University of Augsburg, Institute of Computer Science
Prof. Dr. P. Fischer
J. Kastner, L. Rudenko

SS 2019
17. June 2019
Solution 6

Analyzing Massive Data Sets

Exercise 1: Expressing Similarity (homework)

First of all the word frequencies are determined.

Word	$freq_{D_1}$	$freq_{D_2}$	$freq_{D_3}$	$freq_Q$
to	1	0	0	0
chair	2	3	1	3
cat	2	0	2	1
red	2	1	1	1
blue	0	1	0	1
lies	1	0	0	1
next	1	0	0	0
on	2	0	0	1
the	4	3	3	4
between	0	1	1	1
black	2	1	2	2
and	0	1	1	1
is	0	1	1	0

The similarities of each object to Query Q are calculated as follows:

a) **Manhattan Distance**

- $M(D_1, Q) = 1 + 1 + 1 + 1 + 1 + 0 + 1 + 1 + 0 + 1 + 0 + 1 + 0 = 9$
- $M(D_2, Q) = 0 + 0 + 1 + 0 + 0 + 1 + 0 + 1 + 1 + 0 + 1 + 0 + 1 = 6$
- $M(D_3, Q) = 0 + 2 + 1 + 0 + 1 + 1 + 0 + 1 + 1 + 0 + 0 + 0 + 1 = 8$

The best matching Documents to Query Q are D_2followed by D_3 and D_1.

b) **Canberra Distance**

- $CB(D_1, Q) = 1 + \frac{1}{5} + \frac{1}{3} + \frac{1}{3} + 1 + 0 + 1 + \frac{1}{3} + 0 + 1 + 0 + 1 + 0 = 6.2$
- $CB(D_2, Q) = 0 + 0 + 1 + 0 + 0 + 1 + 0 + 1 + \frac{1}{7} + 0 + \frac{1}{3} + 0 + 1 = 4.5$
- $CB(D_3, Q) = 0 + \frac{1}{2} + \frac{1}{3} + 0 + 1 + 1 + 0 + 1 + \frac{1}{7} + 0 + 0 + 0 + 1 = 5.0$

The best matching Documents to Query Q are D_2 followed by D_3 and D_1.

Exercise 2: Locality Sensitive Hashing (homework)

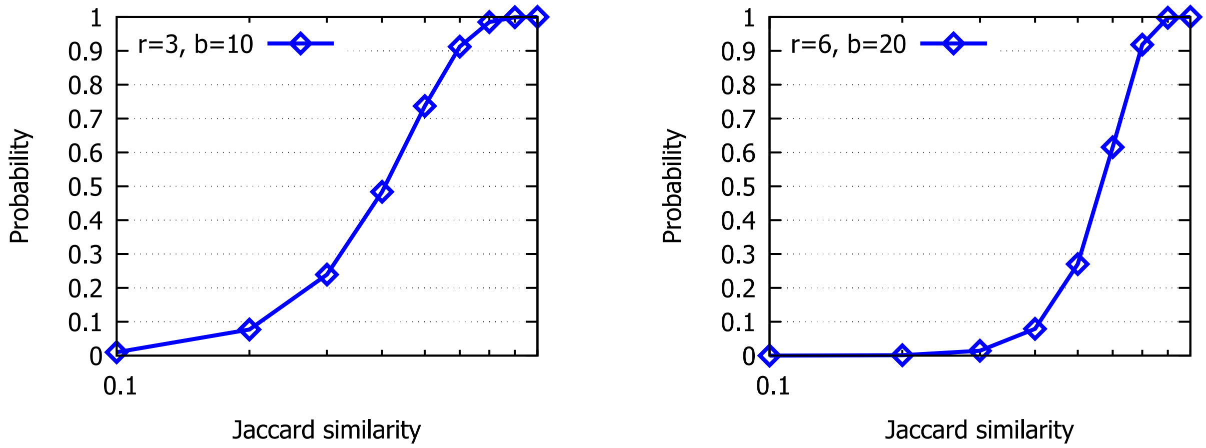
Evaluate the S-curve $1 - (1 - s^r)^b$ for $s = 0.1, 0.2, ..., 0.9$, for the following values of r and b:

a) $r = 3, b = 10$

b) $r = 6, b = 20$

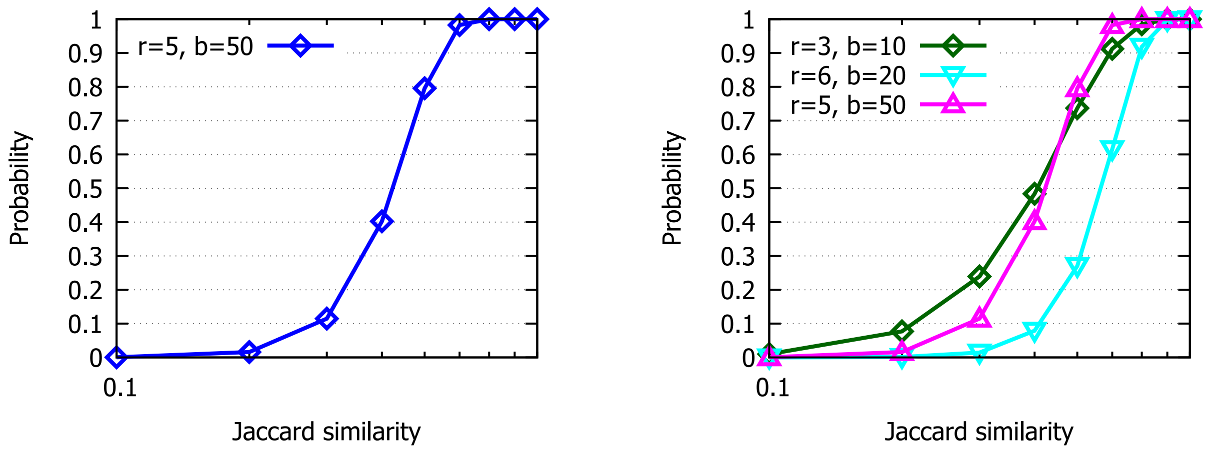
c) $r = 5, b = 50$

s	$1-(1-s^r)^b$	s	$1-(1-s^r)^b$	s	$1-(1-s^r)^b$
	$r=3, b=10$		$r=6, b=20$		$r=5, b=50$
0.1	0.009955	0.1	0.00002	0.1	0.0005
0.2	0.07718	0.2	0.001279	0.2	0.01588
0.3	0.23945	0.3	0.014479	0.3	0.11454
0.4	0.48387	0.4	0.078809	0.4	0.402284
0.5	0.73692	0.5	0.270187	0.5	0.79555
0.6	0.91227	0.6	0.615415	0.6	0.98253
0.7	0.985015	0.7	0.918186	0.7	0.999899
0.8	0.999234	0.8	0.997712	0.8	0.9999999976
0.9	0.999998	0.9	0.999999	0.9	1



(a) r=3, b=10 (b) r=6, b=20

Abbildung 1: S-curves for $r = 3, b = 10$ and $r = 6, b = 20$.



(a) r=5, b=50 (b) comparision of S-curves

Abbildung 2: S-curve for $r = 5, b = 50$ and comparision of 3 S-curves.

Exercise 3: Hierarchical Clustering (live)

1. The subtusks a) to c) were discussed in the exercise.
2. Average distance among all pairs of nodes in each cluster $D(X,Y) = \frac{1}{|X|*|Y|}\sum_{x\in X, y\in Y} d(x,y)$:

   step 1:
   $minD = D(I,J) = d(I,J) = D(J,K) = d(J,K) = \sqrt{2}$
   Combine $J(11,4)$ and $K(12,3)$ to the new cluster $\{JK\}$.

   We compute the **average distance** to this new cluster from the points $H(9,3)$, $I(10,5)$ and $L(12,6)$ (the other points are not close enough to be combined with this cluster):
   - $D(I,\{JK\}) = (d(I,J) + d(J,K))/2 = (\sqrt{2} + 2\sqrt{2})/2 = 2.12$
   - $D(H,\{JK\}) = (d(H,J) + d(H,K))/2 = (\sqrt{5} + 3)/2 = 2.62$
   - $D(L,\{JK\}) = (d(L,J) + d(L,K))/2 = (\sqrt{5} + 3)/2 = 2.62$

   step 2:
   $minD = D(C,D) = d(C,D) = D(C,F) = d(C,F) = 2$
   Combine $C(4,8)$ and $D(4,10)$ to the new cluster $\{CD\}$.

   And we compute **average distance** to this new cluster from the points $F(6,8)$ and $G(7,10)$ (the other points are not close enough to be combined with this cluster):
   - $D(F,\{CD\}) = (d(F,D) + d(F,C))/2 = (2\sqrt{2} + 2)/2 = 2.4$
   - $D(G,\{CD\}) = (d(G,D) + d(G,C))/2 = (3 + \sqrt{13})/2 = 3.3$

   step 3:
   $minD = D(I,\{JK\}) = (d(I,J) + d(J,K))/2 = 2.12$
   Combine $I(4,8)$ and the cluster $\{JK\}$ to the new cluster $\{IJK\}$.

   And we compute **average distance** to this new cluster from the points $H(9,3)$ and $L(12,6)$ (the other points are not close enough to be combined with this cluster):
   - $D(H,\{IJK\}) = (d(H,I) + d(H,J) + d(H,K))/3 = (\sqrt{5} + \sqrt{5} + 3)/3 = 2.49$
   - $D(L,\{IJK\}) = (d(L,I) + d(L,J) + d(L,K))/3 = (\sqrt{5} + \sqrt{5} + 3)/3 = 2.49$

   step 4:
   $minD = D(A,B) = d(A,B) = D(F,G) = d(F,G) = \sqrt{5} = 2.24$
   Combine $A(2,2)$ and $B(3,4)$ to the new cluster $\{AB\}$.

   And we compute **average distance** to this new cluster from the point $E(5,2)$ (the other points are not close enough to be combined with this cluster):
   - $D(E,\{AB\}) = (d(E,A) + d(E,B))/2 = (3 + 2\sqrt{2})/2 = 2.915$

   step 5:
   $minD = D(F,G) = d(F,G) = \sqrt{5} = 2.24$
   Combine $F(6,8)$ and $G(7,10)$ to the new cluster $\{FG\}$.

   And we compute **average distance** to this new cluster from the cluster $\{CD\}$:
   - $D(\{CD\},\{FG\}) = (d(C,F) + d(C,G) + d(D,F) + d(D,G))/4$
     $= (2 + \sqrt{13} + 2\sqrt{2} + 3)/4 = 2.86$

step 6:
$minD = D(H, \{IJK\}) = (d(H,I)+d(H,J)+d(H,K))/3 = D(L, \{IJK\}) = (d(L,I)+d(L,J)+d(L,K))/3 = 2.49$
Combine $H(9,3)$ and the cluster $\{IJK\}$ to the new cluster $\{HIJK\}$.

And we compute **average distance** to this new cluster from the point $L(12,6)$:

- $D(L, \{HIJK\}) = (d(L,H) + d(L,I) + d(L,J) + d(L,K))/4 = (3\sqrt{2} + \sqrt{5} + \sqrt{5} + 3)/4 = 2.93$

step 7:
$minD = D(\{CD\}, \{FG\}) = (d(C,F) + d(C,G) + d(D,F) + d(D,G))/4 = 2.86$
Combine the cluster $\{CD\}$ and the cluster $\{FG\}$ to the new cluster $\{CDFG\}$.

step 8:
$minD = D(E, \{AB\}) = (d(E,A) + d(E,B))/2 = 2.915$
Combine $E(5,2)$ and the cluster $\{AB\}$ to the new cluster $\{ABE\}$.

step 9:
$minD = D(L, \{HIJK\}) = (d(L,H) + d(L,I) + d(L,J) + d(L,K))/4 = 2.93$
Combine $L(12,6)$ and the cluster $\{HIJK\}$ to the new cluster $\{HIJKL\}$.

We have **three clusters**, we are ready. The result is shown in the Figures 5, 4 and 3

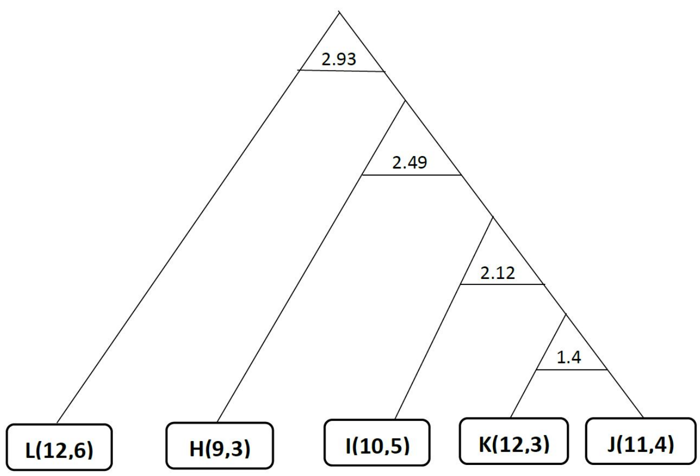


Abbildung 3: Average Distance: Cluster with points H, I, J, K, L

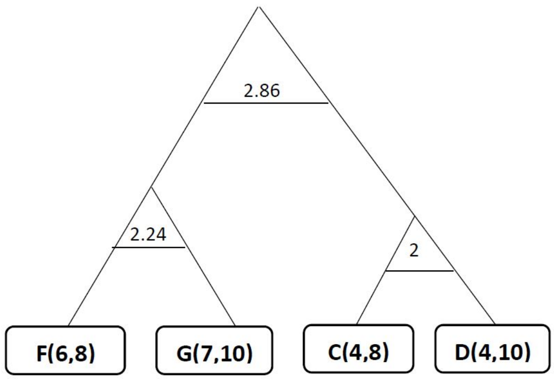


Abbildung 4: Average Distance: Cluster with points C, D, F, G

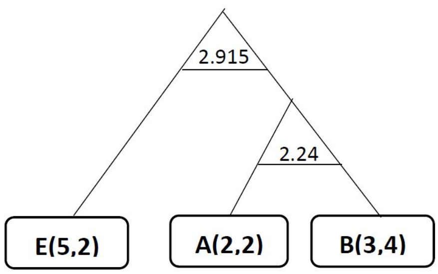


Abbildung 5: Average Distance: Cluster with points A, B, E