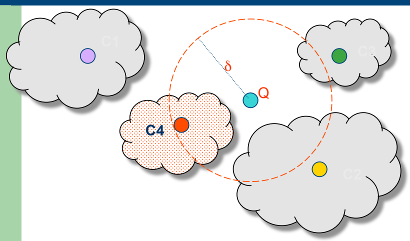
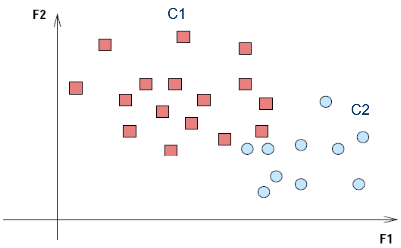
Use of clusters (prune search space)



- …eliminate clusters based on their representatives

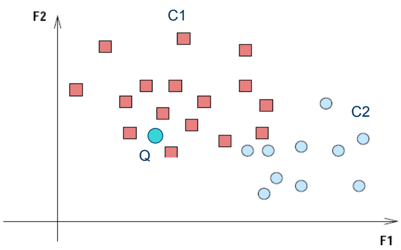
Maria Luisa Sapino (BDM 2018)

Evaluation of clustering methods



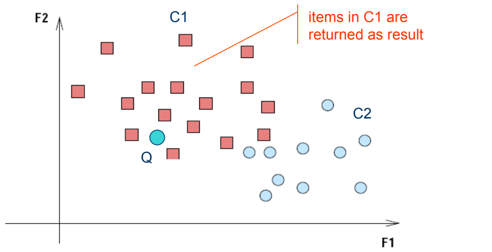
Maria Luisa Sapino (BDM 2018)

Evaluation of clustering methods

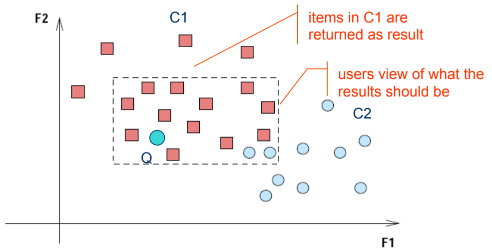


Maria Luisa Sapino (BDM 2018)

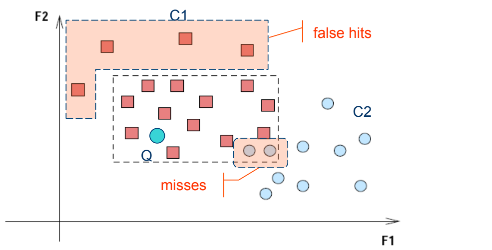
Evaluation of clustering methods



Evaluation of clustering methods

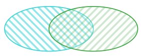


Evaluation of clustering methods



Precision

- Precision

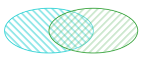


Retrieved and Relevant / Retrieved — measures the effect of false hits

Maria Luisa Sapino (BDM 2018)

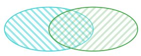
Precision and recall

- Precision



Retrieved and Relevant / Retrieved — measures the effect of false hits

- Recall

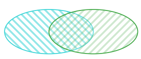


Retrieved and Relevant / Relevant — measures the effect of misses

Maria Luisa Sapino (BDM 2018)

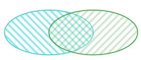
Precision and recall

- Precision



Retrieved and Relevant / Retrieved — measures the effect of false hits

- Recall

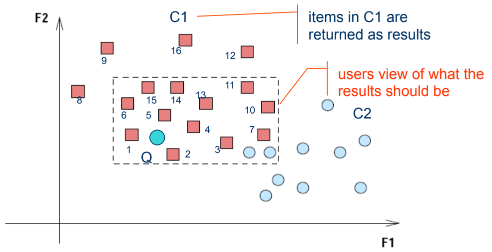


Retrieved and Relevant / Relevant — measures the effect of misses

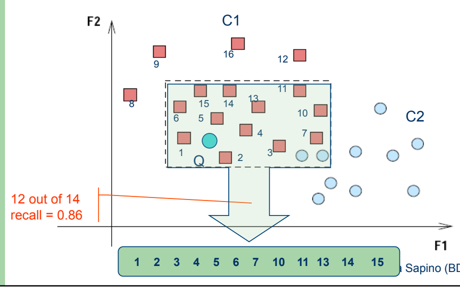
. Both should be closer to 1!!!!

Maria Luisa Sapino (BDM 2018)

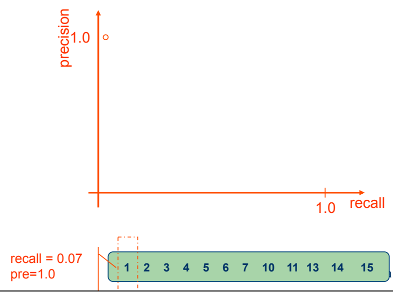
What if we also have rankings in the result???



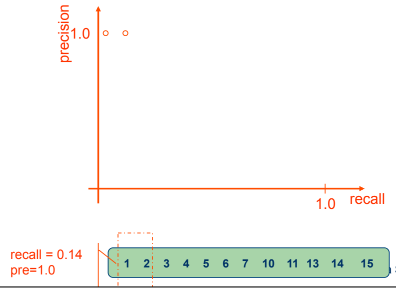
What if we also have rankings in the result???



What if we also have rankings in the result???

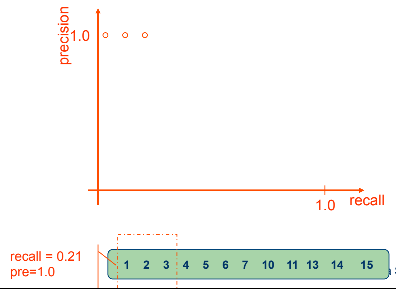


What if we also have rankings in the result???



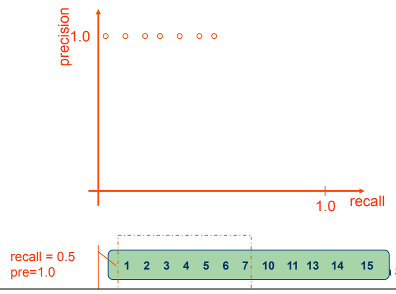
Sapino (BDM 2018)

What if we also have rankings in the result???



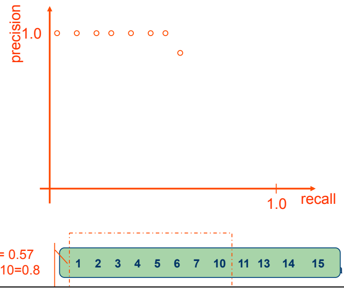
Sapino (BDM 2018)

What if we also have rankings in the result???



Sapino (BDM 2018)

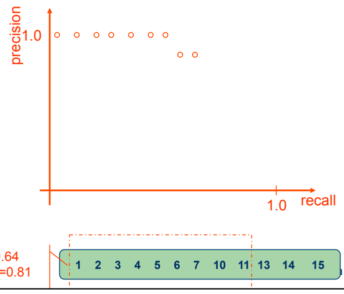
What if we also have rankings in the result???



recall = 0.57
pre=8/10=0.8

Sapino (BDM 2018)

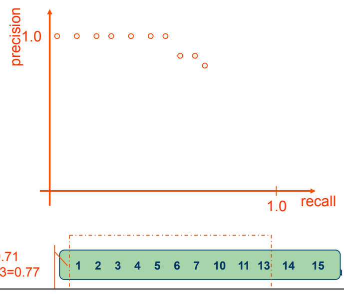
What if we also have rankings in the result???



recall = 0.64
pre=9/11=0.81

Sapino (BDM 2018)

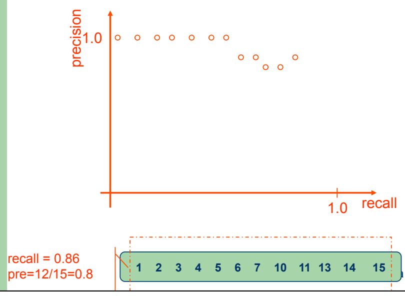
What if we also have rankings in the result???



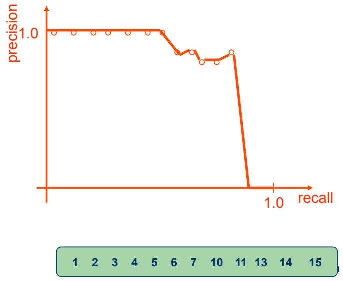
recall = 0.71
pre=10/13=0.77

Sapino (BDM 2018)

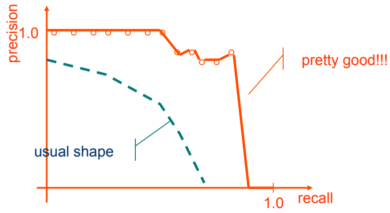
What if we also have rankings in the result???



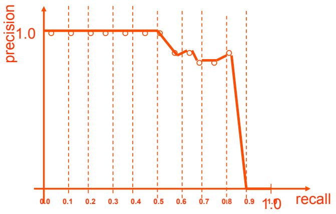
Precision/recall curve



Precision/recall curve



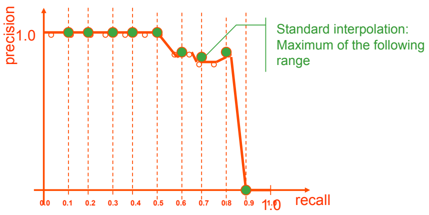
Precision/recall curve



Standard description is with 0.1 increments

Maria Luisa Sapino (BDM 2018)

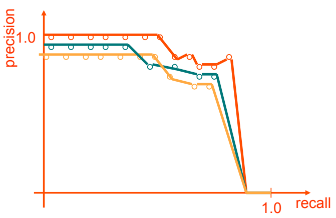
Precision/recall curve



Standard description is with 0.1 increments

Maria Luisa Sapino (BDM 2018)

Precision/recall curve



Given multiple queries: we need to take the average behavior!

Maria Luisa Sapino (BDM 2018)

Single-value summaries

- R-precision
  - # of relevant documents within first R
  - R is the total number of relevant documents in the result

Single-value summaries

- R-precision
  - # of relevant documents within first R
  - R is the total number of relevant documents in the result
- Example:
  - R = 14
  - # of relevant document in the first 14 is 11

    1 2 3 4 5 6 7 10 11 13 14 15

  - R-precision for this query is 11/14 = 0.876

Harmonic mean

- The harmonic mean of *n* numbers (where *i* = 1, ..., *n*) is

$$\frac{1}{H} \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}.$$

- Therefore, harmonic mean of x= P and y= R

H(P,R) = 2PR/ (P+R)

Harmonic mean

- The harmonic mean of n numbers (where i = 1, ..., n) is

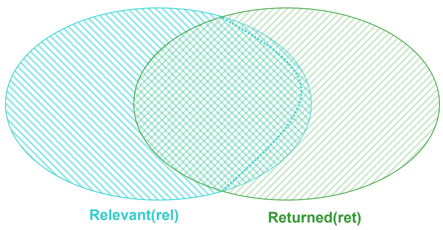
$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}.$$

- Therefore, harmonic mean of x= P and y= R

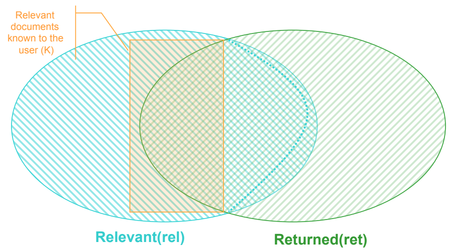
$H(P,R) = 2PR/ (P+R)$

High only when both P and R are high

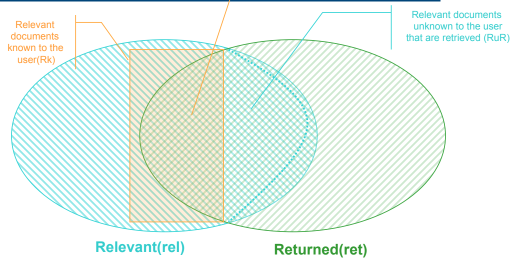
Coverage and Novelty



Coverage and Novelty

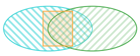


Coverage and Novelty



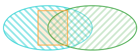
Coverage and novelty

- Coverage



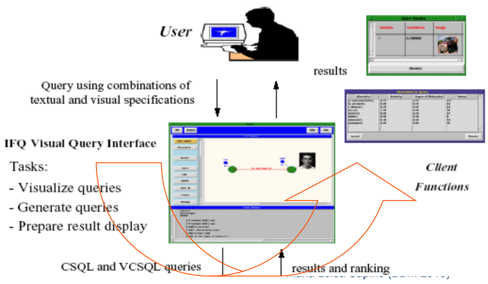
$$\frac{RkR}{Rk}$$

- Novelty

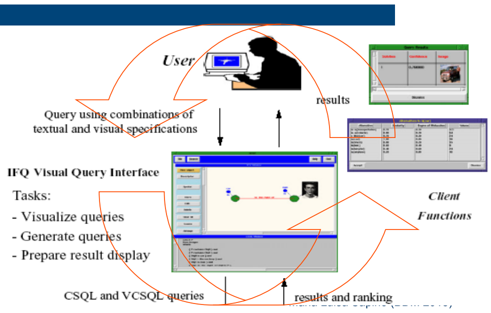


$$\frac{RuR}{RuR + RkR}$$

Relevance feedback



Relevance feedback



Relevance feedback (?)

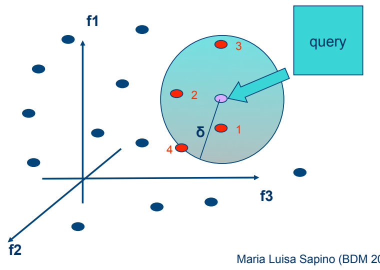
- User submits a query, Q
- System retrieves and ranks a set of objects, S
- User selects a set of relevant (R) and irrelevant (I) objects from S or provides a new ranking
- How can the system improve retrieval results?
  - relevance is subjective
  - distance is objective

Subjectivity vs. objectivity

- *Q*: query
- *R*: event that a document is relevant to the user
- *P(R|Oi)*: given features and weights, the probability that object Oi is relevant to the user
- …then, retrieval is effective iff

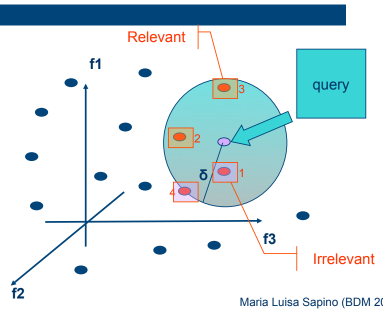
$$p(R \mid O_i) > p(R \mid O_j) \Leftrightarrow sim(Q, O_i) > sim(Q, O_j)$$

Query and results



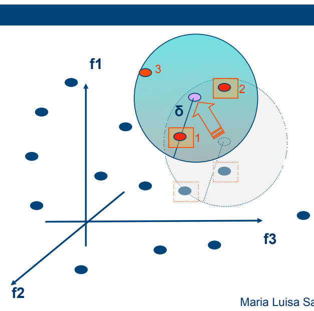
Maria Luisa Sapino (BDM 2018)

Query and results



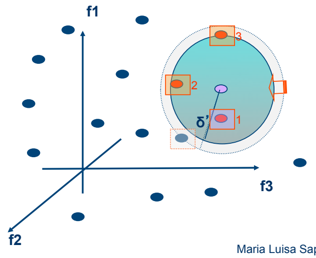
Maria Luisa Sapino (BDM 2018)

..modify query



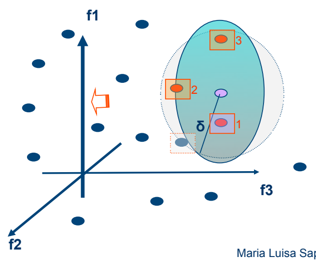
Maria Luisa Sapino (BDM 2018)

…modify radius



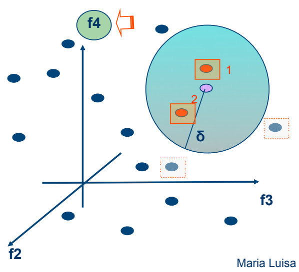
Maria Luisa Sapino (BDM 2018)

…modify importance (weight) of the features



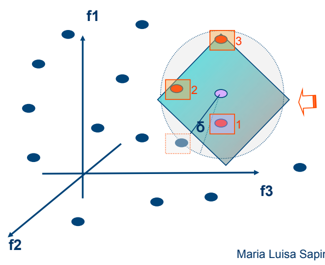
Maria Luisa Sapino (BDM 2018)

..remove add features

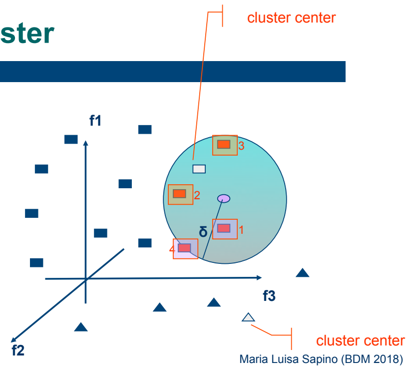


Maria Luisa Sapino (BDM 2018)

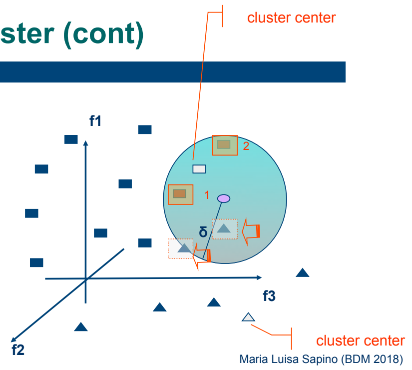
..change the distance measure



..recluster



..recluster (cont)



Basic approach

- An ideal query should separate relevant objects from irrelevant ones

$$sep = \left(\sum_{o_i \in \mathbf{Ir}} dist(Q, o_i) \right) - \left(\sum_{o_i \in \mathbf{Rel}} dist(Q, o_i) \right)$$

should be maximum

Basic approach

- ..assuming linear distance function

$$sep = dist\left(Q, \sum_{o_i \in \mathbf{Ir}} o_i - \sum_{o_i \in \mathbf{Rel}} o_i \right)$$

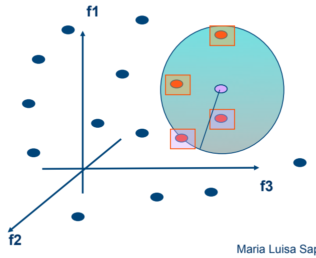
should be maximum

Basic approach

- Move the query
  - closer to the relevant documents and
  - away from irrelevant documents

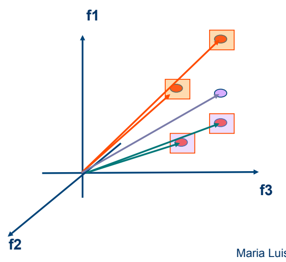
$$Q' = Q + \left(c_{rel} \times \sum_{o_i \in \mathbf{Rel}} o_i \right) + \left(c_{ir} \times \sum_{o_i \in \mathbf{Ir}} o_i \right)$$

Basic Approach



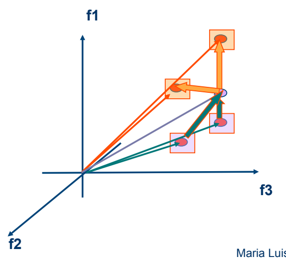
Maria Luisa Sapino (BDM 2018)

Basic Approach



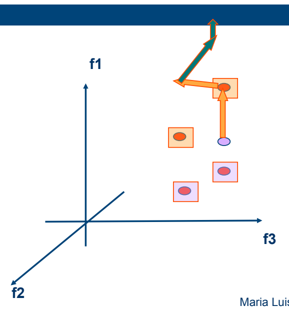
Maria Luisa Sapino (BDM 2018)

Basic Approach



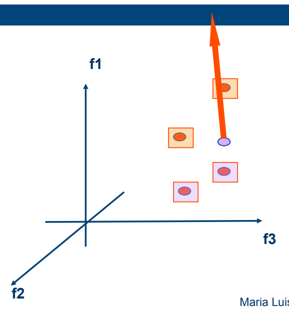
Maria Luisa Sapino (BDM 2018)

Basic Approach



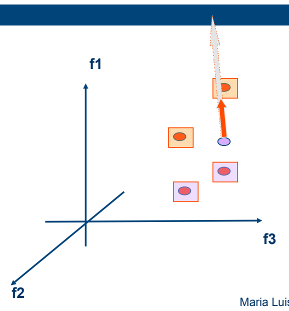
Maria Luisa Sapino (BDM 2018)

Basic Approach



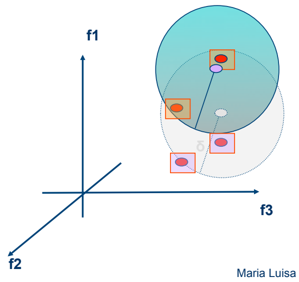
Maria Luisa Sapino (BDM 2018)

Basic Approach



Maria Luisa Sapino (BDM 2018)

Basic Approach



Feature (term) readjustment

- *Q*: query
- *R*: event that a document is relevant to the user
- *P(R|Oi):* given features and weights, the probability that object o_i is relevant to the user
- Significance of a feature f_k is

$$sig(f_k) = \frac{\log\left(\frac{p(f_k \mid R)}{1 - p(f_k \mid R)}\right)}{\log\left(\frac{p(f_k \mid I)}{1 - p(f_k \mid I)}\right)}$$

Feature (term) readjustment

- *Q*: query
- *R*: event that a document is relevant to the user
- *P(R|Oi):* given features and weights, the probability that object o_i is relevant to the user
- Significance of a feature f_k is

$$sig(f_k) = \frac{\log\left(\frac{p(f_k \mid R)}{1 - p(f_k \mid R)}\right)}{\log\left(\frac{p(f_k \mid I)}{1 - p(f_k \mid I)}\right)}$$

How do we estimate these probabilities

Feature (term) readjustment

- Case I
  - If f_k is not a query term (not used in retrieval)

$$p(f_k \mid R) = p(f_k \mid Retrieved \;\&\; Relevant)$$

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

- Case II
  - If f_k is a query term (used in retrieval)

$$p(f_k \mid R) \neq p(f_k \mid Retrieved \;\&\; Relevant)$$

  - There would be bias
    - Most retrieved objects will have f_k

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

- Let us assume binary feature and query
  - $o = \langle f_1, f_2, \ldots f_n \rangle$ f_i = 0 or 1
  - $q = \langle w_1, w_2, \ldots w_n \rangle$ w_i = 0 or 1
- Let us assume dot product as the similarity

$$sim(o,q) = \sum_{i=1}^{n} w_i f_i$$

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

- If a document is returned, then

$$sim(o,q) = \sum_{i=1}^{n} w_i f_i > T$$

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

- Let's focus on a specific feature

$$sim(o,q) = w_j f_j + \sum_{i \in \{1..n\} - \{j\}} w_i f_i > T$$

or

$$sim(o,q) = w_j f_j + sim_{(-j)}(o,q) > T$$

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

$$sim(o,q) = w_j f_j + sim_{(-j)}(o,q) > T$$

	fj=1	fj=0
$sim_{(-j)}(o,q) \leq T$	a	0
$sim_{(-j)}(o,q) > T$	b	c

$$|relevant \,\&\, retrieved| = a + b + c$$

Maria Luisa Sapino (BDM 2018)

Feature (term) readjustment

$$p(f_k \mid \text{Ret} \,\&\, \text{Rel}) = \frac{p\big((f_k = 1) \wedge (sim_{(-k)}(o,q) > T)\big)}{p\big(sim_{(-k)}(o,q) > T\big)}$$

Without bias of fk

Feature (term) readjustment

independent

$$p(f_k \mid \text{Ret} \,\&\, \text{Rel}) = \frac{p\big((f_k = 1) \wedge (sim_{(-k)}(o,q) > T)\big)}{p\big(sim_{(-k)}(o,q) > T\big)}$$

⇓

$$p(f_k \mid \text{Ret} \,\&\text{Rel}) = p(f_k = 1)$$

Feature (term) readjustment

independent

$$p(f_k \mid \text{Ret} \,\&\, \text{Rel}) = \frac{p\big((f_k = 1) \wedge (sim_{(-k)}(o,q) > T)\big)}{p\big(sim_{(-k)}(o,q) > T\big)}$$

⇓

$$p(f_k \mid \text{Ret} \,\&\, \text{Rel}) = p(f_k = 1) = p\big(f_k = 1 \mid sim_{(-k)}(o,q) > T\big)$$

Feature (term) readjustment

$$p(f_k \mid \text{Ret \& Rel}) = \frac{p\left((f_k = 1) \wedge \left(sim_{(-k)}(o,q) > T\right)\right)}{p\left(sim_{(-k)}(o,q) > T\right)}$$

$$p(f_k \mid R) = \frac{b}{b+c}$$

Maria Luisa Sapino (BDM 2018)

Ranking

- *Q*: query
- *R*: event that a document is relevant to the user
- *P(R|Oi)*: given features and weights, the probability that object oi is relevant to the user
- …then, retrieval is effective iff

$$p(R \mid O_i) > p(R \mid O_j) \Leftrightarrow sim(Q, O_i) > sim(Q, O_j)$$

Maria Luisa Sapino (BDM 2018)

Ranking

$$p(R \mid O_i) > p(R \mid O_j) \Leftrightarrow sim(Q, O_i) > sim(Q, O_j)$$

- Let's try to rewrite the first half of the equation using Bayes theorem

Maria Luisa Sapino (BDM 2018)

Bayes Theorem

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Maria Luisa Sapino (BDM 2018)

Bayes Theorem

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B \mid A)\, p(A) + p(B \mid \neg A)\, p(\neg A)}$$

Maria Luisa Sapino (BDM 2018)

Relevance of two objects

$$p(R \mid O_i) = \frac{p(O_i \mid R)\, p(R)}{p(O_i \mid R)\, p(R) + p(O_i \mid I)\, p(I)}$$

>

$$p(R \mid O_j) = \frac{p(O_j \mid R)\, p(R)}{p(O_j \mid R)\, p(R) + p(O_j \mid I)\, p(I)}$$

Maria Luisa Sapino (BDM 2018)

Relevance of two objects

$$\frac{p(O_i \mid R)}{p(O_i \mid I)} > \frac{p(O_j \mid R)}{p(O_j \mid I)}$$

Maria Luisa Sapino (BDM 2018)

…so we have

$$\frac{p(O_i \mid R)}{p(O_i \mid I)} > \frac{p(O_j \mid R)}{p(O_j \mid I)} \Leftrightarrow sim(Q,O_i) > sim(Q,O_j)$$

How do we compute these probabilities???

Maria Luisa Sapino (BDM 2018)

…so we have

- Let us assume features are independent
  - *o= <f1,f2,…..fn>*
  - *q= <w1,w2,…..wn>*

$$p(O_i \mid R) = \prod_{k=1}^{n} p(f_{i,k} \mid R) \qquad p(O_i \mid I) = \prod_{k=1}^{n} p(f_{i,k} \mid I)$$

Maria Luisa Sapino (BDM 2018)

…so we have

- Let us assume features are independent

$$p(O_i \mid R) = \prod_{k=1}^{n} p(f_{i,k} \mid R) \qquad p(O_i \mid I) = \prod_{k=1}^{n} p(f_{i,k} \mid I)$$

$$\frac{p(O_i \mid R)}{p(O_i \mid I)} > \frac{p(O_j \mid R)}{p(O_j \mid I)} \Leftrightarrow \frac{\prod_{k=1}^{n} p(f_{i,k} \mid R)}{\prod_{k=1}^{n} p(f_{i,k} \mid I)} > \frac{\prod_{k=1}^{n} p(f_{j,k} \mid R)}{\prod_{k=1}^{n} p(f_{j,k} \mid I)}$$

Maria Luisa Sapino (BDM 2018)

…so we have

- Let us assume features are independent

$$p(O_i \mid R) = \prod_{k=1}^{n} p(f_{i,k} \mid R) \qquad p(O_i \mid I) = \prod_{k=1}^{n} p(f_{i,k} \mid I)$$

$$\frac{p(O_i \mid R)}{p(O_i \mid I)} > \frac{p(O_j \mid R)}{p(O_j \mid I)} \Leftrightarrow \sum_{k=1}^{n} \log \frac{p(f_{i,k} \mid R)}{p(f_{i,k} \mid I)} > \sum_{k=1}^{n} \log \frac{p(f_{j,k} \mid R)}{p(f_{j,k} \mid I)}$$

Maria Luisa Sapino (BDM 2018)

…so we have

- Let us assume features are independent
  - use dot product as the similarity measure
  - use $\log \frac{p(f_{i,k} \mid R)}{p(f_{i,k} \mid I)}$ as the weight of the k^{th} feature
  - use *<1,1,1,…,1>* as the query!!!

$$\frac{p(O_i \mid R)}{p(O_i \mid I)} > \frac{p(O_j \mid R)}{p(O_j \mid I)} \Leftrightarrow \sum_{k=1}^{n} \log \frac{p(f_{i,k} \mid R)}{p(f_{i,k} \mid I)} > \sum_{k=1}^{n} \log \frac{p(f_{j,k} \mid R)}{p(f_{j,k} \mid I)}$$

Maria Luisa Sapino (BDM 2018)

…what if features are not independent?

- Let us assume features are not independent
  - o= <f1,f2,…..fn>
  - q= <w1,w2,…..wn>

$$p(O_i \mid R) \neq \prod_{k=1}^{n} p(f_{i,k} \mid R) \qquad p(O_i \mid I) \neq \prod_{k=1}^{n} p(f_{i,k} \mid I)$$

Maria Luisa Sapino (BDM 2018)

…what if features are not independent?

- Let us assume features are not independent
  - o= <f1,f2,…..fn>
  - q= <w1,w2,…..wn>
- How can we incorporate term dependence???

Maria Luisa Sapino (BDM 2018)

…so we have

- Let us assume features are not independent
  - o= <f1,f2,…..fn>
  - q= <w1,w2,…..wn>
- How can we incorporate term dependence???
- Degree of approximation…..

$$I(p1, p2) = \sum_x p1(x) \log \frac{p1(x)}{p2(x)}$$

  - *p1=p2* implies that *I=0*
  - *p1<>p2* implies that *I>0*

Maria Luisa Sapino (BDM 2018)

…so we have

- Degree of approximation…..

$$I(p1, p2) = \sum_x p1(x) \log \frac{p1(x)}{p2(x)}$$

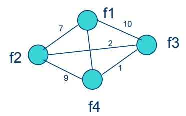
  - *p1=p2* implies that *I=0; p1<>p2* implies that *I>0*

- Degree of dependence between f_i and f_j

$$D_{ij} = I\left(p(f_i \wedge f_j), p(f_i)p(f_j)\right)$$

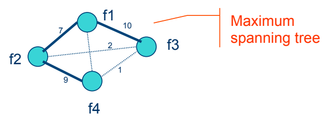
If the two terms are independent, then *Dij* will be 0!!!

Dependence graph



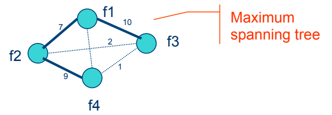
If the two terms are independent, then *Dij* will be 0!!!

Dependence graph



If the two terms are independent, then *Dij* will be 0!!!

Dependence graph

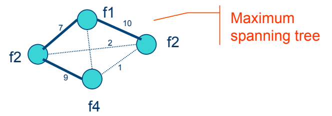


$$p(f_1 \wedge f_2 \wedge f_3 \wedge f_r) = p(f_1)p(f_2 | f_1)p(f_3 | f_1)p(f_4 | f_2)$$

If the two terms are independent, then *Dij* will be 0!!!

Maria Luisa Sapino (BDM 2018)

Dependence graph



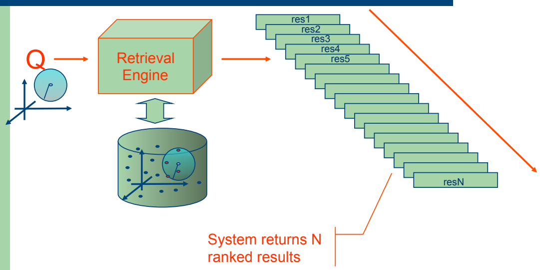
$$p(f_1 \wedge f_2 \wedge f_3 \wedge f_r) = p(f_1)p(f_2 | f_1)p(f_3 | f_1)p(f_4 | f_2)$$

$p(O_i | R)$ can be computed using the distribution of the features in *R*!!!

$p(O_i | I)$ can be computed using the distribution of the features in *I*!!!

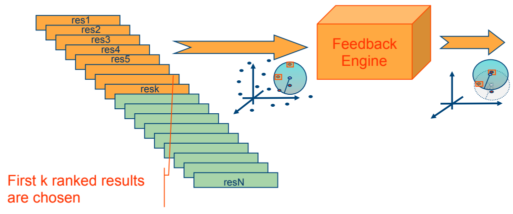
Maria Luisa Sapino (BDM 2018)

Feedback without user's help..



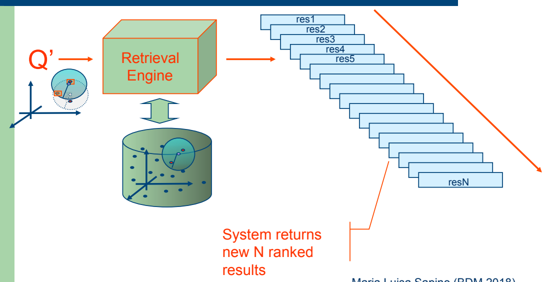
Maria Luisa Sapino (BDM 2018)

Feedback without user's help..



Maria Luisa Sapino (BDM 2018)

Feedback without user's help..



Maria Luisa Sapino (BDM 2018)