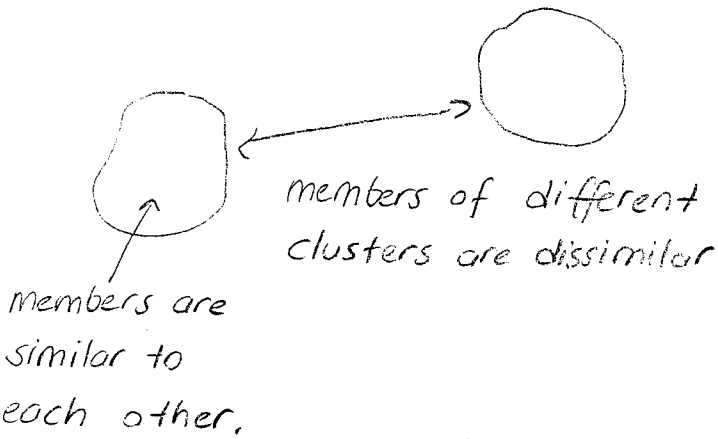
15.03.2019

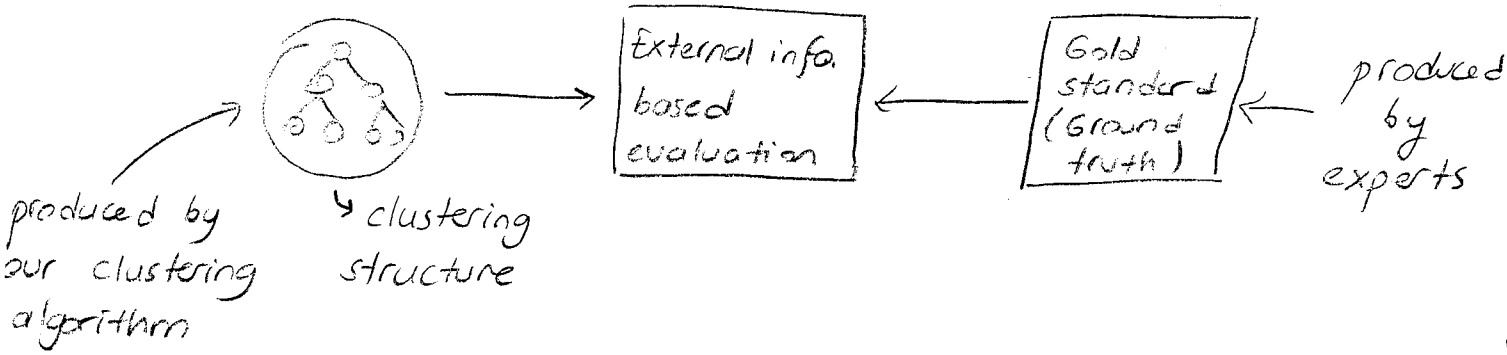
# Cluster Validation / Cluster Quality Measurement

Manning, Raghavan, Schütze
Introduction to Information Retrieval

1. Internal Criterion :



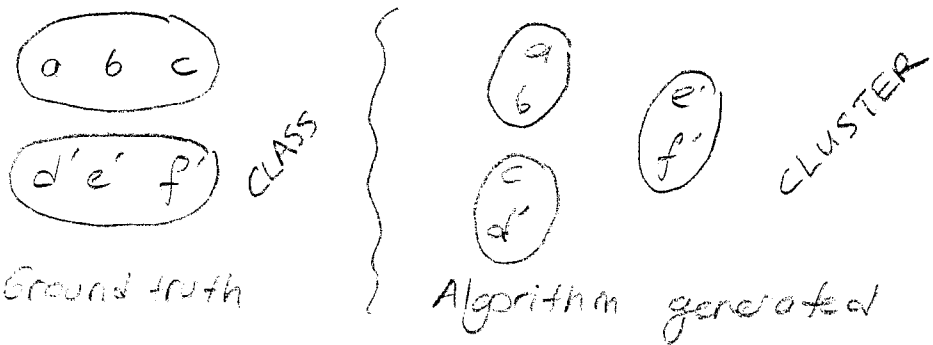

2. External Criterion :




## Rand Coefficient

Consider pairs of documents and find percentage of correct decisions,

Correct clustering structure



$$\binom{6}{2} = \frac{6!}{2!\,4!} = 15$$

| | a | b | c | d' | e' | f' |
|---|---|---|---|---|---|---|
| a | – | ab | ac | ad' | ae' | af' |
| b | – | – | bc | bd' | be' | bf' |
| c | – | – | – | cd' | ce' | cf' |
| d' | – | – | – | – | de' | df' |
| e' | – | – | – | – | – | ef' |
| f' | – | – | – | – | – | – |

| Pair | ab | ac | ad' | ae' | af' | bc | bd' | be' | bf' | cd' | ce' | cf' | de' |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Decision Type | TP | FN | TN | TN | TN | FN | TN | TN | TN | FP | TN | TH | FN |

| df' | ef' |
|---|---|
| FN | TP |

True positive (TP) : two similar objects are assigned to the same cluster.

True negative (TN) : two dissimilar objects are assigned to two different clusters.

False positive (FP) : two dissimilar objects are assigned to the same cluster

False negative (FN) : two similar objects are assigned to different clusters.

$$\text{Rand coefficient} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{2+8}{15} = \underline{\underline{0.66}}$$

| | same cluster | different clusters | |
|---|---|---|---|
| same class | TP: 2 | FN: 4 | 6 |
| different classes | FP: 1 | TN: 8 | 9 |
| | 3 | 12 | 15 |

Corrected Rand:
see Jain & Dubes
Clustering Algorithms for Data
p. 177

# F-Measure

van Rijsbergen, Information Retrieval

$$F = \frac{2PR}{P+R}$$

$$\text{Precision } (P) = \frac{TP}{TP+FP} = 2/3$$

$$\text{Recall } (R) = \frac{TP}{TP+FN} = 2/6$$

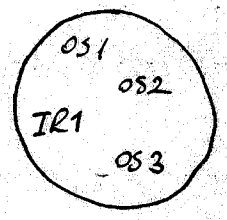
$$F = \frac{2 \cdot \frac{2}{3} \cdot \frac{2}{6}}{\frac{2}{3} + \frac{2}{6}} = \frac{4}{9} = 0.44$$

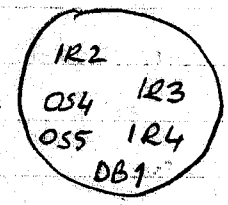
17.07.2010

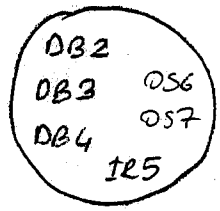
**Purity**: Each cluster is assigned to the class which is most frequent in the cluster, count the no. of correct assignments and divide it by the total no. of elements.

Example: DB, IR, OS

Clusters



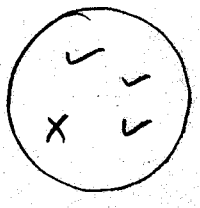



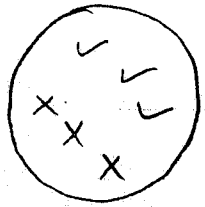

IR : 1
OS : 3
DB : 0

IR : 3
OS : 2
DB : 1

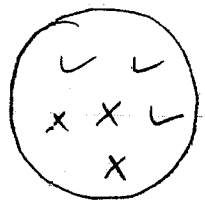
IR : 1
OS : 2
DB : 3



OS cluster



IR cluster



DB cluster

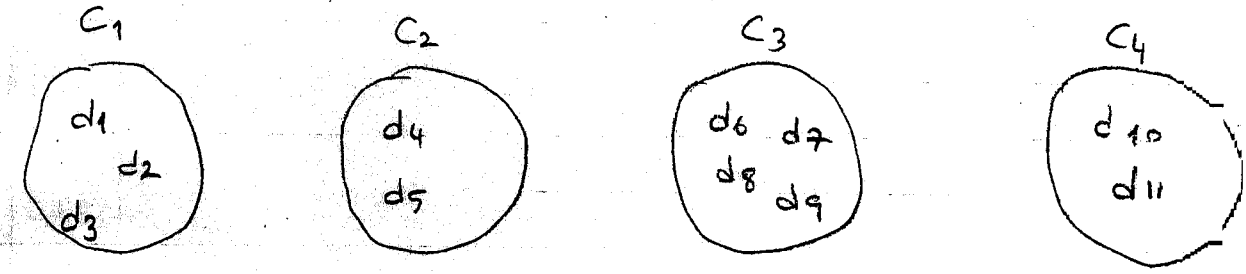
$$Purity = \frac{3+3+3}{4+6+6} = \frac{9}{16} \simeq \underline{0.56}$$



This is practical for small collections, Because we can identify the class of each document. This approach is not scalable.

## Cluster validation using Yao's Formula

This approach is scalable.
It needs an IR collection. → documents
queries
relevant documents for queries




relevant docs
$q_1 \rightarrow d_1, d_5, d_6, d_7$
$q_2 \rightarrow d_1, d_2, d_6, d_{11}$

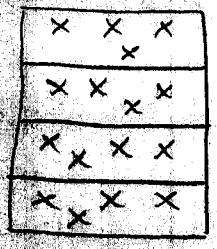
* Target cluster for a query contains at least one relevant document.

Find the target cluster for $q_1$? $C_1, C_2, C_3$
for $q_2$? $C_1, C_3, C_4$

Average number of target clusters = $\frac{3+3}{2} = 3$

This number should be small. How small? It should be at least smaller than the random case.

Yao's formula:
$q \rightarrow$ relevant docs
Approximating block accesses in database organizations
S.B. Yao
Comm. of the ACM
1977



Database

← disk page (disk block)

$n$ = no. of records

$m$ = no. of blocks

block size = $n/m$

$k$ = no. of records to be accessed

$p = n/m$ no. of records in the $j$th block

$n-p$ = no. of records in the other blocks

$C_k^n$ : # of combinations that we can have if we try to select $k$ records out of $n$ records.

$$C_k^n = \frac{n!}{k!(n-k)!} \qquad C_2^3 = \frac{3!}{2!(3-2)!} = 3 \qquad \frac{a,b,c}{ab,\ ac,\ bc}$$

$C_k^{n-p}$ : different ways of selecting $k$ documents from $(n-p)$ documents

The probability that no documents are selected from the $j$th block $\rightarrow$ $C_k^{n-p} / C_k^n$

Let $d = 1 - \frac{1}{m}$

$n - p = n - n/m = n\left(1 - 1/m\right) = nd$

$E(I_j)$ = probability of selecting at least a record from the $j$th block

$= 1 - C_k^{nd} / C_k^n$

Expected number of blocks to be accessed $= \sum_{j=1}^{m} E(I_j) =$

$= m \times \left(1 - C_k^{nd} / C_k^n\right)$

$= m \times \left[1 - \frac{\frac{nd!}{k!(nd-k)!}}{\frac{n!}{k!(n-k)!}}\right] =$

$$= m \times \left[ 1 - \frac{nd!}{k!\,(nd-k)!} \cdot \frac{k!\,(n-k)!}{n!} \right]$$

$$= m \times \left[ 1 - \frac{1.2.3.\ldots nd}{1.2.\ldots (nd-k)} \times \frac{1 \ldots (n-k)}{1.2.\ldots n} \right]$$

$$\frac{(nd-k+1)(nd-k+2)\ldots nd}{1} \qquad \frac{1}{(n-k+1)(n-k+2)\ldots n}$$

$$= m \times \left[ 1 - \prod_{i=1}^{k} \left( \frac{nd-i+1}{n-i+1} \right) \right]$$

$m$ = no. of records

$m_j = m - |C_j|$ where $|C_j|$ is the no. of records in $C_j$

$$P_j = \left[ 1 - \prod_{i=1}^{k} \frac{m_j - i + 1}{m - i + 1} \right]$$

No. of clusters to be accessed $= \sum_{j=1}^{n} P_j$

example: $k = 3$

$m = 100$ $\qquad m_j = 100 - 5 = 95$

$|C_j| = 5$

$$P_j = \left[ 1 - \left( \frac{95-1+1}{100-1+1} \right) \left( \frac{95-2+1}{100-2+1} \right) \left( \frac{95-3+1}{100-3+1} \right) \right]$$

$$= 1 - \left( \frac{95}{100} \right) \left( \frac{94}{99} \right) \left( \frac{93}{98} \right)$$

$$\approx 1 - 0.86 = \underline{0.14}$$

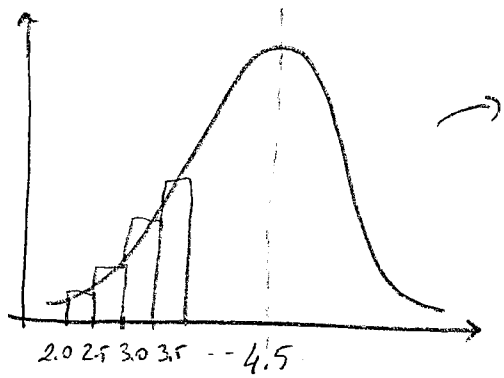
$n_{tr}$ = no. of target clusters to be accessed for random clustering
$n_t$ = no. of target clusters to be accessed.

If $n_t < n_{tr}$ then our clustering structure is <u>possibly</u> valid

else our clustering structure is no good.

If this is true, now show that the difference between $n_{tr}$ and $n_t$ is significant.

Find the baseline for $n_{tr}$ ;
* Monte Carlo
  - Keep the clustering structure the same (i.e. keep the number of clusters the same and the cluster sizes the same)
  - Assign documents randomly to clusters
  - Obtain $n_{tr}$ value
  - do it all again




→ probability density function for $n_{tr}$