13.02.2012

STAIRS

IBM search engine

How to rank?

Value of a term in

a retrieved document $= f(a, b, c)$

$a \rightarrow$ the freq. of the term in the doc.

$b \rightarrow$ the freq. of the term in the retrieved set

$c \rightarrow$ the no. of docs in the retrieved set in which the term appears

Value of a term = $\boxed{\dfrac{a \times b}{c}}$

Score of a document

$\sum$ value of all query terms which appear in the document

example

$a = 16$ ⇒ appears this many no. of times in this particular document

$b = 1247$ appears this many no. of times in the retrieved documents

$c = 152$ it appears in this many no. of distinct documents

$$\text{Value} = \frac{16 \times 1247}{152} = 131{,}26$$

# of docs. that contain the term

$$tf \cdot idf \leftarrow \log\left(\frac{n}{N}\right) + 1$$

no. of occurences in a particular document

total # of docs.

Van Rijsbergen 2 chapters
↳ Information Retrieval

Evaluation

Effectiveness

Efficiency

↳ space & time

Search engines

ease of use

coverage of database

upto date

no broken links

response time

resource requirements

Effectiveness

How to measure?

traceval

TREC = Text Retrieval Conference

Precision

Recall

↳ Assumes the availability of a test collection

- a set of documents
- a set of queries
- a set of relevant documents for each queries

$$\text{Recall} = \frac{\#\text{ of retrieved and relevant docs.}}{\text{Total }\#\text{ of relevant docs in collection}}$$

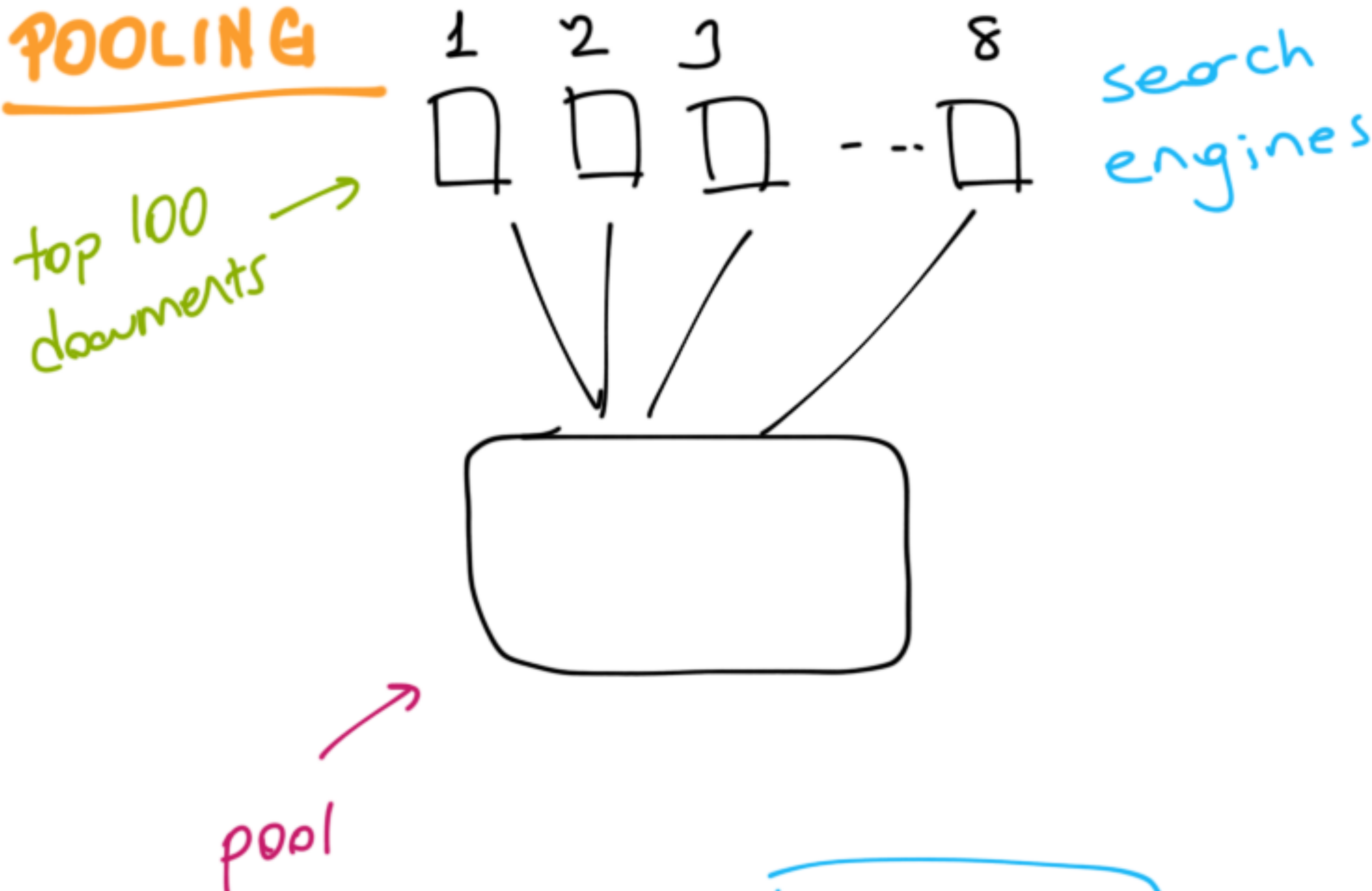
$$\frac{a}{a+c}$$

$$\text{Precision} = \frac{\#\text{ of retrieved and relevant docs.}}{\text{Total }\#\text{ of retrieved docs.}}$$

$$\frac{a}{a+b}$$

	Relevant	Not relevant
Retrieved	a	b
Not Retrieved	c	d

POOLING



Is pooling reliable?

Problem of pooling

Pooling assumes that documents which are not evaluated are not relevant.

They may be relevant.

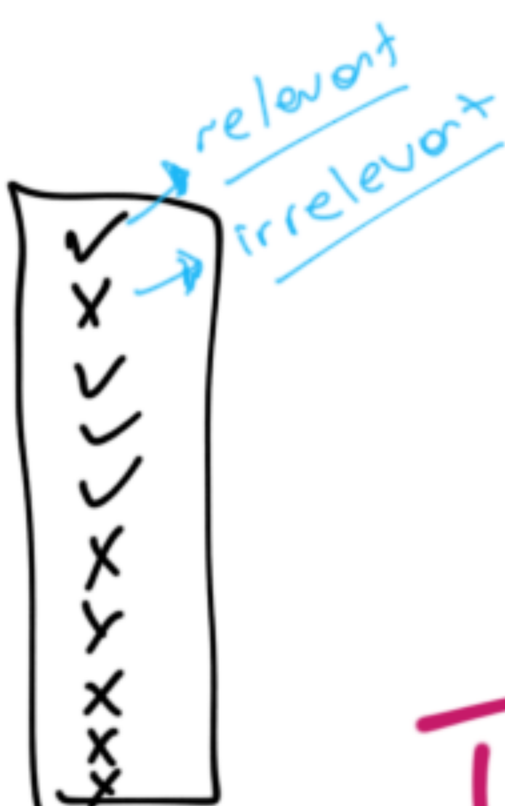
15.02.2012

Precision proportion of retrieved documents which are relevant.

Recall proportion of relevant documents retrieved

example

For a given query let's assume that we have all together 20 relevant documents in the collection.



First 10 result

$$P@10 \rightarrow P = \frac{4}{10} = 0.4$$

$$R@20 \rightarrow R = \frac{4}{20} = 0.2$$

$$\boxed{\text{F-measure} = \frac{2PR}{P+R}}$$

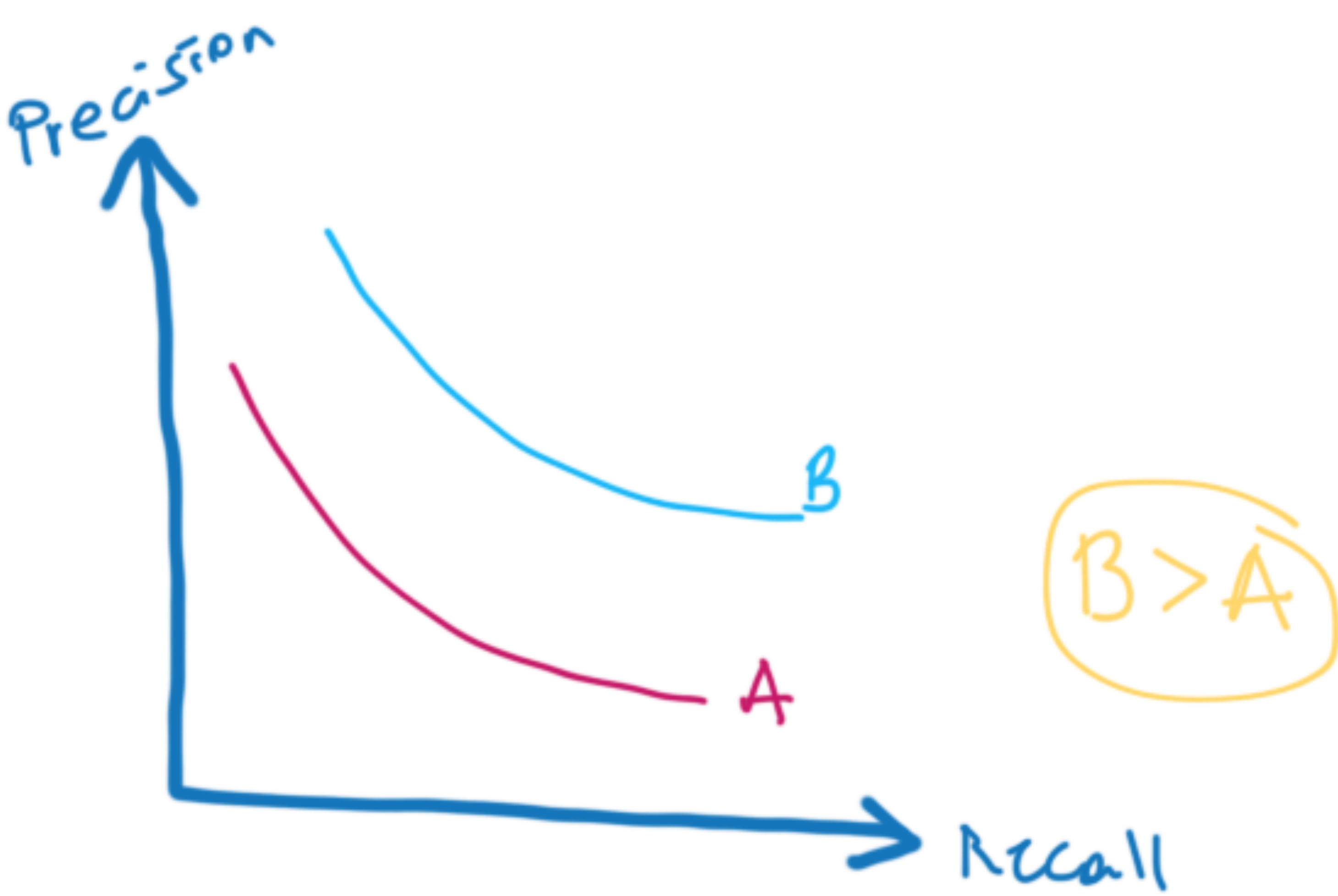
harmonic mean

A document is decided as relevant or irrelevant by pooling.

b-pref (Binary Preference)

It is an effectiveness measure which ignores documents which are not assesed by annotators.

annotators → pooling'de query'nin gönderildiği search engine'ler.



Rank	1	2	3	4	5	6	7	8	9	10
Relevance	0	1	0	1	1	1	1	0	0	1
Precision	0/1	1/2	1/3	2/4						
Recall	0/10	1/10	1/10	2/10						

Total relevance = 10

Similarity Measures and Calculations

Q : D ← query document matching

matching func.

$s(d_i, d_j) = s(d_j, d_i)$ symmetric

similarity of document i to document j.

not normalized

	Binary	Weighted
Inner product	$\|X \cap Y\|$	$\sum_{i=1}^{n} x_i \cdot y_i$
Dice coefficient	$\frac{2\|X \cap Y\|}{\|X\| + \|Y\|}$	$\frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$

	Binary	Weighted
Cosine Coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i}$

example (binary)

$X = (1\ 0\ 1\ 1\ 1)$ $\quad |X| = 4 = \sum x_i$

$Y = (1\ 1\ 0\ 1\ 0)$ $\quad |Y| = 3$

monotonic w/ respect to each other

Inner product = 2

Dice coef. $= \frac{2 \times 2}{4+3}$

Cosine coef. $= \frac{2}{\sqrt{4} + \sqrt{3}}$

example (weighted)

$X = (2 \quad 0 \quad 1 \quad 3 \quad 2)$

$Y = (1 \quad 0 \quad 2 \quad 1 \quad 5)$

$$\text{Dice coef.} = \frac{2 \times (2\times1 + 1\times2 + 3\times1 + 2\times5)}{(4+1+9+4) + (1+4+1+25)} \cong 0.69$$

$$\text{Cosine coef.} = \frac{(2\times1 + 1\times2 + 3\times1 + 2\times5)}{[(18)\cdot(31)]^{1/2}} = 0.72$$

Cluster hypothesis

documents similar to each other are relevant to the same query.

Construction of Similarity (proximity) Matrices

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$S = \begin{bmatrix} 1.0 & S_{12} & S_{13} & S_{14} & S_{15} \\ \times & 1.0 & S_{23} & S_{24} & S_{25} \\ \times & \times & 1.0 & S_{34} & S_{35} \\ \times & \times & \times & 1.0 & S_{45} \\ \times & \times & \times & \times & 1.0 \end{bmatrix}$$

n: # of objects

$$(n-1)+(n-2)\ldots 1 = \frac{n(n-1)}{2} = O(n^2)$$

Most documents don't have a common term.
Calculate similarity between docs that contain a common term.

↑ avoid unnecessary calculations

How to calculate similarities using the knowledge of term distribution in documents?

$t_1 \rightarrow d_1, d_2$

$t_2 \rightarrow d_1, d_2$

$t_3 \rightarrow d_4, d_5$

$t_4 \rightarrow d_2, d_5$

$t_5 \rightarrow d_1, d_2$

$t_6 \rightarrow d_3, d_4, d_5$

$\Longrightarrow$

Consider d_1

t_1, t_2, t_5

$\{d_1, d_2\} \vee \{d_1, d_2\} \vee \{d_1, d_2\} = (d_1, d_2)$

Just calculate s_{12}

Consider d_2

$t_1 \; t_2 \; t_3 \; t_5$

$\{d_1, d_2\} \vee \{d_1, d_2\} \vee \{d_2, d_3\} \vee \{d_1, d_2\} = (d_1, d_2, d_3)$

Just calculate s_{23} ?

How to calculate similarity btw 2 docs?

$X: \boxed{5} \boxed{7} \boxed{8} \boxed{\ldots}$

$Y: \boxed{1} \boxed{3} \boxed{10} \boxed{\ldots}$

should be sorted!

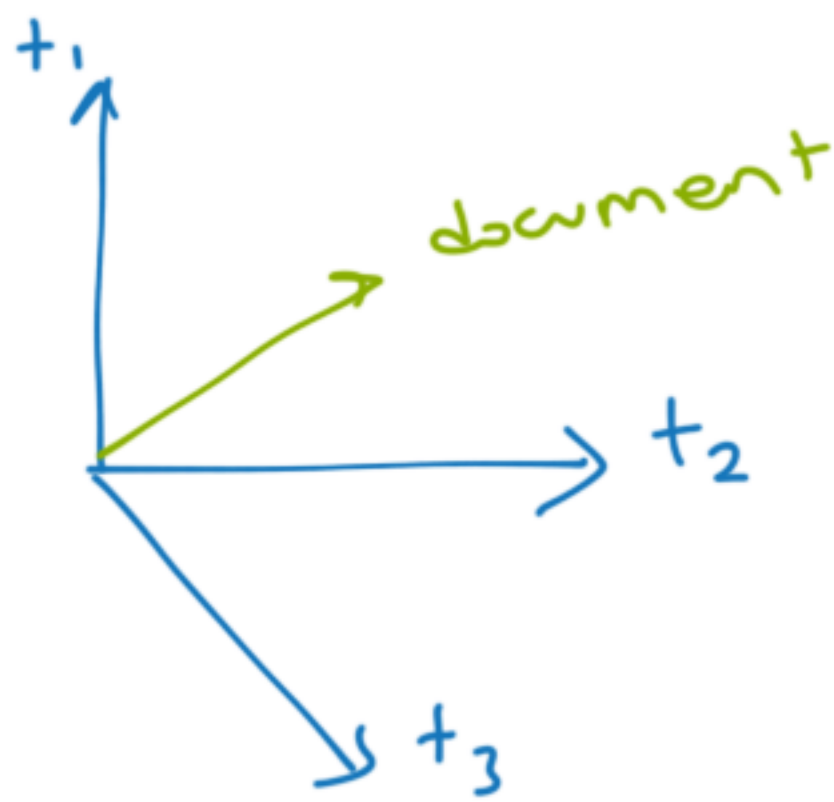
documents

terms appear in that doc.

20.02.2012

Similarity Matrix Calculation

Vector space



1. Straightforward approach
   Compare all documents with each other.

2. Calculate similarity among documents that contain at least one common term.

3. Use inverted indexes

$$D = \begin{array}{cc} & \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{matrix} \\ \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} & \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

posting list

$t_1 \rightarrow \langle 1,1 \rangle \ \langle 2,1 \rangle$

$t_2 \rightarrow \langle 1,1 \rangle \ \langle 2,1 \rangle$

$t_3 \rightarrow \langle 4,1 \rangle \ \langle 5,1 \rangle$

$t_4 \rightarrow \langle 2,1 \rangle \ \langle 5,1 \rangle$

$t_5 \rightarrow \langle 1,1 \rangle \ \langle 2,1 \rangle$

$t_6 \rightarrow \langle 3,1 \rangle \ \langle 4,1 \rangle \ \langle 5,1 \rangle$

Dice coef. $\rightarrow \dfrac{2|X \cap Y|}{|X| + |Y|}$

Document length info.

d_1	d_2	d_3	d_4	d_5
3	4	1	2	3

Consider d_1

contains t_1, t_2, t_3

	S_{12}	S_{13}	S_{14}	S_{15}
X	0	0	0	0

mail box

process t_1

X	1	0	0	0

process t_2

X	2	0	0	0

process t_3

X	3	0	0	0

similarity of a document with the rest of the collection

$$\frac{2 \times 3}{3+4} = \frac{6}{7}$$

Consider d_2

S_{21}		S_{23}	S_{24}	S_{25}
X	X	0	0	0

↳ already calculated in the prev. doc.

$$\left.\begin{array}{l} t_1 \\ t_2 \end{array}\right\rangle \text{ no change}$$

$t_4 \rightarrow$

X	X	0	0	1

$$s = \frac{2\times 1}{4+3} = \frac{2}{7}$$

$\downarrow$

...

x_d: depth of indexing
=
avg. # of terms per document

t_g: term generality
=
avg # of documents per term

$$\underbrace{x_d \cdot t_g}_{d_1} + \underbrace{x_d \cdot t_g}_{d_2} + \text{———} \;\text{- -}\; + \underbrace{x_d \cdot t_g}_{d_m}$$

m: # of docs
n: # of terms

⇒ If we store posting lists in the decreasing order of doc. numbers

$$x_d \cdot t_g \cdot \frac{m-1}{m} + x_d \cdot t_g \cdot \frac{m-2}{m} + \ldots + x_d \cdot t_g \cdot \frac{1}{m}$$

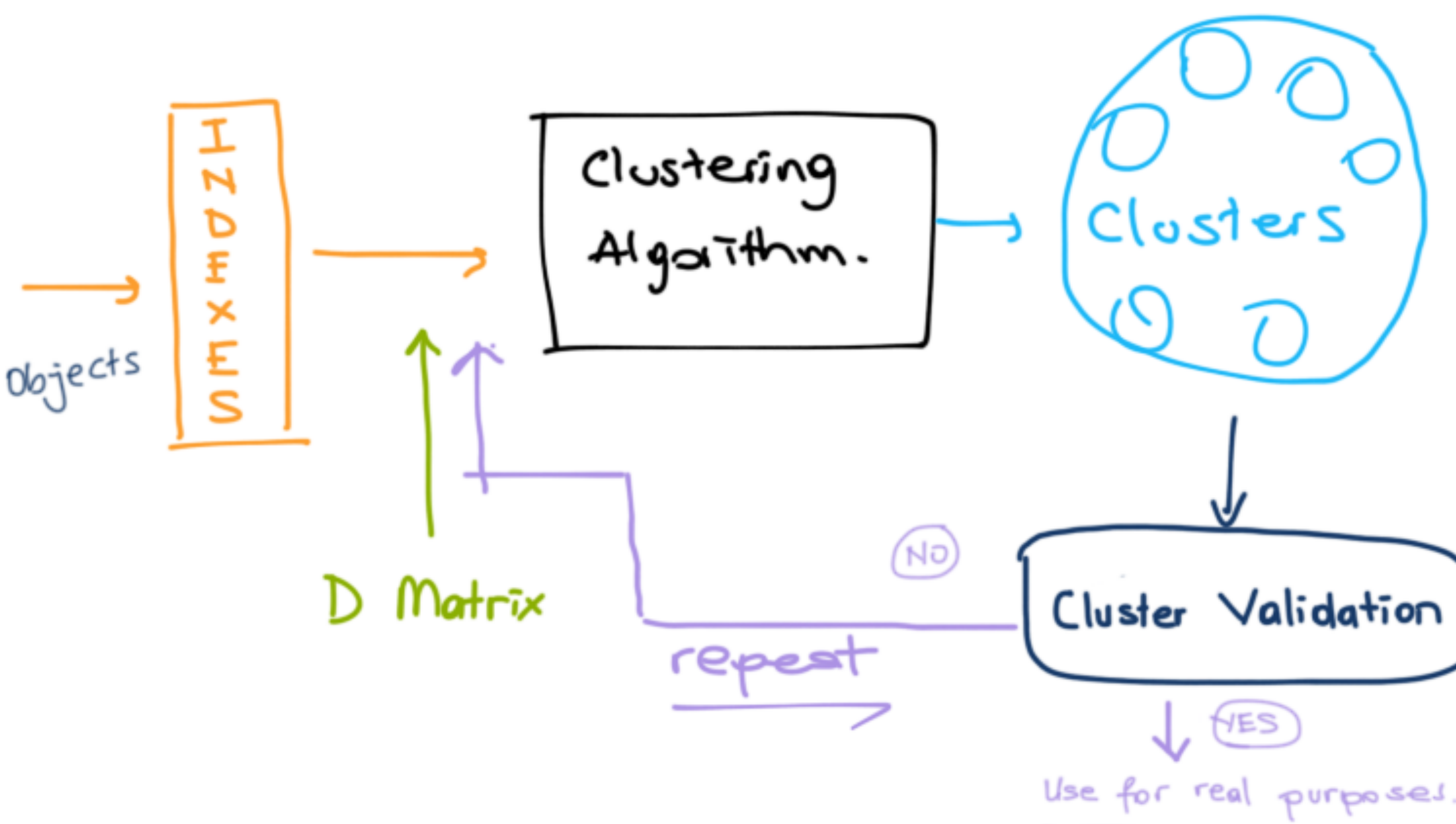
22.02.2012

Document Clustering

Clustering is grouping objects into groups that contain similar objects.

unsupervised

(without using previous knowledge)



How to use clusters?

1. Choose the best cluster and present all of its contents to the user.

2. Cluster-based retrieval:
   Choose a number of best-matching clusters and rank their documents according to their similarity to the query.

3. Use clustering for browsing

What is our expectation?

- increase efficiency
- increase effectiveness

Classification of Clustering Algorithms

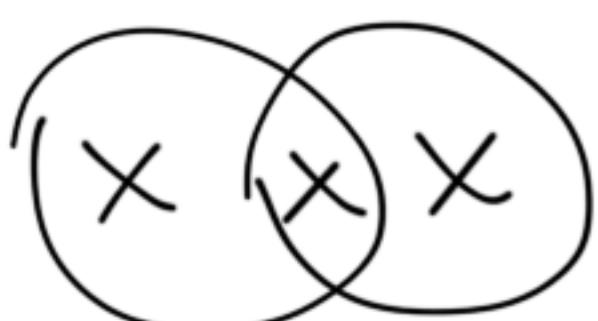
A) According to structure

1. Partitioning $\quad c_i \cap c_j = \emptyset$

$i \neq j$

clusters do not have common members.

2. Overlapping

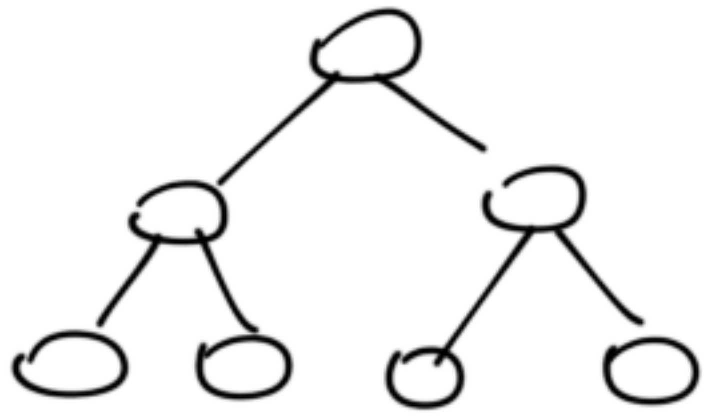


$c_i \cap c_j \neq \emptyset$

fuzzy version

Objects can be a member of different clusters with a certain possibility.

3. Hierarchical



B. According to the working principles of Algorithms

1. Single pass
2. Multi pass
3. Graph threotical

   (nodes and similarity or link among nodes)

4. Using user queries

Single pass Algorithms

a) seed oriented approach

Choose some objects as cluster initiators (cluster seeds)

Assign non-seed objects to cluster seeds.

How many clusters (seeds)?

$n_c = \sqrt{m}$

← # of objects

# of clusters

How to assign objects to seeds?

- choose a similarity measure
- in overlapping clusters choose a threshold based on this threshold, we may assign an object to more than one seed.

we may need to generate a rag bag cluster for mis-fits.

How to choose seeds?

1. Choose objects that contain most # of keywords.

Clusters may not be seperated from each other.

2. Seeds that contain most # of unique keywords.

Seperates clusters from each other.

3. random

4. Use synthetic seeds

It brings supervision

5. Use inverted indexes

b. Heuristic apprCaches

From van Rijsbergen

1. process objects one by one in the order they are numbered
2. the first one starts its own cluster.
3. 2nd or later objects are compared with already existing clusters and if they are different enough they start their own cluster.

Questions

1. With what similarity value are we going to assign an object to a cluster?
2. How to define cluster representations (centroids)

   first object
   last object
   average object

order dependent

no problem if we are processing news articles or if we are doing new event detection.

an article which is different from all existing articles.

3. What are the characteristics of a good clustering algorithm?

1. order independence
2. cat-clustering algorithms

=

non-parametric

effectiveness
efficiency

3. robustness

errors in input do not affect the result.

4. Stable

new additions do not significantly change the output.

5. easy to maintain

Multi pass Algorithms

1. Generate an initial clustering structure using a quick & dirty method.

2. Try to make cluster better by iterative processing.

$\boxed{\text{k-means}}$

Graph Theoretical Clustering

Agglomerative

up

obtain initial clusters
using the objects then
glue (join) existing clusters
to obtain cluster of clusters.
(super clusters)

bottom

single-link, complete-link, avg.-link

Single-link

The similarity between a pair of clusters is taken to be the similarity between the most similar pair of objects, one of which appears in each cluster; thus each cluster member will be _ _ similar to at least one member that some cluster than to any members of another clustes.

$$S = \begin{bmatrix} 1.0 & 0.3 & 0.5 & 0.6 \\ - & 1.0 & 0.4 & 0.5 \\ - & - & 1.0 & 0.3 \\ - & - & - & 1.0 \end{bmatrix}$$

Pair	Similarity
AD	0.6
AC	0.5
BD	0.5
BC	0.4
AB	0.3
CD	0.3

Complete-link

The similarity between the least similar pair of items from the two cluster is used as the cluster similarity. Each cluster member is more similar to the most dissimilar member of its own cluster than the most similar member of any other cluster.

5.3.2012

Seed based

Method: Calculate n_c (# of clusters)

Find n_c no. of cluster seeds

Assign non-seed documents to cluster seeds.

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Inverse of row nums

$\alpha_1 = 1/3$, $\alpha_2 = 1/4$
$\alpha_3 = 1/3$, $\alpha_4 = 1/2$
$\alpha_5 = 1/2$

Columns

$\beta_1 = 1/4$, $\beta_2 = 1/1$
$\beta_3 = 1/2$, $\beta_4 = 1/2$
$\beta_5 = 1/2$, $\beta_6 = 1/3$

$$S = \begin{bmatrix} 1/3 & 0 & 0 & 1/3 & 0 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ & & \vdots & & & \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \end{bmatrix}$$

$$S' = \begin{bmatrix} 1/4 & 0 & & 1/3 \\ 1/4 & 1 & & 0 \\ 1/4 & 0 & \cdots & 1/3 \\ 0 & 0 & & 1/3 \\ 1/4 & 0 & & 0 \end{bmatrix}$$

$$S'^{T} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & & & & \\ 0 & & & & \\ 1/2 & & & & \\ 0 & & & & \\ 1/3 & & & & \end{bmatrix}$$

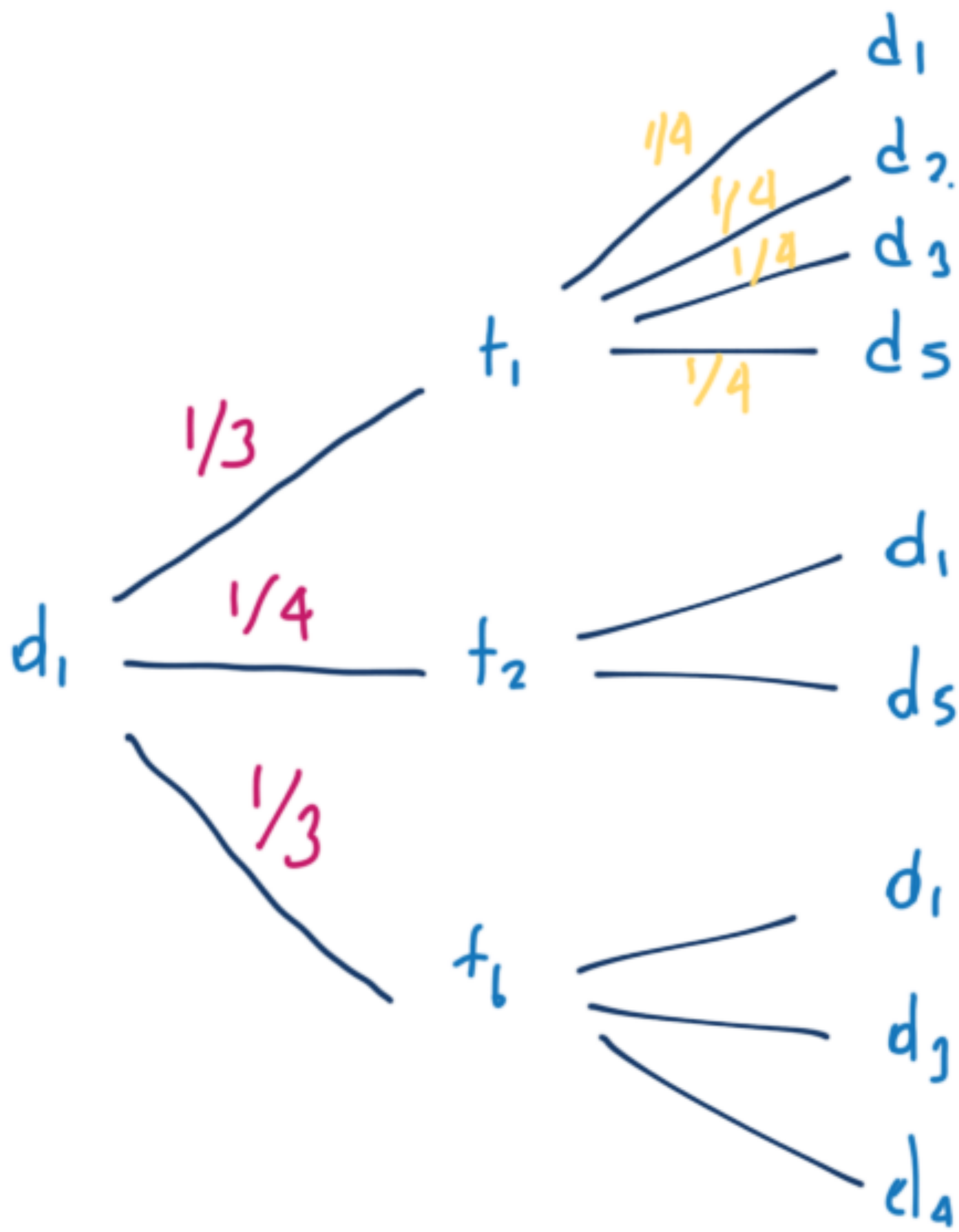
$$C = S \cdot S'^{T}$$

Cover Coefficient

$$C_{m \times m}$$

$$\begin{bmatrix} c_{11} & \cdots & c_{15} \\ c_{21} & \cdots & c_{25} \\ & \vdots & \\ c_{51} & \cdots & c_{55} \end{bmatrix}$$

$$C = \begin{bmatrix} c_{11} = \frac{1}{3} \times \frac{1}{4} + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{3} = \frac{13}{36} = 0.361 \end{bmatrix}$$


d1
1/3
1/4
1/3
t1
t2
t6
1/4
1/4
1/4
1/4
d1
d2
d3
d5
d1
d5
d1
d3
d4

$$C_{ij} = \alpha_i \times \sum_{k=1}^{n} \left(d_{ik} \times \beta_k \times d_{jk} \right)$$

$$C_{11} = 1/3 \times 1/4 + 1/3 \times 1/2 + 1/3 \cdot 1/3$$

C_{ij} ⇒ probability of selecting any term of d_i from d_j.

$$C = \begin{bmatrix} 0.361 & 0.250 & 0.194 & 0.111 & 0.083 \\ \underline{0.188} & \underline{0.563} & \underline{0.063} & \underline{0} & \underline{0.188} \\ 0.194 & 0.083 & 0.361 & 0.277 & 0.083 \\ 0.167 & 0.000 & 0.417 & 0.417 & 0.000 \\ 0.125 & 0.375 & 0.125 & 0.000 & 0.375 \end{bmatrix}$$

$$0 < c_{ii} \leq 1 \;,\; 0 \leq c_{ij} < 1 \;.\; i \neq j$$

$$c_{ii} \geq c_{ij}$$

EKSiK