PS Ähnlichkeitssuche in großen Datenbanken

Set Similarity Queries

Thomas Hütter, WS 2020

Motivation

Set Similarity Queries

- Suppose you are the CTO of bookface, a social media platform.

- Since most of the users have only a small number of friends, the CEO asks you to increase it.

- Being puzzled for several days, you have the brilliant idea to recommend users with similar hobbies to one another.

Motivation

Set Similarity Queries

- All users are stored in a database with the following schema:

ID	Name	Hobbies
1	Peter	{guitar, biking, swimming}
2	Sarah	{skiing, hiking, singing}
3	John	{singing, hiking}
4	Kate	{guitar, skiing, swimming, running}
…	…	…

- Open questions we will address in this lab:
  - How can we define similarity between sets?
  - How can we find pairs of similar sets?
  - How to scale to large numbers of large sets.

Definition

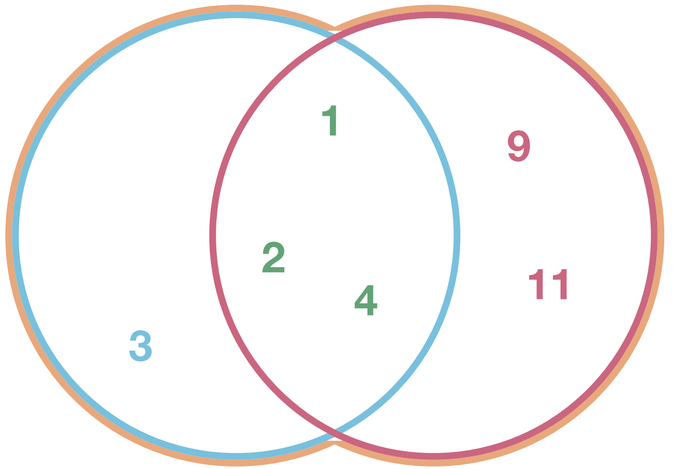
Set Similarity Functions

- Given: two sets *r* and *s*.

r =	1	2	3	4

s =	1	2	4	9	11

- Goal: a similarity measure for sets.
- Existing measures:
  - Overlap: $O(r, s) = |r \cap s| = 3$
  - Jaccard: $J(r, s) = |r \cap s| \,/\, |r \cup s| = 0.5$



|r ∩ s| / |r ∪ s|

Further Definitions

Set Similarity Functions

Table 2 Definitions of Measures for binary data

$$S_{JACCARD} = \frac{a}{a+b+c} \quad (1)$$

$$S_{DICE} = \frac{2a}{2a+b+c} \quad (2)$$

$$S_{CZEKANOWSKI} = \frac{2a}{2a+b+c} \quad (3)$$

$$S_{SW-JACCARD} = \frac{3a}{3a+b+c} \quad (4)$$

$$S_{NEI\&LI} = \frac{2a}{(a+b)+(a+c)} \quad (5)$$

$$S_{SOKAL\&SNEATH-I} = \frac{a}{a+2b+2c} \quad (6)$$

$$S_{SOKAL\&MICHENER} = \frac{a+d}{a+b+c+d} \quad (7)$$

$$S_{SOKAL\&SNEATH-II} = \frac{2(a+d)}{2a+b+c+2d} \quad (8)$$

$$S_{ROGER\&TANIMOTO} = \frac{a+d}{a+2(b+c)+d} \quad (9)$$

$$S_{FAITH} = \frac{a+0.5d}{a+b+c+d} \quad (10)$$

$$S_{GOWER\&LEGENDRE} = \frac{a+d}{a+0.5(b+c)+d} \quad (11)$$

$$S_{INTERSECTION} = a \quad (12)$$

$$S_{INNERPRODUCT} = a+d \quad (13)$$

$$S_{RUSSELL\&RAO} = \frac{a}{a+b+c+d} \quad (14)$$

$$D_{HAMMING} = b+c \quad (15)$$

$$D_{EUCLID} = \sqrt{b+c} \quad (16)$$

$$D_{SQUARED-EUCLID} = \sqrt{(b+c)^2} \quad (17)$$

$$D_{CANBERRA} = (b+c)^{\frac{2}{2}} \quad (18)$$

$$D_{MANHATTAN} = b+c \quad (19)$$

$$D_{MEAN-MANHATTAN} = \frac{b+c}{a+b+c+d} \quad (20)$$

$$D_{CITYBLOCK} = b+c \quad (21)$$

$$D_{MINKOWSKI} = (b+c)^{\frac{1}{1}} \quad (22)$$

$$D_{VARI} = \frac{(b+c)}{4(a+b+c+d)} \quad (23)$$

$$D_{SIZEDIFFERENCE} = \frac{(b+c)^2}{(a+b+c+d)^2} \quad (24)$$

$$D_{SHAPEDIFFERENCE} = \frac{n(b+c)-(b-c)^2}{(a+b+c+d)^2} \quad (25)$$

$$D_{PATTERNDIFFERENCE} = \frac{4bc}{(a+b+c+d)^2} \quad (26)$$

$$D_{LANCE\&WILLIAMS} = \frac{b+c}{(2a+b+c)} \quad (27)$$

$$D_{BRAY\&CURTIS} = \frac{b+c}{(2a+b+c)} \quad (28)$$

$$D_{HELLINGER} = 2\sqrt{\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)} \quad (29)$$

$$D_{CHORD} = \sqrt{\left(2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)\right)} \quad (30)$$

$$S_{COSINE} = \frac{a}{\sqrt{(a+b)(a+c)}^2} \quad (31)$$

$$S_{GILBERT\&WELLS} = \log a - \log n - \log\left(\frac{a+b}{n}\right) - \log\left(\frac{a+c}{n}\right) \quad (32)$$

$$S_{OCHIAI-I} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (33)$$

$$S_{FORBESI} = \frac{na}{(a+b)(a+c)} \quad (34)$$

$$S_{FOSSUM} = \frac{n(a-0.5)^2}{(a+b)(a+c)} \quad (35)$$

$$S_{SORGENFREI} = \frac{a^2}{(a+b)(a+c)} \quad (36)$$

$$S_{MOUNTFORD} = \frac{a}{0.5(ab+ac)+bc} \quad (37)$$

$$S_{OTSUKA} = \frac{a}{((a+b)(a+c))^{0.5}} \quad (38)$$

$$S_{MCCONNAUGHEY} = \frac{a^2-bc}{(a+b)(a+c)} \quad (39)$$

$$S_{TARWID} = \frac{na-(a+b)(a+c)}{na+(a+b)(a+c)} \quad (40)$$

$$S_{KULCZYNSKI-II} = \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} \quad (41)$$

$$S_{DRIVER\&KROEBER} = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right) \quad (42)$$

$$S_{JOHNSON} = \frac{a}{a+b}+\frac{a}{a+c} \quad (43)$$

$$S_{DENNIS} = \frac{ad-bc}{\sqrt{n(a+b)(a+c)}} \quad (44)$$

$$S_{SIMPSON} = \frac{a}{\min(a+b,a+c)} \quad (45)$$

$$S_{BRAUN\&BANQUET} = \frac{a}{\max(a+b,a+c)} \quad (46)$$

$$S_{FAGER\&MCGOWAN} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2} \quad (47)$$

$$S_{FORBES-II} = \frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)} \quad (48)$$

$$S_{SOKAL\&SNEATH-IV} = \frac{\frac{a}{(a+b)}+\frac{a}{(a+c)}+\frac{d}{(b+d)}+\frac{d}{(b+d)}}{4} \quad (49)$$

$$S_{GOWER} = \frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (50)$$

$$S_{PEARSON-I} = \chi^2 \text{ where } \chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} \quad (51)$$

$$S_{PEARSON-II} = \left(\frac{\chi^2}{n+\chi^2}\right)^{1/2} \quad (52)$$

$$S_{PEARSONIII} = \left(\frac{\rho}{n+\rho}\right)^{1/2} \text{ where } \rho = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (53)$$

$$S_{PEARSON\&HERON-I} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (54)$$

$$S_{PEARSON\&HERON-II} = Cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right) \quad (55)$$

$$S_{SOKAL\&SNEATH-III} = \frac{a+d}{b+c} \quad (56)$$

$$S_{SOKAL\&SNEATH-V} = \frac{ad}{(a+b)(a+c)(b+d)(c+d)^{0.5}} \quad (57)$$

$$S_{COLE} = \frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}} \quad (58)$$

$$S_{STILES} = \log_{10}\frac{n(|ad-bc|-\frac{n}{2})^2}{(a+b)(a+c)(b+d)(c+d)} \quad (59)$$

$$S_{OCHIAI-II} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (60)$$

$$S_{YULEQ} = \frac{ad-bc}{ad+bc} \quad (61)$$

$$D_{YULEQ} = \frac{2bc}{ad+bc} \quad (62)$$

$$S_{YULEw} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \quad (63)$$

$$S_{KULCZYNSKI-I} = \frac{a}{b+c} \quad (64)$$

$$S_{TANIMOTO} = \frac{a}{(a+b)+(a+c)-a} \quad (65)$$

$$S_{DISPERSON} = \frac{ad-bc}{(a+b+c+d)^2} \quad (66)$$

$$S_{HAMANN} = \frac{(a+d)-(b+c)}{a+b+c+d} \quad (67)$$

$$S_{MICHAEL} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2} \quad (68)$$

$$S_{GOODMAN\&KRUSKAL} = \frac{\sigma-\sigma'}{2n-\sigma'} \text{ where } \sigma = \max(a,b)+\max(c,d)+\max(a,c)+\max(b,d),\ \sigma' = \max(a+c,b+d)+\max(a+b,c+d) \quad (69)$$

$$S_{ANDERBERG} = \frac{\sigma-\sigma'}{2n} \quad (70)$$

$$S_{BARONI-URBANI\&BUSER-I} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c} \quad (71)$$

$$S_{BARONI-URBANI\&BUSER-II} = \frac{\sqrt{ad}+a-(b+c)}{\sqrt{ad}+a+b+c} \quad (72)$$

$$S_{PEIRCE} = \frac{ab+bc}{ab+2bc+cd} \quad (73)$$

$$S_{EYRAUD} = \frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} \quad (74)$$

$$S_{TARANTULA} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)} \quad (75)$$

$$S_{AMPLE} = \left|\frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}}\right| = \left|\frac{a(c+d)}{c(a+b)}\right| \quad (76)$$

Choi, Seung-Seok, Sung-Hyuk Cha, and Charles C. Tappert. "A survey of binary similarity and distance measures." *Journal of Systemics, Cybernetics and Informatics* 8.1 (2010): 43-48.

Exercise

Set Similarity Functions

- Given: two sets *r* and *s*.

r = | 6 | 10 | 24 | 49 | 65 | 71 |

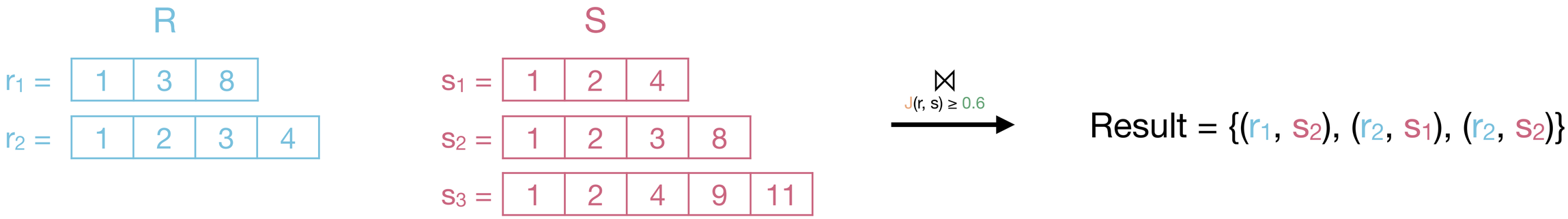
s = | 1 | 6 | 11 | 21 | 34 | 49 | 51 | 65 |

- Task: compute Jaccard and Overlap for *r* and *s*.

Definition

Set Similarity Joins

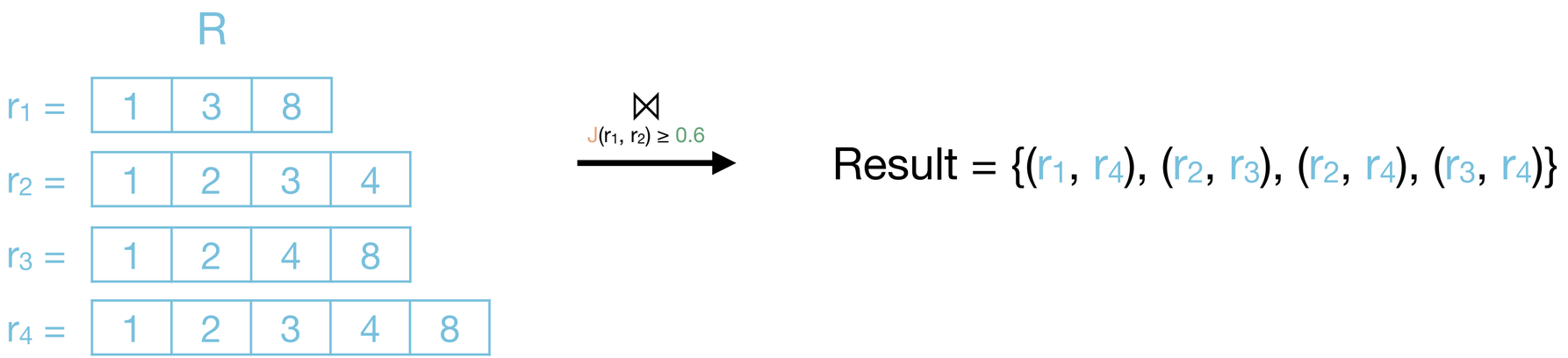
- Given: two collections of sets R and S, a similarity function sim, and a similarity threshold t.
- Goal: compute all pairs of similar sets, denoted $R \underset{sim(r, s) \geq t}{\bowtie} S = \{(r, s) \in R \times S \mid sim(r, s) \geq t\}$.
- Example: sim = Jaccard, t = 0.6



Definition

Set Similarity Self Joins

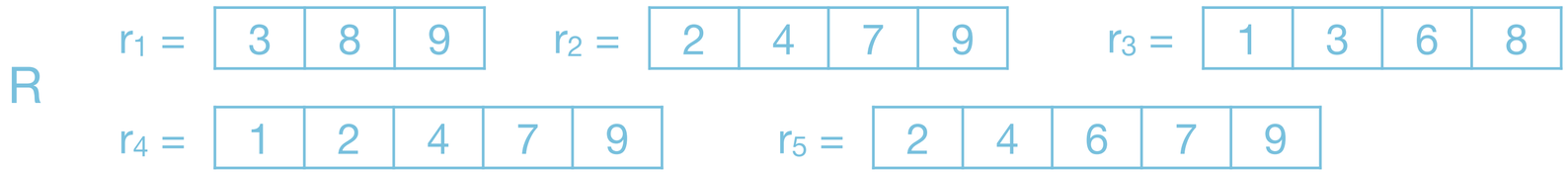
- Given: a collection of sets *R*, a similarity function *sim*, and a similarity threshold *t*.
- Goal: compute all pairs of similar sets, denoted $R \underset{sim(r_1, r_2) \geq t}{\bowtie} R = \{(r_1, r_2) \in R \times R \mid r_1 \neq r_2 \wedge sim(r_1, r_2) \geq t\}$.
- Example: sim = Jaccard, t = 0.6



Exercise

Set Similarity Self Joins

- Given: a collections of sets *R*, a similarity function *Jaccard*, and a similarity threshold *0.75*.



- Task: perform a self similarity join.

Motivation

Set Similarity Self Joins

- All users are stored in a database with the following schema:

ID	Name	Hobbies
1	Peter	{guitar, biking, swimming}
2	Sarah	{skiing, hiking, singing}
3	John	{singing, hiking}
4	Kate	{guitar, skiing, swimming, running}
…	…	…

- Open questions we will address in this lab:
  - How can we define similarity between sets? → Jaccard, overlap, …
  - How can we find pairs of similar sets? → set similarity self join
  - How to scale to large numbers of large sets. → our semester project