Information Retrieval Systems Class Notes (Week 6)

Prepared by: Hamed Rezanejad

δ: Decoupling coefficient

δ_i => for documents $1 <= i <= m$ Average δ

δ_i' => for terms $1 <= i <= n$ Average δ'

$n_c = \sum \delta_i = m \times \delta$, $n_c' = \sum \delta_i = n \times \delta'$ => $n_c = n_c'$

- From previous example:

D:

	t_1	t_2	t_3	t_4	t_5	t_6
d_1	1	0	0	1	0	1
d_2	1	1	1	1	0	0
d_3	1	0	0	0	1	1
d_4	0	0	0	0	1	1
d_5	1	0	1	0	0	0

C:

0.361	0.250	0.194	0.111	0.083
0.188	0.563	0.063	0.000	0.188
0.194	0.083	0.361	0.277	0.088
0.167	0.000	0.417	0.417	0.000
0.125	0.375	0.125	0.000	0.375

$n_c = \sum c_{ii} = \sum \delta_i \approx 2$

$C_{ij} = \alpha_i \times \sum_{k=1}^{n} d_{ik} \times \beta_k \times d_{jk}$, $\delta_i = C_{ii}$, $\psi_i = 1 - C_{ii}$

Seed power of d_i: $P_i = \delta_i \times \psi_i \times X_{di}$, X_{di} = No. of terms in d_i (depth of indexing)

$P_1 = 0.361 \times (1-0.361) \times 3 = 0.692$

$P_2 = 0.563 \times (1-0.563) \times 4 = 0.984$

$P_3 = 0.692$

$P_4 = 0.484$

$P_5 = 0.469$

⇨ Select the documents with the highest seed value as the cluster seeds: choose 2 of them => n_c=2

C_{11} (0.361) = C_{33} (0.361) ≠ C_{13} (0.199) = C_{31} (0.199) => they are not identical

- Which one to choose? d_1 or d_3 ?
  - ✓ Look at the terms used by cluster seeds chosen so far (d_2)
    $d_2 = t_1, t_2, t_3, t_4$
    Two candidate: d_1: t_1, t_4, t_6
    d_3: t_5, t_6 => select d_3 as the next seed

  - ✓ Assign d_1 to one of the seeds
    C_{12} = 0.250 > C_{13} = 0.194
    So, it is better to join cluster of d_2

Computational cost:

Find all C_{ii} values = m

Assign non-seeds to seeds = $f(m,n_c) = (m-n_c) \times n_c$

$m + mlogm + (m - n_c) \times n_c \approx m \times n_c$,because ($n_C \ll m$)

$O(m \times n_c \times d)$, d : average depth of indexing (average number of terms / document)

Inverted index for seed document (IISD):

t_1 : <2,1> <3,1>

t_2 : <2,1>

t_3 : <2,1>

t_4 : <2,1>

t_5 : <3,1>

t_6 : <3,1>

C_{12} = 0 , C_{13} = 0

- t1 for d1:

$C_{12} = C_{12} + \alpha_1 \times (d_{11} \times \beta_1 \times d_{21}) = 0 + 1/3\ (1 \times 1/4 \times 1) = 1/12$

$C_{13} = C_{13} + \alpha_1 \times (d_{11} \times \beta_1 \times d_{31}) = 0 + 1/3\ (1 \times 1/4 \times 1) = 1/12$

- t_4 for d_1 :

$C_{12} = C_{12} + \alpha_1 \times (d_{14} \times \beta_4 \times d_{24}) = 1/12 + 1/3\ (1 \times 1/2 \times 1) = 0.250$

- t_6 for d_1 :

$C_{13} = C_{13} + \alpha_1 \times (d_{16} \times \beta_6 \times d_{36}) = 1/12 + 1/3\ (1 \times 1/3 \times 1) = 0.194$

Computational cost:

$(m\text{-}n_c) \approx m$ => $O(m \times X_d \times tg_s)$, X_d = average number of term/documents

Indexing clustering relation implied by C^3M

$n_C = \sum_{i=1}^{m} \delta_i = \sum_{i=1}^{n} \delta_i'$

$n_c = (m \times n)/t = (5 \times 6)/14 = 30/14 \approx 2$,t: number of terms (non-zero terms) in D

X_d = depth of indexing D (average number of terms/documents) = t/m

t_g = average number of document / term = t/n

⇨ $n_c = n/X_d = m/t_g$

Cluster validation (cluster quality measurement)

➢ Are the clusters meaningful??

Cluster Hypothesis: documents similar to each other would be relevant to the same query and would appear in the same cluster.

0. User evaluation: difficult
1. Internal criterion
    a. High intra cluster similarity
        i. Documents within a cluster are highly similar
    b. Low intra cluster similarity
        i. Documents in different clusters are highly dissimilar
2. External criterion

Classification		Clustering
Gold standard (ground truth) =clustering structures produced by human experts	Check the similarity	Clustering structure generated by algorithms

Rand Index (RI): consider pairs of objects and find ratio of correct decision.

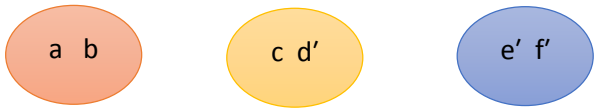
Total number of this pairs: $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$

- ✓ TP (true positive) = two "similar" documents are assigned to the "same" clusters
- ✓ TN (true negative) = two "dissimilar" documents are assigned to "different" clusters
- ✓ FP (false positive) = two "dissimilar" documents are assigned to "same" clusters
- ✓ FN (false negative) = two "similar" documents are assigned to "different" clusters

$$RI = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Ground truth:

Clustering:



Number of pairs: $\binom{6}{2} = 15$

ab	ac	ad'	ae'	af'	bc	bd'	be'	bf'	cd'	ce'	cf'	d'e'	d'f'	e'f'
TP	FN	TN	TN	TN	FN	TN	TN	TN	FP	TN	TN	FN	FN	TP

RI = (2+8)/15 = 10/15 = 2/3 = 0.66

F-measure -> F= $\frac{2 \times P \times R}{P+R}$

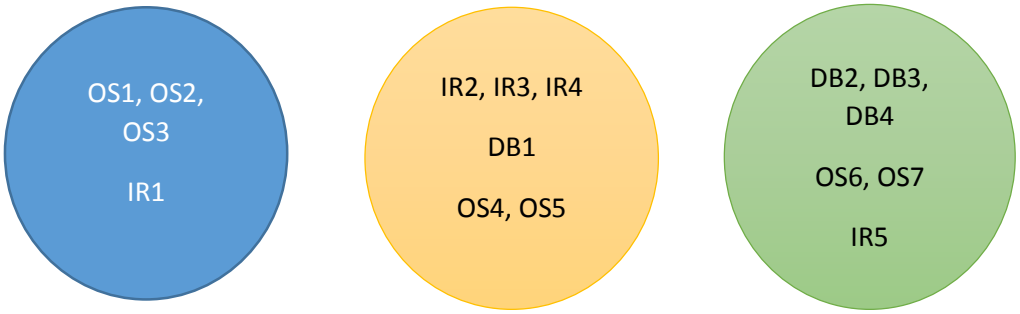
Harmonic mean of Recall (R) and Precision (P):

$P = \frac{TP}{TP+FP} = 2/3$, $R = \frac{TP}{TP+FN} = 2/6$

⇨ F-measure = $\frac{2 \times \frac{2}{3} \times \frac{2}{6}}{\frac{2}{3}+\frac{2}{6}}$ = 0.44

Purity: each cluster is assigned to the class which is most frequent in the cluster. Number the correct assignments and then divide it by the total number of elements.

- DB: 4, IR: 5, OS: 7
- Clusters:



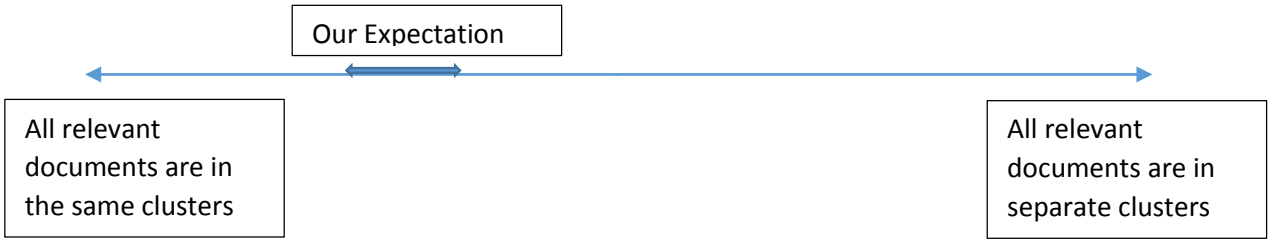
Purity = $\frac{3+3+3}{4+6+6}$ = 9/16 = 0.56

Purity is another approach that we can use with IR test collections.

For a particular query consider clusters that contain at least one relevant document (target cluster).

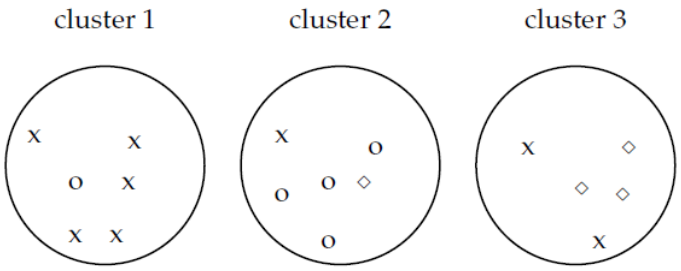
Cluster Hypothesis: documents similar to each other would be relevant to the same query.

Expectation: number of target clusters for a query should be small.



Questions:

- Consider below clusters:



1. Compute Purity for this clustering.

   Answer:

   Number of members of the majority class for the three clusters are:

   x: 5 (cluster 1);
   o: 4 (cluster 2);
   ◇: 3 (cluster 3);

   Purity = $\frac{5+4+3}{17}$ = 0.7

2. Compute Rand Index for this clustering.

   Answer:

   We first compute TP +FP. The three clusters contain 6, 6, and 5 points, respectively, so the total number of "positives" or pairs of documents that are in the same cluster is:

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

   Of these, the x pairs in cluster 1, the o pairs in cluster 2, the ◇ pairs in cluster 3, and the x pair in cluster 3 are true positives:

$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

Thus, FP= 40-20 = 20. FN and TN are computed similarly, resulting in the following contingency table:

	Same cluster	Different clusters
Same class	TP = 20	FN = 24
Different classes	FP = 20	TN = 72

RI = (20+72)/(20+ 20+24 +72) ≈ 0.68

3. Compute F-measure for this clustering.

Answer:

Based on previous question calculations, we can compute P and R:
P = 20/40 = 0.5, R= 20/44 ≈ 0.455

So, we can calculate F-measure easily: $F \approx 0.48$

- Consider below D matrix:

D:

	t1	t2	t3	t4	t5
d1	0	0	0	1	1
d2	1	1	0	1	0
d3	1	1	1	1	0
d4	0	1	0	1	1

4. Compute the clusters using C3M:

S:

	t1	t2	t3	t4	t5
d1	0	0	0	0.5	0.5
d2	0.33	0.33	0	0.33	0
d3	0.25	0.25	0.25	0.25	0
d4	0	0.33	0	0.33	0.33

S′:

	t1	t2	t3	t4	t5
d1	0	0	0	0.25	0.5
d2	0.5	0.33	0	0.25	0
d3	0.5	0.33	1	0.25	0
d4	0	0.33	0	0.25	0.5

$C = S \times S'^{T} =$

0.375	0.125	0.125	0.375
0.083	0.356	0.356	0.191
0.063	0.270	0.520	0.145
0.248	0.191	0.191	0.356

$n_c = 5*4/12 = 1.66 \approx 2$

P1 = (0.375)*(0.625)*2 = 0.468
P2 = (0.356)*(0.644)*3 = 0.687
P3 = (0.520)*(0.480)*4 = 0.998
P4 = (0.356)*(0.644)*3 = 0.687

Now we choose **d3** as our first seed of cluster and then we have: d3->t1, t2, t3, t4
And d2-> t1, t2, t4 , d4-> t2, t4, t5. So, we choose **d4** as the next seed.

d2 => C_{23} = 0.356 > C_{24}=0.191 => we join d2 to cluster of d3
d1 => C_{13} = 0.125 < C_{14}=0.375 => we join d1 to cluster of d4

5. Compute the clusters using single link dendrogram (compute similarity matrix using dice coefficient):

S12= (2*1)/(2+3) = 0.4, S13= (2*1)/(2+4) = 0.33, S14= (2*2)/(2+3) = 0.8
S23= (2*3)/(3+4) = 0.85, S24= (2*2)/(3+3) = 0.66, S34= (2*2)/(4+3) = 0.57

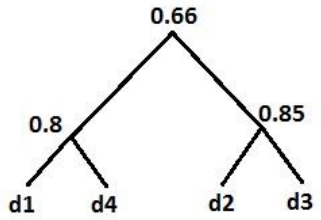
S :

1	0.4	0.33	0.8
-	1	0.85	0.66
-	-	1	0.57
-	-	-	1

Sort these numbers:

1	d2, d3	0.85
2	d1, d4	0.8
3	d2, d4	0.66
4	d3, d4	0.57
5	d1, d2	0.4
6	d1, d3	0.33

Then our dendrogram must be like below:



We can see that most probable two clusters is (d1, d4) and (d2, d3) which is same with the question no.4 result with C3M solution.