Remerciements: Damien DUDOGNON

Le Traitement Documentaire
Rappel des points abordés

• Quelques définitions
  – Notions de documents et de collections
  – L’accès à l’information

• Représenter les documents
  – Indexation manuelle
  – Indexation automatique
  – Stockage et représentation

• Sélectionner les documents pertinents
Le traitement documentaire

• Modèle en U

Collection

Indexation

Représentation des contenus de documents

Représentation de requête

Résesemblance

Documents retournés
Le traitement documentaire

• Représenter les documents
La traitement documentaire

• Définitions
  – Document
  – Collection

Indexation

Représentation des contenus de documents
Le traitement documentaire

• Phase d’indexation
  – Obtenir une représentation des documents
  – 3 étapes
    • Extraction
    • Normalisation
      – Troncature
      – Radicalisation
      – Lemmatisation
    • Pondération
Le traitement documentaire

- Stockage et représentation
  - Représentation logique
    - Représentation vectorielle
  - Représentation physique
    - Fichiers inverses
Indexation automatique

- **Pondération des mots normalisés**
  - $tf$: term frequency
    - Fréquence du terme dans le document
  - $df$: document frequency
    - Nombre de documents dans lesquels apparaît le terme
  - $idf$: inverse document frequency

\[
idf_t = \log \frac{N}{df}
\]

- où $N$ est le nombre de documents dans la collection

\[
Poids(t) = tf_t \cdot idf_t
\]
Indexation automatique

• Pondération des mots normalisés
  – Lois de Zipf (1949) et de Luhn (1958)
    1. Un terme apparaissant trop fréquemment dans un texte ne joue qu’un rôle syntaxique et ne doit pas être utilisé dans le langage d’indexation
    2. Un terme présent dans l’ensemble des documents n’apporte aucun pouvoir discriminant au document
    3. Un terme de fréquence moyenne est considéré comme significatif. Il représente le contenu sémantique du document et appartient au langage d’indexation
Indexation automatique

- Pondération des mots normalisés
Indexation automatique

• Pondération des mots normalisés
Indexation automatique

- Pondération des mots normalisés
Indexation automatique

- Pondération des mots normalisés
Représentation et stockage

• Deux formes de représentation
  – Logique
    • Représentation vectorielle
  – Physique
    • Indexes inversés
Représentation et stockage

- Représentation logique : représentation vectorielle
  - Chaque document est représenté par un vecteur
  - Espace vectoriel
    - t dimensions
      - t = nombre de termes dans le langage d’indexation
    - Pondération
      - Binaire (0 ou 1) : reflète la présence/absence d’un terme dans le document
      - Poids : reflète le pouvoir de caractérisation du terme pour le document
        » Par exemple avec tf . idf
Représentation et stockage

- Représentation logique : représentation vectorielle
  - Exemple
    - 4 documents (N = 4)
      - La taille du Web
      - Le Web caché
      - Technologies du Web
      - Les nouvelles technologies et technologies du Web

- Langage d’indexation après lemmatisation

<table>
<thead>
<tr>
<th>termes</th>
<th>caché</th>
<th>nouvel</th>
<th>taille</th>
<th>technologie</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Représentation et stockage

• Représentation logique : représentation vectorielle
  – Exemple
    • Représentations des documents et pondération des termes
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cachet</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>nouvel</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>taille</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>technologie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N = \text{nombre de documents} = 4$

Terme « cachet » :
$df = 1$
$Idf = \log (4/1) = \log(4) = 0,6$
<table>
<thead>
<tr>
<th>Termes</th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>( t_4 )</th>
<th>( t_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cacher</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>

N = nombre de documents = 4

Terme « cacher » (idem pour « nouvel » et « taille ») :
\( df = 1 \)
\( idf = \log \left( \frac{4}{1} \right) = \log(4) = 0,6 \)
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cachet</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>

$N = \text{nombre de documents} = 4$

Terme « cachet » (idem pour « nouvel » et « taille ») :
- $df = 1$
- $idf = \log(4/1) = \log(4) = 0,6$

Terme « technologie » :
- $df = 2$
- $idf = \log(4/2) = \log(2) = 0,3$
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cacher</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>

$N =$ nombre de documents $= 4$

Terme « cacher » (idem pour « nouvel » et « taille ») :
df $= 1$
Idf $= \log (4/1) = \log(4) = 0,6$

Terme « technologie » :
df $= 2$
Idf $= \log (4/2) = \log(2) = 0,3$

Terme « web » :
df $= 1$
Idf $= \log (4/4) = \log(1) = 0$
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cachet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nouvel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taille</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>technologie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $df$      | 1     | 1     | 1     | 2     | 4     |
| $idf$     | 0,6   | 0,6   | 0,6   | 0,3   | 0     |
| **D1**    | **Présent ?** |       |       |       |       |

**D1 : La taille du Web**
D1 : La **taille** du **Web**

Termes présents : « taille » et « web »
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cache</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>D1</td>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

D1 : La **taille** du **Web**
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$tf$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

D1 : La **taille** du **Web**

Poids(taille) = $tf \cdot Idf = 1 \times 0,6 = 0,6$

Poids(web) = $1 \times 0 = 0$
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td><strong>Présent ?</strong></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>tf</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>poids</em></td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**D1 : La *taille* du *Web***

Représentations vectorielles :

En reflétant la présence/absence des termes dans le document :

$$d_1 = 0t_1 + 0t_2 + 1t_3 + 0t_4 + 1t_5$$

$$d_1 = 1t_3 + 1t_5$$

$$d_1 = t_3 + t_5$$
<table>
<thead>
<tr>
<th>Termes</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cache</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nouvelle</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taille</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>technologie</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>web</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

D1 : La taille du Web

Représentations vectorielles :

En reflétant le pouvoir de caractérisation des termes pour le document :

\[ d₁ = 0t₁ + 0t₂ + 0,6t₃ + 0t₄ + 0t₅ \]

\[ d₁ = 0,6t₃ \]
Représentation et stockage

- **Représentation logique : représentation vectorielle**
  
  - **Exercice**
    
    - Calculer le poids des termes pour chaque document
    - Donner les représentations vectorielles de D2, D3 et D4
      - En tenant compte uniquement de la présence/absence des termes dans le document
      - En tenant compte du poids des termes
    - **On considère D1 déjà indexé**
    - **Documents**
      - D2 : Le Web caché
      - D3 : Technologies du Web
      - D4 : Les nouvelles technologies et technologies du Web
<table>
<thead>
<tr>
<th>Termes</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cache</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>D1</td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

D1 : La taille du Web

D2 : Le Web caché
D3 : Technologies du Web
D4 : Les nouvelles technologies et technologies du Web
<table>
<thead>
<tr>
<th>Termes</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cachet</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Présent ?</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Termes</td>
<td>$t_1$</td>
<td>$t_2$</td>
<td>$t_3$</td>
<td>$t_4$</td>
<td>$t_5$</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>cachet</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td><strong>D1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>D2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>D3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>Termes</td>
<td>$t_1$</td>
<td>$t_2$</td>
<td>$t_3$</td>
<td>$t_4$</td>
<td>$t_5$</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>cachet</td>
<td>nouvel</td>
<td>taille</td>
<td>technologie</td>
<td>web</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>D1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0,6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>D4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Présent ?</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>tf</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>poids</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
<td>0,6</td>
<td>0</td>
</tr>
</tbody>
</table>
**Représentation et stockage**

- **Représentation logique : représentation vectorielle**
  - **Exercice : correction**
    - **Présence/absence**
      \[
      \begin{align*}
      \vec{d}_2 &= t_1 + t_5 \\
      \vec{d}_3 &= t_4 + t_5 \\
      \vec{d}_4 &= t_2 + t_4 + t_5 \\
      \end{align*}
      \]
    - **Poids**
      \[
      \begin{align*}
      \vec{d}_2 &= 0,6t_1 \\
      \vec{d}_3 &= 0,3t_4 \\
      \vec{d}_4 &= 0,6t_2 + 0,6t_4 \\
      \end{align*}
      \]
Représentation et stockage

• Comment obtenir tous les documents qui contiennent un terme particulier ?
  – Parcours séquentiel de l’ensemble des documents ?
Représentation et stockage

• Comment obtenir tous les documents qui contiennent un terme particulier ?
  – Parcours séquentiel de l’ensemble des documents ?
    • Très couteux
    • Pas viable pour une collection importante
  – Une solution : les indexes inversés
Représentation et stockage

• Représentation physique : indexes inversés

  – Index
    • Fichier informatique
    • Permet d’accélérer l’accès aux données

  – Index inversé
    • Pour un terme donné, indique l’ensemble des documents où il apparaît
Représentation et stockage

• Représentation physique : indexes inversés
  – Index inversé simple

<table>
<thead>
<tr>
<th>Terme</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>cacher</td>
<td>D2</td>
</tr>
<tr>
<td>nouvel</td>
<td>D4</td>
</tr>
<tr>
<td>taille</td>
<td>D1</td>
</tr>
<tr>
<td>technologie</td>
<td>D3,D4</td>
</tr>
<tr>
<td>web</td>
<td>D1, D2, D3, D4, D5</td>
</tr>
</tbody>
</table>
Représentation et stockage

- Représentation physique : indexes inversés
  - Index inversé plus complet
    - Donne en plus la (les) position(s) du terme dans les documents

<table>
<thead>
<tr>
<th>Terme</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>cacher</td>
<td>D2{3}</td>
</tr>
<tr>
<td>nouvel</td>
<td>D4{2}</td>
</tr>
<tr>
<td>taille</td>
<td>D1{2}</td>
</tr>
<tr>
<td>technologie</td>
<td>D3{1}, D4{3,5}</td>
</tr>
<tr>
<td>web</td>
<td>D1{4}, D2{2}, D3{3}, D4{7}</td>
</tr>
</tbody>
</table>
Représentation et stockage

• Représentation physique : indexes inversés
  – Il existe d’autres formes d’indexes
    • Par exemple les B-Tree
      – Arbres « équilibrés »
Représentation et stockage

• Représentation physique : indexes inversés

  – Exercice

    • Soit la collection suivante :
      – D1 : « La cuisine de référence »
      – D2 : « Le Larousse de la cuisine »
      – D3 : « Mon cours de cuisine : apprendre la cuisine »
      – D4 : « Le cours cuisine des enfants »

    • Construire l’index inversé résultant de l’indexation de cette collection (en indiquant la position des termes au sein des documents)
Représentation et stockage

• Représentation physique : indexes inversés
  – Exercice : solution
    – D1 : « La cuisine de référence »
    – D2 : « Le Larousse de la cuisine »
    – D3 : « Mon cours de cuisine : apprendre la cuisine »
    – D4 : « Le cours de cuisine des enfants »

<table>
<thead>
<tr>
<th>Termes</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>apprendre</td>
<td>D3{5}</td>
</tr>
<tr>
<td>cours</td>
<td>D3{2}, D4{2}</td>
</tr>
<tr>
<td>cuisine</td>
<td>D1{2}, D2{5}, D3{4, 7}, D4{4}</td>
</tr>
<tr>
<td>enfant</td>
<td>D4{6}</td>
</tr>
<tr>
<td>larousse</td>
<td>D2{2}</td>
</tr>
<tr>
<td>référence</td>
<td>D1{4}</td>
</tr>
</tbody>
</table>
La tâche de recherche

• Modèle en U
La tâche de recherche

• La phase de recherche
  – Analyser la requête
  – Représenter la requête
  – Évaluer la ressemblance entre
    • La représentation de la requête
    • Les représentations des documents
La tâche de recherche

• Analyser la requête

Exemple « Quelles sont les technologies du Web ? »

– Traitement analogue à l’indexation

– Construire la représentation de la requête

• Même modèle que les représentations des documents

<table>
<thead>
<tr>
<th>Termes</th>
<th>ëtre</th>
<th>technologie</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correspondance langage d’indexation</td>
<td>?</td>
<td>t₄</td>
<td>t₅</td>
</tr>
<tr>
<td>df</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>tf</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Poids (tf . idf)</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>
La tâche de recherche

- **Analyser la requête**

\[ q = t_4 + t_5 \]
\[ q = 0,3t_4 \]

<table>
<thead>
<tr>
<th>Termes</th>
<th>Être</th>
<th>technologie</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correspondance langage d’indexation</td>
<td>?</td>
<td>( t_4 )</td>
<td>( t_5 )</td>
</tr>
<tr>
<td>df</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>idf</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
<tr>
<td>tf</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Poids (tf . idf)</td>
<td>0</td>
<td>0,3</td>
<td>0</td>
</tr>
</tbody>
</table>
La tâche de recherche

• Modèle en U
La tâche de recherche

• **Evaluer la ressemblance**
  
  – **Mesures de ressemblance**
    
    • Basées sur l’appartenance des termes aux textes (binaire)
      
      – Mesure de simple appariement
        
        \[
        \text{Card}(X \cap Y)
        \]
      
      – Mesure de Jaccard
        
        \[
        \frac{\text{Card}(X \cap Y)}{\text{Card}(X \cup Y)}
        \]
      
      – Fonction Cosinus
        
        \[
        \frac{\text{Card}(X \cap Y)}{\sqrt{\text{Card}(X) \times \text{Card}(Y)}}
        \]
La tâche de recherche

• **EVALUER LA RESSEMBLANCE**
  
  – **Mesures de ressemblance**
  
  • Basées sur l’appartenance des termes aux textes et sur les poids sémantiques des termes
    
    – Pour le modèle vectoriel
    
    => Dérivée de la fonction cosinus
    
    \[
    \text{Res}(D_i, Q_j) = \frac{\sum_{k=1}^{t} d_{ik} q_{jk}}{\sqrt{\sum_{k=1}^{t} (d_{ik})^2 \times \sum_{k=1}^{t} (q_{jk})^2}}
    \]
    
    – \(d_{ik}\) : poids du kième terme représentatif du doc \(D_i\)
    
    – \(q_{jk}\) : poids du kième terme représentatif de la requête
    
    – \(t\) : nombre de termes du document et de la requête
<table>
<thead>
<tr>
<th>Termes</th>
<th>(t_1)</th>
<th>(t_2)</th>
<th>(t_3)</th>
<th>(t_4)</th>
<th>(t_5)</th>
<th>(q_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>idf</td>
<td>0,6</td>
<td>0,6</td>
<td>0,6</td>
<td>0,3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Simple appariement

\[
\text{Card}(Q \cap D_1) = 1
\]

Jaccard

\[
\frac{\text{Card}(Q \cap D_1)}{\text{Card}(Q \cup D_1)} = \frac{1}{3} = 0,33
\]

Cosinus

\[
\frac{\text{Card}(Q \cap D_1)}{\sqrt{\text{Card}(Q) \times \text{Card}(D_1)}} = \frac{1}{\sqrt{2} \times \sqrt{2}} = 0,5
\]
La tâche de recherche

- **Evaluer la ressemblance**
  
  - **Exercice**
    
    - Déterminer la ressemblance entre la requête « Quelles sont les technologies du Web ? » avec les documents de la collection précédente
    1. Avec la mesure de simple appariement
    2. Avec la mesure de Jaccard
    3. Avec la fonction Cosinus (sans pondération)
La tâche de recherche

• Evaluer la ressemblance
  – Exercice : correction
    • Requête : « Quelles sont les technologies du Web »
    • Documents
      – D1 : La taille du Web
      – D2 : Le Web caché
      – D3 : Technologies du Web
      – D4 : Les nouvelles technologies et technologies du Web

<table>
<thead>
<tr>
<th>Mesure</th>
<th>Simple</th>
<th>Jaccard</th>
<th>Cosinus</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1</td>
<td>0,33</td>
<td>0,5</td>
</tr>
<tr>
<td>D2</td>
<td>1</td>
<td>0,33</td>
<td>0,5</td>
</tr>
<tr>
<td>D3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D4</td>
<td>2</td>
<td>0,66</td>
<td>0,81</td>
</tr>
</tbody>
</table>
La tâche de recherche

• Evaluer la ressemblance
  – Exercice : correction détaillée D2

**Simple appariement**

\[ \text{Card}(Q \cap D_2) = 1 \]

**Jaccard**

\[ \frac{\text{Card}(Q \cap D_2)}{\text{Card}(Q \cup D_2)} = \frac{1}{3} = 0,33 \]

**Cosinus**

\[ \frac{\text{Card}(Q \cap D_2)}{\sqrt{\text{Card}(Q) \times \text{Card}(D_2)}} = \frac{1}{\sqrt{2} \times \sqrt{2}} = 0,5 \]
La tâche de recherche

• Evaluer la ressemblance
  
  — Exercice : correction détaillée D3

Simple appariement

\[
\text{Card}(Q \cap D_3) = 2
\]

Jaccard

\[
\frac{\text{Card}(Q \cap D_3)}{\text{Card}(Q \cup D_3)} = \frac{2}{2} = 1
\]

Cosinus

\[
\frac{\text{Card}(Q \cap D_3)}{\sqrt{\text{Card}(Q) \times \text{Card}(D_3)}} = \frac{2}{\sqrt{2} \times \sqrt{2}} = 1
\]
La tâche de recherche

- Evaluer la ressemblance

  - Exercice : correction détaillée D4

Simple appariement

\[ \text{Card}(Q \cap D_4) = 2 \]

Jaccard

\[ \frac{\text{Card}(Q \cap D_4)}{\text{Card}(Q \cup D_4)} = \frac{2}{2} = 1 \]

Cosinus

\[ \frac{\text{Card}(Q \cap D_4)}{\sqrt{\text{Card}(Q) \times \text{Card}(D_4)}} = \frac{2}{\sqrt{2} \times \sqrt{3}} = 1 \]
La tâche de recherche

• **Evaluation de la tâche de recherche**
  
  — *Silence/Bruit*
  
  • Silence : documents pertinents non retrouvés
  • Bruit : documents non pertinents retrouvés

  — *Rappel/Précision*
  
  • Taux de rappel
    
    — Proportion de documents pertinents retrouvés parmi l’ensemble des documents pertinents

  • Taux de précision
    
    — Proportion de documents pertinents parmi les documents retrouvés
La tâche de recherche

- Evaluation de la tâche de recherche
  - Courbe Rappel/Précision
La tâche de recherche

• Evaluation de la tâche de recherche
  – Autres critères
    • Temps de réponse
    • Présentation des résultats
    • Facilité d’utilisation
    • ...
La tâche de recherche

• Evaluation de la tâche de recherche
  – Plateformes d’évaluation
    • Collection d’entraînement
    • Requêtes
    • Jugements de pertinence
      – Document/requête
  – Objectifs
    • Concevoir une méthode ou une application
    • Évaluer ses performances (qualitative)
    • Permettre la comparaison avec d’autres méthodes/appli
  – Exemple : TREC (Text Retrieval Conference)
La tâche de recherche

- Evolution
La tâche de recherche

• Evolution
  – Reformulation ou modification de requêtes
    • Expansion de requêtes
      – Ajout ou suppression de terme dans la requête initiale
    • Repondération des termes
    • Expansion + repondération
    • Réinjection de pertinence
      – Selon les jugements de l’utilisateur
  – Objectifs
    • Améliorer les performances
Quelques systèmes d’accès à l’information
Points abordés

• Les moteurs de recherche
• Les systèmes de recommandation
  – Filtrage collaboratif
  – Filtrage basé sur le contenu
  – Systèmes hybrides
• Les systèmes de classification
Moteurs de recherche

• Modèle en U
Moteurs de recherche

• Besoins
  – Exprimé par une requête en langage naturel
• Processus de restitution des documents
  – Interactif
    • L’usager est actif
    • Saisie de la requête
Systèmes de recommandation

• Besoins de l’usager
  – Exprimé par les contenus que l’usager
    • Consulte
    • A déjà consultés (historique) ou évalués
      – Construction d’un profil

• Processus de restitution
  – Automatique
    • Aucune action additionnelle de l’usager n’est requise
Systèmes de recommandation

- Modèles existants
  - Filtrage collaboratif
  - Filtrage basé sur le contenu
  - Hybride
Systèmes de recommandation

- Quelques systèmes de recommandation
  - Amazon
Systèmes de recommandation

- Quelques systèmes de recommandation

- LinkedIn

**LES GROUPES QUE VOUS POURRIEZ AIMER**

- Université Toulouse III - Paul Sabatier - FRANCE
  - Rejoindre - Groupe d'anciens élèves

- Hyridic Earth: Dr. Larin's concept
  - Rejoindre - Groupe professionnel

- AFUP - L'Association Française des Utilisateurs de PHP
  - Rejoindre - Groupe d'animation de réseau

**LES CONNAISSEZ-VOUS ?**

- Adrien Plat, CEO & Co-Founder at Keemix
  - Se connecter

- Candice Tissot, Chef de projet Google chez Digital Local
  - Se connecter

- Anthony Rey, Ingénieur en informatique chez Akka
  - Se connecter

**OFFRES D'EMPLOI QUI POURRAIENT VOUS INTÉRESSER**

- Software Architect (#616270) Intel Corporation - Toulouse
- INGENIEUR RESEAU H/F Storia - TOULOUSE
- 3GPP System Test Expert (#624436) Intel Corporation - Toulouse

Feedback | Voir plus »
Systèmes de recommandation

• Quelques systèmes de recommandation
  – Facebook
Systèmes de recommandation

• Quelques systèmes de recommandation
  – Google Scholar
Systèmes de recommandation

• Filtrage collaboratif (Collaborative Filtering)
  – Repose sur les jugements de pertinence émis par les autres usagers
    • Jugements explicites
      – Evaluation du document/du produit
        » Echelle de 1 à 5
        » Binaire (ex : Like FB)
      – Achat d’un produit
      – ...

Systèmes de recommandation

• Filtrage collaboratif
  – Repose sur les jugements de pertinence émis par les autres usagers
    • Jugements implicites
      – Consultation/clic
      – Temps de navigation
      – Taux de rebond
      – ...

Systèmes de recommandation

• Filtrage collaboratif
Systèmes de recommandation

- Filtrage collaboratif
Systèmes de recommandation

• Filtrage collaboratif
Systèmes de recommandation

• Filtrage collaboratif
Systèmes de recommandation

• Filtrage collaboratif
Systèmes de recommandation

- Filtrage collaboratif
Systèmes de recommandation

• Filtrage collaboratif
Systèmes de recommandation

- Filtrage collaboratif
  - Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
### Systèmes de recommandation

- **Filtrage collaboratif**
  
  – Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
Systèmes de recommandation

- **Filtrage collaboratif**
  - Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>user1</strong></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>user2</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>user3</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>user4</strong></td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
Systèmes de recommandation

• Filtrage collaboratif
  – Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
**Systèmes de recommandation**

- **Filtrage collaboratif**
  - Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
Systèmes de recommandation

• Filtrage collaboratif
  – Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
• Filtrage collaboratif

– Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
**Systèmes de recommandation**

- **Filtrage collaboratif**
  - Déterminer les corrélations

<table>
<thead>
<tr>
<th></th>
<th>obj1</th>
<th>obj2</th>
<th>obj3</th>
<th>obj4</th>
</tr>
</thead>
<tbody>
<tr>
<td>user1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>user4</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>
Systèmes de recommandation

• Filtrage basé sur le contenu (Content-based Filtering)
  – Repose sur les caractéristiques des contenus/produits
    • Contenu textuel
      – Description, titre
      ⇒ Problématique comparable à la recherche d’information
    • Autres caractéristiques
      – Couleur, prix, taille, fournisseur, ...
      – Popularité, nouveauté, ...
      – ...
Systèmes de recommandation

• Filtrage basé sur le contenu
  – Déterminer les corrélations
    • Entre les contenus disponibles
    • Et le contenu visité

=> Mesures de ressemblance
Algorithmes de classification

• Classification
  – Principe
    • Découper la collection en groupes
  – Objectifs
    • Regrouper les contenus proches
      – Groupe = Cluster
    • Identifier des ensembles thématiques, des catégories, ...
    • Selon leur proximité/ressemblance
      – Caractéristiques communes
      – Contenu textuel similaire
Algorithmes de classification

• Plusieurs algorithmes
  – K-Moyennes (KMeans)
  – Classification ascendante hiérarchique (HAC)
Algorithmes de classification

• K-Moyennes
  – Comment diviser la collection en K groupes ?

• Démarche (simplifiée)
  – Définir K, le nombre de clusters attendus
  – Créer un premier cluster contenant un document
  – Calculer le représentant du cluster
  – Pour chaque document restant dans la collection
    • Calculer la ressemblance entre le document et les représentants des clusters existants
    • Choisir le cluster le plus proche
    • Si la ressemblance avec ce cluster est au dessus d’un seuil
      – Affecter le document à ce cluster
    • Sinon
      – Créer un nouveau cluster
    • Mettre à jour le représentant du cluster
Algorithmes de classification

• Classification ascendante hiérarchique
  – Comment construire des classes hiérarchique de document semi automatiquement ?

• Démarche
  – Placer chaque document dans un cluster
  – Calculer la matrice de proximité entre clusters (2 à 2)
  – Tant qu’il reste au moins 2 clusters faire
    • Rechercher les 2 clusters les plus similaires
    • Fusionner ces deux clusters
    • Mettre à jour la matrice de proximité
Algorithmes de classification

• Classification ascendante hiérarchique
  – Mise à jour de la matrice de similarité
  – 2 approches
    • Single link
      – Similarité max entre 2 clusters
      – Production rapide de gros clusters
    • Complete link
      – Similarité min entre 2 clusters
      – Production de clusters compacts et de petite taille

– Exemple
  • Complete link
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
### Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td><strong>0,9</strong></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td><strong>0,2</strong></td>
<td><strong>0,8</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td><strong>0,6</strong></td>
<td><strong>0,4</strong></td>
<td><strong>0,1</strong></td>
<td></td>
</tr>
</tbody>
</table>
# Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

Sim(C1,C) = min( sim(C,B), sim(C,A) )

Sim(C1,C) = min (0,8, 0,2)

Sim(C1,C) = 0,2
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
</tr>
<tr>
<td>D</td>
<td>0,4</td>
</tr>
</tbody>
</table>

Sim(C1,D) = min( sim(D,B), sim(D,A) )
Sim(C1,D) = min (0,4, 0,6)
Sim(C1,D) = 0,4
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.4</td>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>

Note: C2 indicates a specific classification category.
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>

C2
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,1</td>
</tr>
</tbody>
</table>

Sim(C2,C) = min( sim(C,C1), sim(C,D) )
Sim(C2,C) = min (0,2, 0,1)
Sim(C2,C) = 0,1
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>C2</td>
<td>C3</td>
</tr>
<tr>
<td>C</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
Algorithme de classification

• Représentation
  – Dendrogramme
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>C1</td>
<td>0,9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,2</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,4</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td>0.4</td>
<td>0.1</td>
<td></td>
</tr>
</tbody>
</table>

C1
# Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,8</td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Sim}(C1,C) = \max( \text{sim}(C,B), \text{sim}(C,A) )
\]

\[
\text{Sim}(C1,C) = \max(0,8, 0,2)
\]

\[
\text{Sim}(C1,C) = 0,8
\]
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.8</td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[
\text{Sim}(C1, D) = \max(\text{Sim}(D, B), \text{Sim}(D, A))
\]

\[
\text{Sim}(C1, D) = \max(0.4, 0.6)
\]

\[
\text{Sim}(C1, D) = 0.6
\]
Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,6</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>
Algorithmes de classification

\[
\begin{array}{|c|c|c|}
\hline
 & C2 & D \\
\hline
C2 & 0,6 & \text{Sim}(C2,D) = 0,6 \\
\hline
\end{array}
\]

\textbf{Sim}(C2,D) = \text{max} ( \text{Sim}(D,C1), \text{Sim}(D,C) )

\textbf{Sim}(C2,D) = \text{max} (0,6, 0,1)

\textbf{Sim}(C2,D) = 0,6
## Algorithmes de classification

<table>
<thead>
<tr>
<th></th>
<th>C2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td></td>
<td>C3</td>
</tr>
<tr>
<td>C</td>
<td>0,6</td>
<td></td>
</tr>
</tbody>
</table>

115
## Algorithme de classification

- **Exercice – HAC**
  - Soit la matrice de départ suivante
    - Appliquer l’algorithme HAC (complete link)
    - Construire le dendrogramme

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,4</td>
<td>0,2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,1</td>
<td>0,6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0,3</td>
<td>0,5</td>
<td>0,8</td>
<td>0,7</td>
<td></td>
</tr>
</tbody>
</table>
Algorithme de classification

- Exercice – HAC
  - Correction

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0,4</td>
<td>0,2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0,1</td>
<td>0,6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0,3</td>
<td>0,5</td>
<td>0,8</td>
<td>0,7</td>
<td></td>
</tr>
</tbody>
</table>
Algorithme de classification

- **Exercice – HAC**
  - **Correction**

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>A</th>
<th>B</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0,3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0,2</td>
<td>0,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0,1</td>
<td>0,6</td>
<td></td>
</tr>
</tbody>
</table>

**C2**
**Exercice – HAC**

- **Correction**

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0,3</td>
<td>0,1</td>
<td></td>
</tr>
</tbody>
</table>

**Algorithm of classification**
Algorithme de classification

- **Exercice – HAC**
  - Correction

<table>
<thead>
<tr>
<th></th>
<th>C3</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

C4