpha$

Building Inverted Lists & Counting Candidates

- With each record,
  - Insert the identifier of the record (RID) into the inverted list entries of its items in prefix only

$\sigma = 0.5$

RID	Items	$\lvert x\rvert - \lceil \sigma \cdot \lvert x\rvert \rceil + 1$
R_1	C, D, F	2
R_2	A, B, E, G	3
R_3	A, B, C, D, E	3
R_4	B, C, D, E, F	3
R_5	A, E, G	2

Item	RIDs
A	R_2, R_3, R_5
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_4
E	R_2, R_5
F	
G	

We cannot count exact overlapping tokens between candidates

Building Inverted Lists

- With each record,
  - Insert the identifier of the record (RID) into the inverted list entries of its items in prefix only

$\sigma = 0.5$

RID	Items	$\lvert x\rvert - \lceil \sigma \cdot \lvert x\rvert \rceil + 1$
R_1	C, D, F	2
R_2	A, B, E, G	3
R_3	A, B, C, D, E	3
R_4	B, C, D, E, F	3
R_5	A, E, G	2

Index items with actual records

Item	RIDs
A	R_2={A,B,E,G}, R_3={A,B,C,D,E} R_5={A,E,G}
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_4
E	R_2, R_5
F	
G	

Building Inverted Lists

- With each inverted list,
  - Investigate Jaccard similarity between all possible pairs of sets if they satisfy the threshold

$\sigma = 0.5$

Item	RIDs
A	R_2={A,B,E,G}, R_3={A,B,C,D,E} R_5={A,E,G}
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	R_1, R_4
E	R_2, R_5
F	
G	

RID	Items
R_1	C, D, F
R_2	A, B, E, G
R_3	A, B, C, D, E
R_4	B, C, D, E, F
R_5	A, E, G

Candidate pair	Sim
(R_2, R_3)	3/6
~~(R_1, R_3)~~	~~2/6~~
(R_1, R_4)	3/5
(R_2, R_5)	3/4

A Heuristic

- Sort items in the increasing order of frequency
  - Prefix may include more infrequent items and it can result in producing smaller candidate pairs
  - G(2), F(2), A(3), B(3), C(3), D(3), E(4)

$\sigma = 0.5$

Item	RIDs
A	R_2={G,A,B,E}, R_3={A,B,C,D,E} R_5={G,A,E}
B	R_2, R_3, R_4
C	R_1, R_3, R_4
D	
E	
F	
G	R_2, R_5

RID	Items
R_1	F, C, D
R_2	G, A, B, E
R_3	A, B, C, D, E
R_4	F, B, C, D, E
R_5	G, A, E

Candidate pair	Sim
(R_2, R_3)	3/6
~~(R_1, R_3)~~	~~2/6~~
(R_2, R_5)	3/4