How can we estimate the number of clusters in a database?

- …some clustering algorithms require the expected number of clusters as input

How can we estimate the number of clusters in a database?

- …some clustering algorithms requires the expected number of clusters as input
- covering

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k \mid o_i)\, p(o_j \mid k)$$

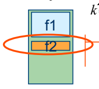
importance of feature f_k in o_i

probability that o_j is a document having feature f_k

How can we estimate the number of clusters in a database?

- …some clustering algorithms requires the expected number of clusters as input
- covering

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k \mid o_i)\, p(o_j \mid k)$$

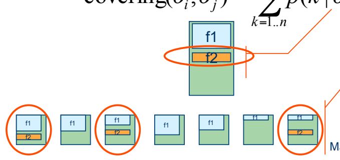


probability that o_j is a document having feature f_k

How can we estimate the number of clusters in a database?

- …some clustering algorithms requires the expected number of clusters as input
- covering

$$\text{covering}(o_i,o_j)=\sum_{k=1..n} p(k\mid o_i)p(o_j\mid k)$$



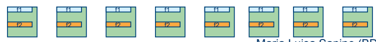
How can we estimate the number of clusters in a database?

$$\text{covering}(o_i,o_j)=\sum_{k=1..n} p(k\mid o_i)p(o_j\mid k)$$

importance of feature f_k in o_i

probability that o_j is a document having feature f_k

- Suppose the database is a perfect cluster
  - features are uniformly distributed
  - all document are equally likely to be selected



How can we estimate the number of clusters in a database?

$$\text{covering}(o_i,o_j)=\sum_{k=1..n} p(k\mid o_i)p(o_j\mid k)$$

importance of feature f_k in o_i

probability that o_j is a document having feature f_k

- Suppose the database is a single cluster

$$\text{covering}(o_i,o_j)=\sum_{k=1..n}\frac{1}{n}\frac{1}{D}=n\frac{1}{n}\frac{1}{D}=\frac{1}{D}$$

How can we estimate the number of clusters in a database?

- Suppose the database is a single cluster

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} \frac{1}{n}\frac{1}{D} = n\frac{1}{n}\frac{1}{D} = \frac{1}{D}$$

- Let's sum up all self-coverings, then

$$\sum_{o_i} \text{covering}(o_i, o_i) = \sum_{D} \frac{1}{D} = 1$$

Maria Luisa Sapino (BDM 2018)

How can we estimate the number of clusters in a database?

- Suppose the database is a single cluster

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} \frac{1}{n}\frac{1}{D} = n\frac{1}{n}\frac{1}{D} = \frac{1}{D}$$

- Let's sum up all self-coverings, then

$$\sum_{o_i} \text{covering}(o_i, o_i) = \sum_{D} \frac{1}{D} = 1$$

Maria Luisa Sapino (BDM 2018)

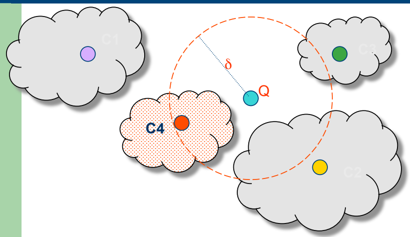
How can we estimate the number of clusters in a database?

$$\sum_{o_i} \text{covering}(o_i, o_i) = p$$

There are approximately p clusters

Maria Luisa Sapino (BDM 2018)

Use of clusters (prune search space)



- …eliminate clusters based on their representatives

Maria Luisa Sapino (BDM 2018)

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

Maria Luisa Sapino (BDM 2018)

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

- Goal: for each cluster, estimate # of documents having *t* or more matching keywords with a query with *k* keywords

Maria Luisa Sapino (BDM 2018)

Use of clusters
Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

$$o_i = \langle f_{i,1}, f_{i,2}, \dots f_{i,n} \rangle;$$

Probability that a document in the cluster has this keyword

$$R_O = \langle r_1, r_2, \dots r_n \rangle = \frac{\sum_{o_i \in O} o_i}{|O|}$$

$$q = \langle 1,1,1,\dots,1,0,0,\dots,0 \rangle; with \quad k \quad 1s$$

Maria Luisa Sapino (BDM 2018)

Use of clusters
Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

$$o_i = \langle f_{i,1}, f_{i,2}, \dots f_{i,n} \rangle; \qquad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle r_1, r_2, \dots r_n \rangle = \frac{\sum_{o_i \in O} o_i}{|O|}$$

$$q = \langle 1,1,1,\dots,1,0,0,\dots,0 \rangle; with \quad k \quad 1s$$

Maria Luisa Sapino (BDM 2018)

Use of clusters
Binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots f_{i,n} \rangle; \qquad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle r_1, r_2, \dots r_n \rangle = \frac{\sum_{o_i \in O} o_i}{|O|}$$

$$p(o^k \in O) = \prod_{j=1}^{k} (r_j)^{f_j} (1 - r_j)^{1-f_j}$$

Maria Luisa Sapino (BDM 2018)

Use of clusters
Binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, .. f_{i,n} \rangle; \qquad o^k = \langle f_1, f_2, ..., f_k \rangle;$$

$$R_O = \langle r_1, r_2, .. r_n \rangle = \frac{\sum_{o_i \in O} o_i}{|O|}$$

$$num(t,Q) = \sum_{o^k \text{ with } t \text{ 1s}} \left(p(o^k \in O) \right)$$

Use of clusters
Non-binary, independent features

- Each document is a non-binary vector
- Documents are organized into clusters
- Each cluster has a representative

- Goal: for each cluster, find the probability that one object in the cluster will be more than S similiar to the query

Use of clusters
Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, .. f_{i,n} \rangle; \qquad o^k = \langle f_1, f_2, ..., f_k \rangle;$$

Use of clusters
Non-binary independent features

$o_i = \langle f_{i,1}, f_{i,2}, \ldots f_{i,n} \rangle;$ $\quad o^k = \langle f_1, f_2, \ldots, f_k \rangle;$

$R_O = \langle [r_1, w_1][r_2, w_2] \ldots, [r_n, w_n] \rangle$

$q = \langle q_1, q_2, \ldots, q_k \rangle$

Probability that a document in the cluster has this keyword

Maria Luisa Sapino (BDM 2018)

Use of clusters
Non-binary independent features

$o_i = \langle f_{i,1}, f_{i,2}, \ldots f_{i,n} \rangle;$ $\quad o^k = \langle f_1, f_2, \ldots, f_k \rangle;$

$R_O = \langle [r_1, w_1][r_2, w_2] \ldots, [r_n, w_n] \rangle$

The average weight of the keyword in the documents that have this keyword

$q = \langle q_1, q_2, \ldots, q_k \rangle$

Probability that a document in the cluster has this keyword

Maria Luisa Sapino (BDM 2018)

Use of clusters
Non-binary independent features

$o_i = \langle f_{i,1}, f_{i,2}, \ldots f_{i,n} \rangle;$ $\quad o^k = \langle f_1, f_2, \ldots, f_k \rangle;$

$R_O = \langle [r_1, w_1][r_2, w_2] \ldots, [r_n, w_n] \rangle$

$q = \langle q_1, q_2, \ldots, q_k \rangle$

$cont(i, Q) = w_i q_i;$ with r_i probability

Maria Luisa Sapino (BDM 2018)

Use of clusters
Non-binary independent features

$o_i = \langle f_{i,1}, f_{i,2}, \ldots f_{i,n} \rangle;$ $o^k = \langle f_1, f_2, \ldots, f_k \rangle;$

$R_O = \langle [r_1, w_1][r_2, w_2] \ldots [r_n, w_n] \rangle$

$q = \langle q_1, q_2, \ldots, q_k \rangle$

$$p(sim(O,Q) = s) = coef\left(x^s, \prod_{i=1}^{k} \left(r_i x^{w_i q_i} + (1 - r_i)\right)\right)$$

Maria Luisa Sapino (BDM 2018)

Use of clusters
Non-binary independent features

$o_i = \langle f_{i,1}, f_{i,2}, \ldots f_{i,n} \rangle;$ $o^k = \langle f_1, f_2, \ldots, f_k \rangle;$

$R_O = \langle [r_1, w_1][r_2, w_2] \ldots [r_n, w_n] \rangle$

A generating function!..not evaluated

$q = \langle q_1, q_2, \ldots, q_k \rangle$

$$p(sim(O,Q) = s) = coef\left(x^s, \prod_{i=1}^{k} \left(r_i x^{w_i q_i} + (1 - r_i)\right)\right)$$

Maria Luisa Sapino (BDM 2018)

What if features are not independent?

- Metric spaces assume that features are independent (orthogonal to each other)
- …what if they are not?

Maria Luisa Sapino (BDM 2018)

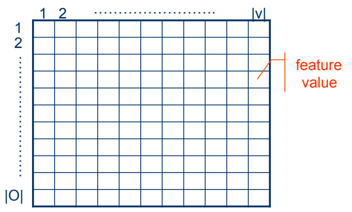
Latent Semantic Indexing

- Used for hidden (latent) concepts in a given collection
  - mostly for text collections (cosine similarity)!
- Let us have
  - |O| objects
  - Each object o is represented with a vector of size |v| (number of features)

Maria Luisa Sapino (BDM 2018)

Document-feature vector

OF =



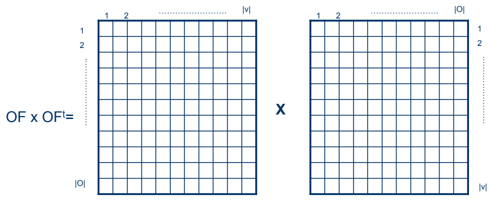
Maria Luisa Sapino (BDM 2018)

How can we use this matrix?

- This matrix is the database!!!
- Can we use it to find
  - object-object similarities?
  - feature-feature correlation?
  - independent concepts in the collection?
- Can we use it for efficient indexing?

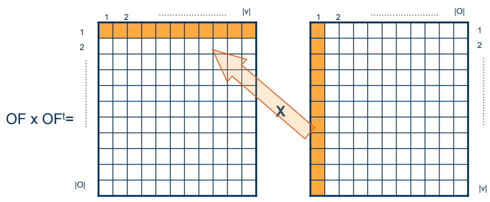
Maria Luisa Sapino (BDM 2018)

Obj-feature X feature-obj



Maria Luisa Sapino (BDM 2018)

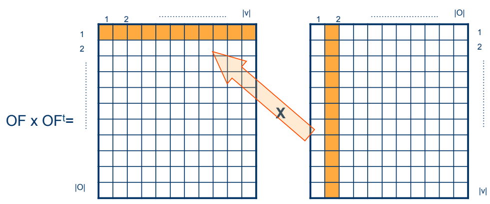
Obj-feature X feature-obj



vector multiplication (dot)

Maria Luisa Sapino (BDM 2018)

Obj-feature X feature-obj

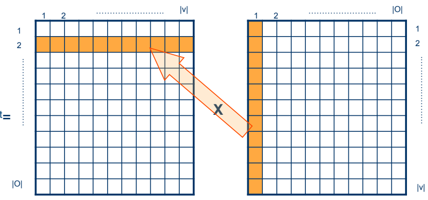


vector multiplication (dot)

Maria Luisa Sapino (BDM 2018)

Obj-feature X feature-obj

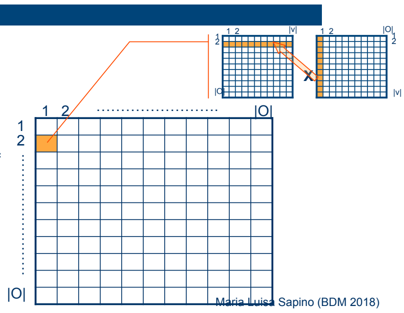
$OF \times OF^t =$



vector multiplication (dot)

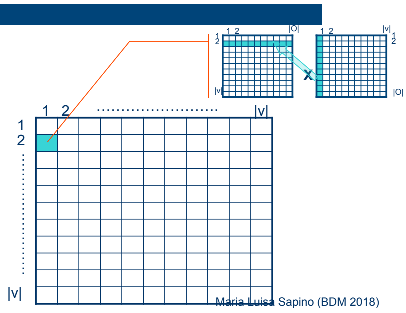
Obj-obj similarity matrix!!!!

$OF \times OF^t = OO =$

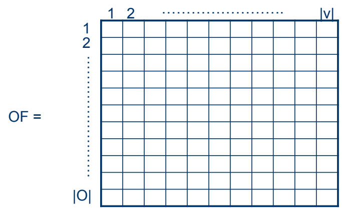


Feature-feature correl. matrix!!!!

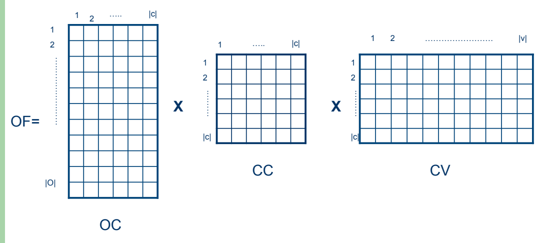
$FO \times FO^t = FF =$



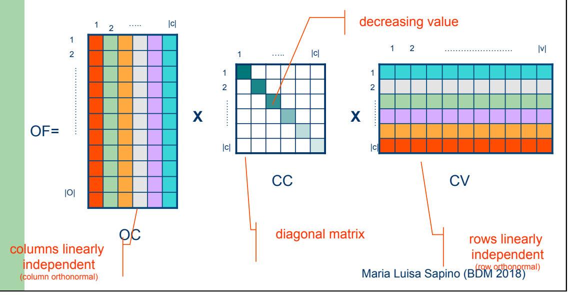
Singular valued decomposition



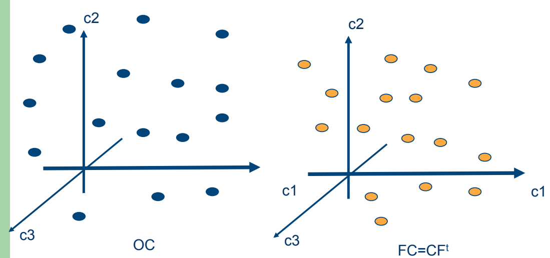
Singular valued decomposition



Singular valued decomposition

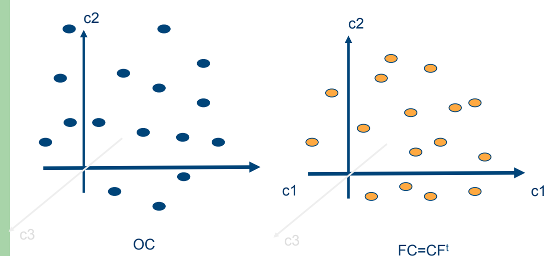


Concept space…and importance



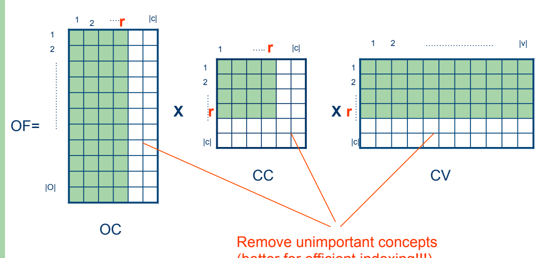
Maria Luisa Sapino (BDM 2018)

We can ignore less important concepts….



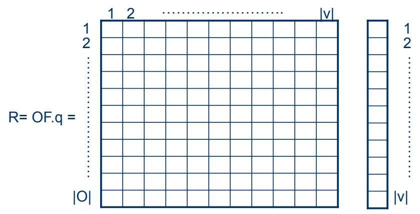
Maria Luisa Sapino (BDM 2018)

Singular valued decomposition



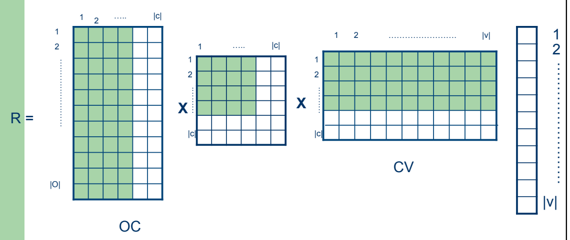
Maria Luisa Sapino (BDM 2018)

Query processing



Cost = |O|*|v|*|v|

Query processing



Cost = |O|*|r|*|r| + |r|*|r|*|r| + |r|*|v|*|v|