

All Pairs

Nikolaus Augsten

University of Salzburg, Austria

October 24, 2017

Similarity functions

Function	Definition	$eqo(r,s)$	lb_r	ub_r
Jaccard	$\frac{ r \cap s }{ r \cup s }$	$\frac{t}{1+t}(r + s)$	$t \cdot r $	$\frac{ r }{t}$
Cosine	$\frac{ r \cap s }{\sqrt{ r \cdot s }}$	$t \sqrt{ r \cdot s }$	$t^2 r $	$\frac{ r }{t^2}$
Dice	$\frac{2 \cdot r \cap s }{ r + s }$	$\frac{t(r + s)}{2}$	$\frac{t r }{2-t}$	$\frac{(2-t) r }{t}$
Overlap	$ r \cap s $	t	t	∞

Table: Similarity functions and set size bounds for set r and s .

All Pairs

Collection:

r_1

1	3	5
---	---	---

r_2

1	2	3	4
---	---	---	---

r_3

1	2	4	9	11
---	---	---	---	----

r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

All Pairs

Collection:

r_1	1	3	5
-------	---	---	---

r_3	1	2	4	9	11
-------	---	---	---	---	----

r_2	1	2	3	4
-------	---	---	---	---

r_4	1	3	5	8	9	10	11	12	13	14
-------	---	---	---	---	---	----	----	----	----	----

- Jaccard: $J = \frac{|r \cap s|}{|r \cup s|}$
- $lb_r = t \cdot |r|$
- $\pi_r^i = |r| - \lceil \text{eqo}(r, r) \rceil + 1$
- $\text{eqo}(r, s) = \frac{t}{1+t} \cdot (|r| + |s|)$
- $\pi_r = |r| - \lceil lb_r \rceil + 1$

All Pairs

Threshold: $t = 0.5$

	$ r_i $	$\text{eqo}(r, r)$	lb_{r_i}	π_{r_i}	$\pi_{r_i}^i$
r_1	3	2	1.5	2	2
r_2	4	$2.\dot{6}$	2	3	2
r_3	5	$3.\dot{3}$	2.5	3	2
r_4	10	$6.\dot{6}$	5	6	4

Algorithm

Algorithm 1: AllPairs(R, t)

input : R collection of sets, t similarity threshold

output: res set of result pairs (similarity at least t)

```
 $I = \{\}$ ;          /*  $I$  inverted list index covering prefix of sets */
foreach  $r$  in  $R$  do      /* process in ascending length order of  $r$  */
     $M = \{\}$ ;          /* dictionary for candidate set. Key: candidate,
                        value: number of intersecting tokens found so far. */
    for  $p \leftarrow 0 < \pi_r - 1$  do /*  $\pi_r$ : probing prefix length of  $r$  */
        for  $s$  in  $I_{r[p]}$  do
            if  $|s| < lb_r$  then /*  $lb_r$ : length bound */
                | remove index entry with  $s$  from  $I_{r[p]}$ ;
            else
                | if  $s$  not in  $M$  then  $M[s] = 0$ ;
                |  $M[s] = M[s] + 1$ ;
    for  $p \leftarrow 0 < \pi_r^i - 1$  do /*  $\pi_r^i$ : indexing prefix length of  $r$  */
        |  $I_{r[p]} = I_{r[p]} \circ r$ ;          /* Add set  $r$  to index */
    /* Verify() verifies the candidates in  $M$  */
     $res = res \cup Verify(r, M, t)$ ;
```

Round 1

Current probing set: r_1

1	3	5
---	---	---

$$lb_r = 1.5, \quad \pi_r = 2, \quad \pi_r^i = 2 \quad M = \{\}$$

Round 1

Current probing set: r_1

1	3	5
---	---	---

$$lb_r = 1.5, \quad \pi_r = 2, \quad \pi_r^i = 2 \quad M = \{\}$$

$$I = \{\}$$

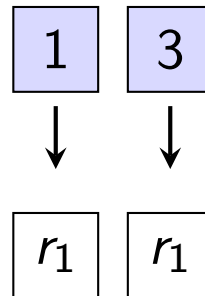
Round 1

Current probing set: r_1

1	3	5
---	---	---

$$lb_r = 1.5, \quad \pi_r = 2, \quad \pi_r^i = 2 \quad M = \{\}$$

Index updated:



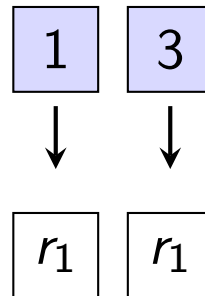
Round 1

Current probing set: r_1

1	3	5
---	---	---

$$lb_r = 1.5, \quad \pi_r = 2, \quad \pi_r^i = 2 \quad M = \{\}$$

Index updated:



No verification

$$res = \{\}$$

Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{\}$$

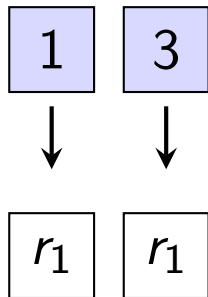
Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{\}$$

Index:



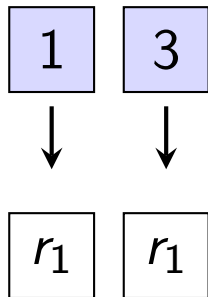
Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 2)\}$$

Index:



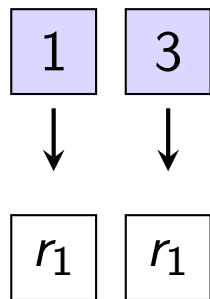
Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 2)\}$$

Index:



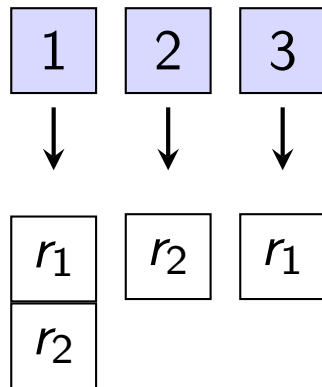
Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 2)\}$$

Index updated:



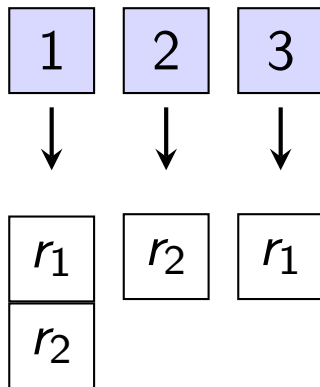
Round 2

Current probing set: r_2

1	2	3	4
---	---	---	---

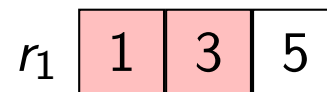
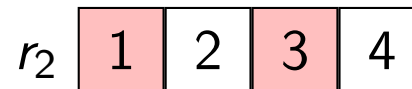
$$lb_r = 2, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 2)\}$$

Index updated:



Verification:

$$|r_2 \cap r_1| \not\geq \lceil \text{eqo}(r_2, r_1) \rceil \iff (2 \not\geq 3)$$



$$res = \{\}$$

Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{\}$$

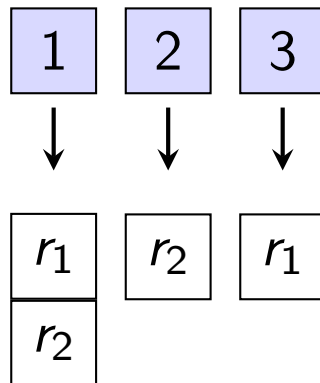
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{\}$$

Index:



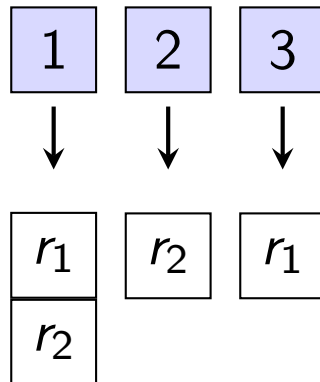
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 1), (r_2, 2)\}$$

Index:



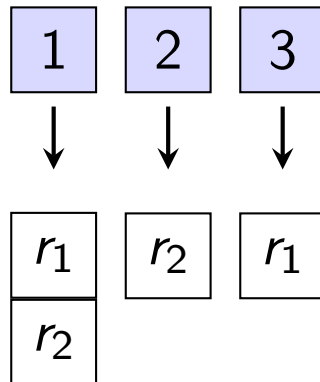
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 1), (r_2, 2)\}$$

Index:



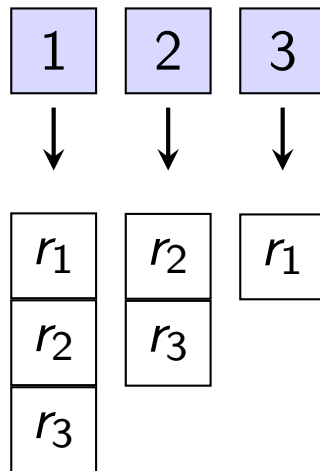
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 1), (r_2, 2)\}$$

Index updated:



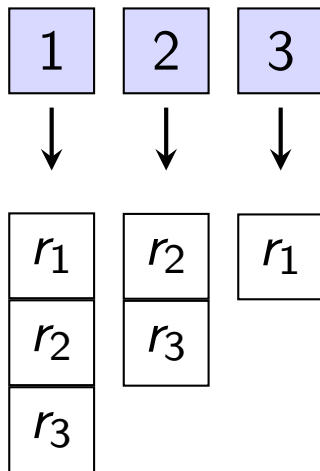
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

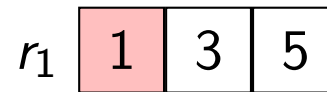
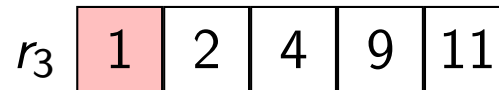
$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 1), (r_2, 2)\}$

Index updated:



Verification 1:

$|r_3 \cap r_1| \not\geq \lceil \text{eqo}(r_3, r_1) \rceil \iff (1 \not\geq 3)$



$res = \{\}$

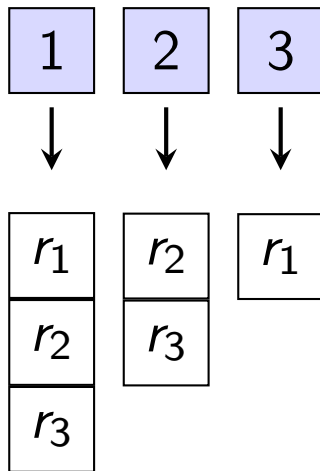
Round 3

Current probing set: r_3

1	2	4	9	11
---	---	---	---	----

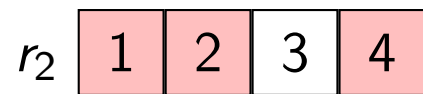
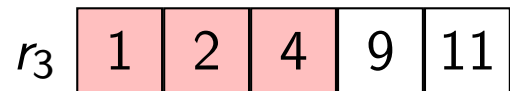
$$lb_r = 2.5, \quad \pi_r = 3, \quad \pi_r^i = 2 \quad M = \{(r_1, 1), (r_2, 2)\}$$

Index updated:



Verification 2:

$$|r_3 \cap r_2| \geq \lceil \text{eqo}(r_3, r_2) \rceil \iff (3 \geq 3)$$



$$res = \{(r_3, r_2)\}$$

Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{\}$$

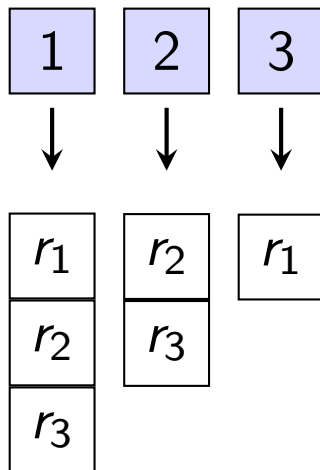
Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{\}$$

Index:



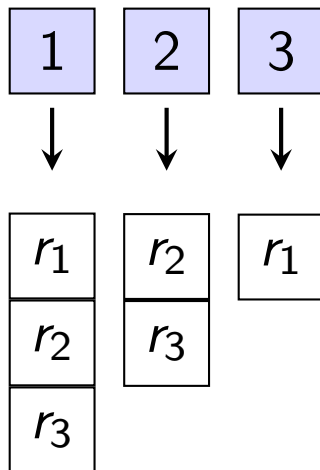
Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{(r_3, 1)\}$$

Index:



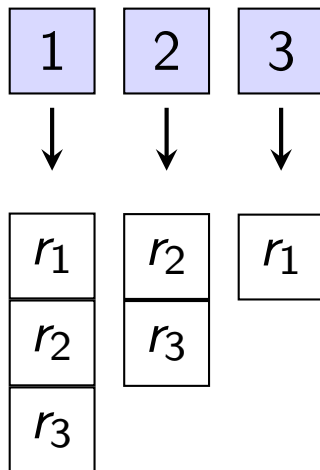
Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{(r_3, 1)\}$$

Index:



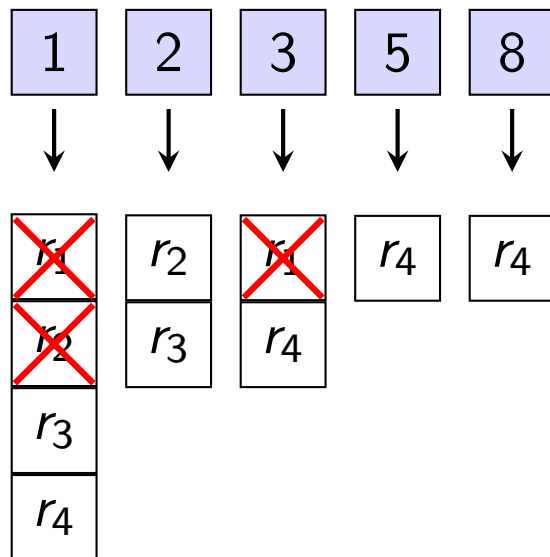
Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{(r_3, 1)\}$$

Index updated:



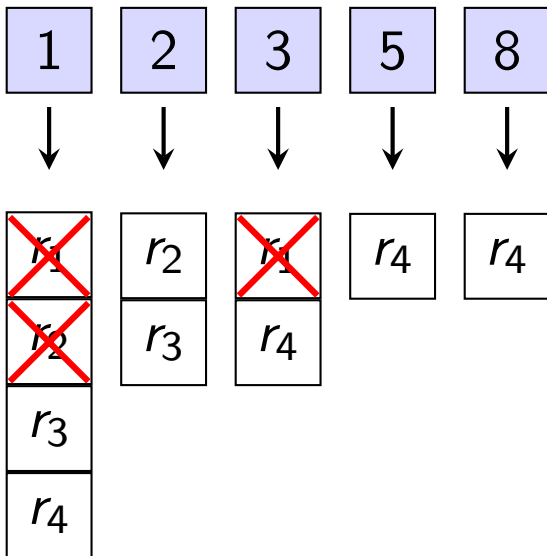
Round 4

Current probing set: r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

$$lb_r = 5, \quad \pi_r = 6, \quad \pi_r^i = 4 \quad M = \{(r_3, 1)\}$$

Index updated:



Verification:

$$|r_4 \cap r_3| \not\geq \lceil \text{eqo}(r_4, r_3) \rceil \iff (2 \not\geq 5)$$

r_4

1	3	5	8	9	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

r_3

1	2	4	9	11
---	---	---	---	----

$$res = \{(r_3, r_2)\}$$