Génération d'adaptateurs web intelligents à l'aide de techniques de fouille de texte

Huaizhong KOU

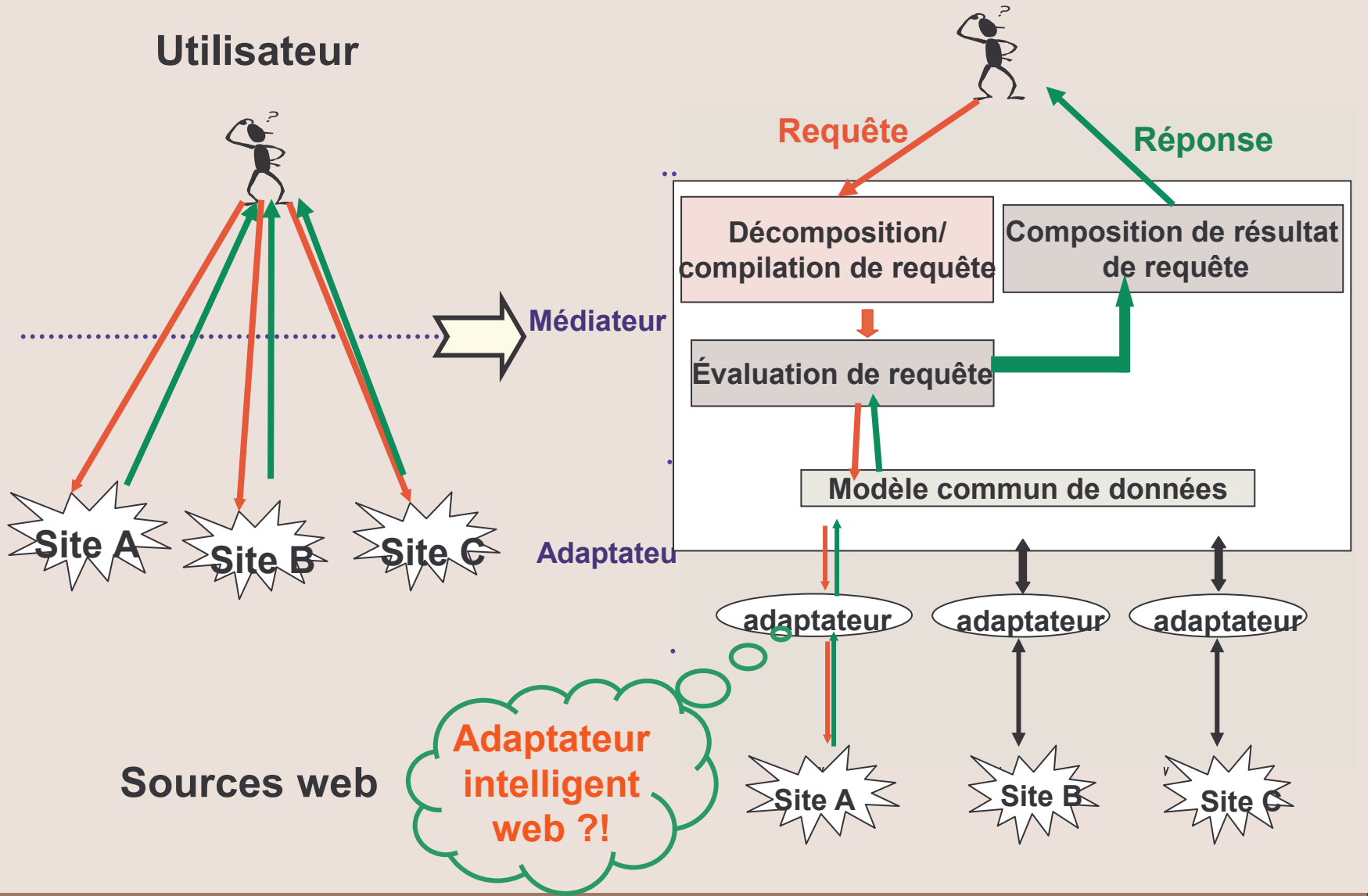
Directeur: Prof. Georges GARDARIN

IFDOCS, Laboratoire PRiSM, Université de Versailles St-Quentin

Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
5. Conclusion

1.Contexte : architecture de médiation web



Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
5. Conclusion

2. Objectif

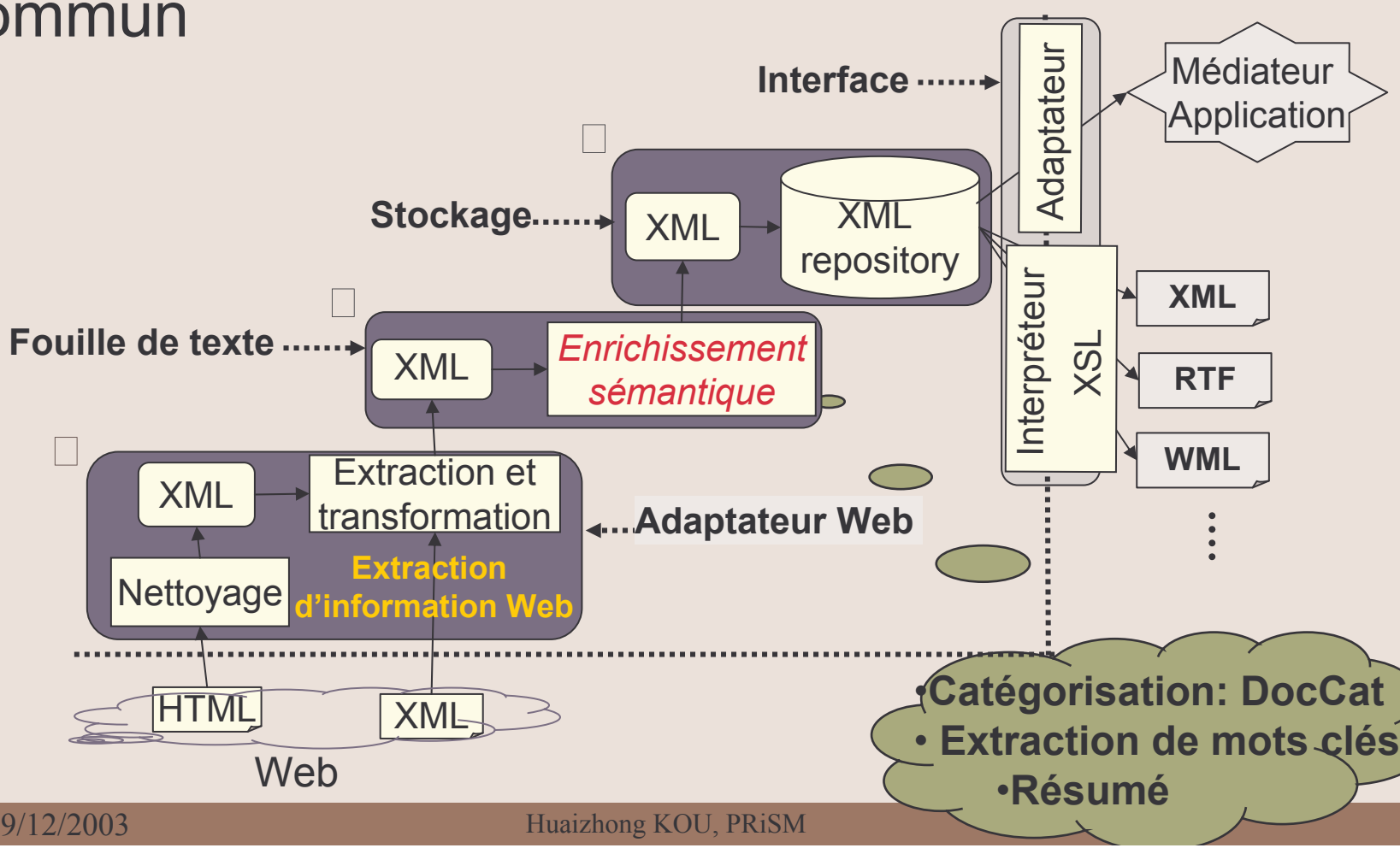
- Définir un cadre de système pour intégrer sémantiquement des informations web
  - Adaptateur web : extraire des informations web
  - Fouille de texte : découvrir des sémantiques contenues dans les informations textuelles
    - Catégorisation documentaire, extraction des mots clés, résumé automatique de documents
  - Enrichissement sémantique : baliser des informations avec leur sémantiques

Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
5. Conclusion

3.SEWISE : adaptateur web sémantique

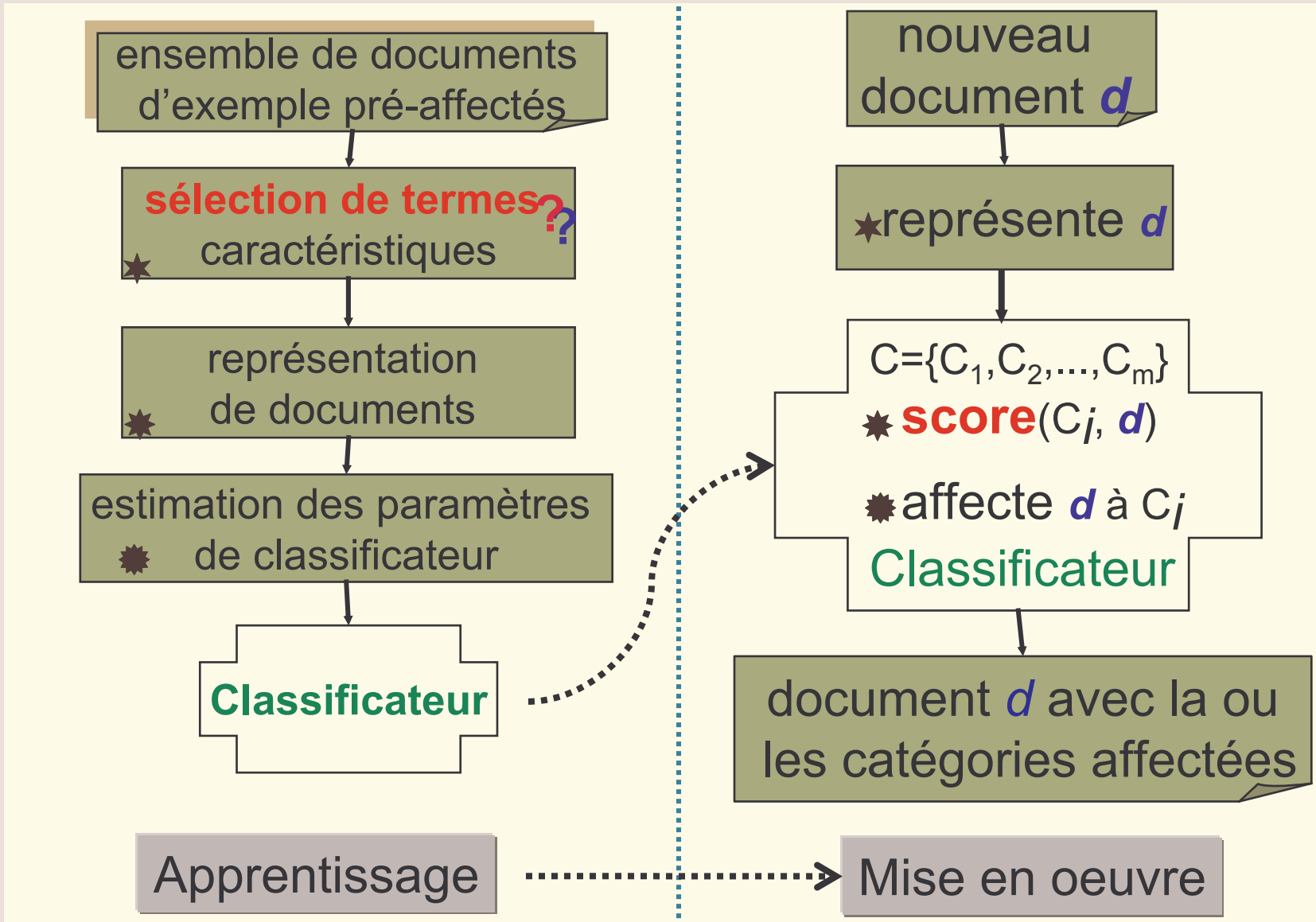
- Objectif: extraire des informations web, les enrichir sémantiquement et les transformer en un modèle commun



Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - 4.2. Modèles de présentation de documents
   - 4.3. Algorithmes de catégorisation
   - 4.4. Approches de sélection de termes
   - 4.5. Modèle de similarité et association de termes
   - 4.6. Calcul de score de catégories pour k-NN
5. Conclusion

4.1 Processus de catégorisation



Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - Modèles de présentation de documents
   - 4.3. Algorithmes de catégorisation
   - 4.4. Approches de sélection de termes
   - 4.5. Modèle de similarité et association de termes
   - 4.6. Calcul de score de catégories pour k-NN
5. Conclusion

4.2. Modèles de représentation

- Modèle vectoriel
  - ensemble de termes e
  - coordonnées, poids d
  - modèles de similarités

la direction d'Airbus a confirmé les information contenues dans le Financial Times mentionnant l'élaboration d'un plan de réduction de coûts...

(airbus 0.13, coût 0.053, direction 0.02, élaboration 0.03, information 0.01, plan 0.1, réduction 0.06, ...)

- Distance Euclidienne

$$simil(d_1, d_2) = \|d_1 - d_2\|^2$$

- Fonction cosinus

$$simil(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\|^2 \times \|d_2\|^2}$$

- Produit scalaire

$$simil(d_1, d_2) = d_1 \bullet d_2$$

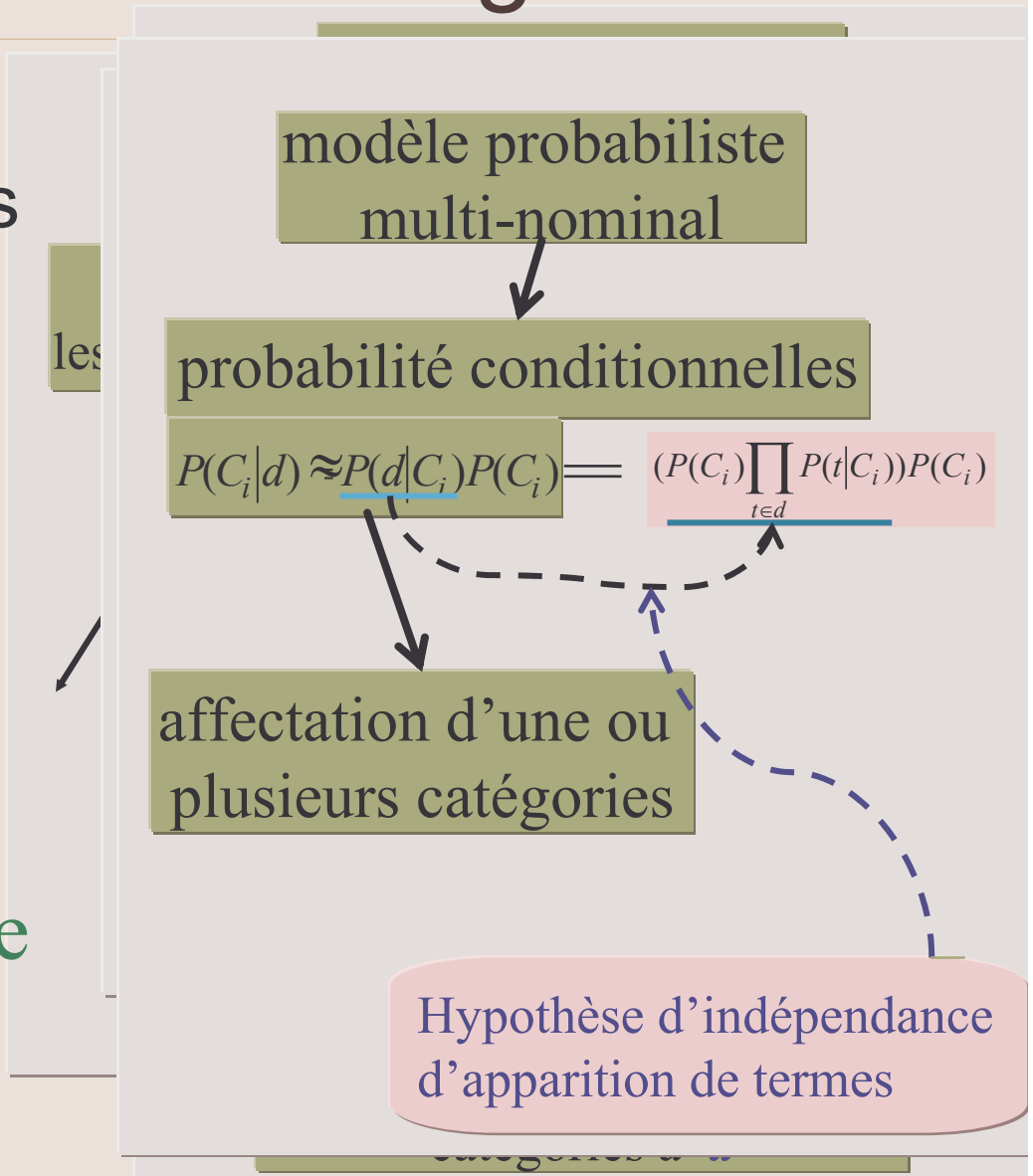
- Hypothèse d'orthogonalité sur des vecteurs des dimensions ⇨ néglige des associations des termes

Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - 4.2 Modèles de présentation de documents
   - Algorithmes de catégorisation
   - 4.4. Approches de sélection de termes
   - 4.5. Modèle de similarité et association de termes
   - 4.6. Calcul de score de catégories pour k-NN
5. Conclusion

4.3. Algorithmes de catégorisation

- Algorithme des k plus proches voisins (k-NN)
- Algorithme de Centroide
- Algorithme de naïve Bayes (NB)
- Algorithme hybride proposé (CKNN): Cetroide + k-NN

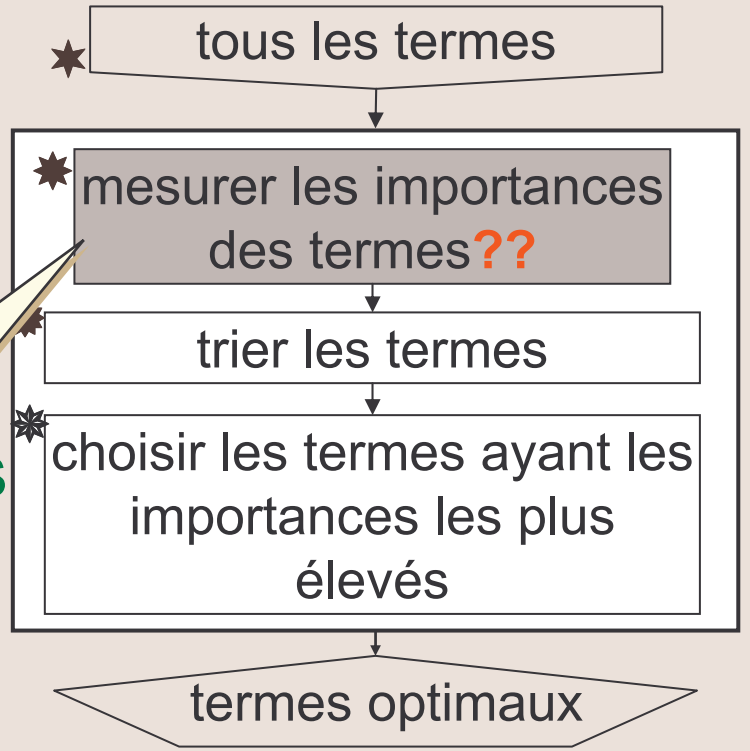


Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - 4.2 Modèles de présentation de documents
   - 4.3. Algorithmes de catégorisation
   - Approches de sélection de termes
   - 4.5. Modèle de similarité et association de termes
   - 4.6. Calcul de score de catégories pour k-NN
5. Conclusion

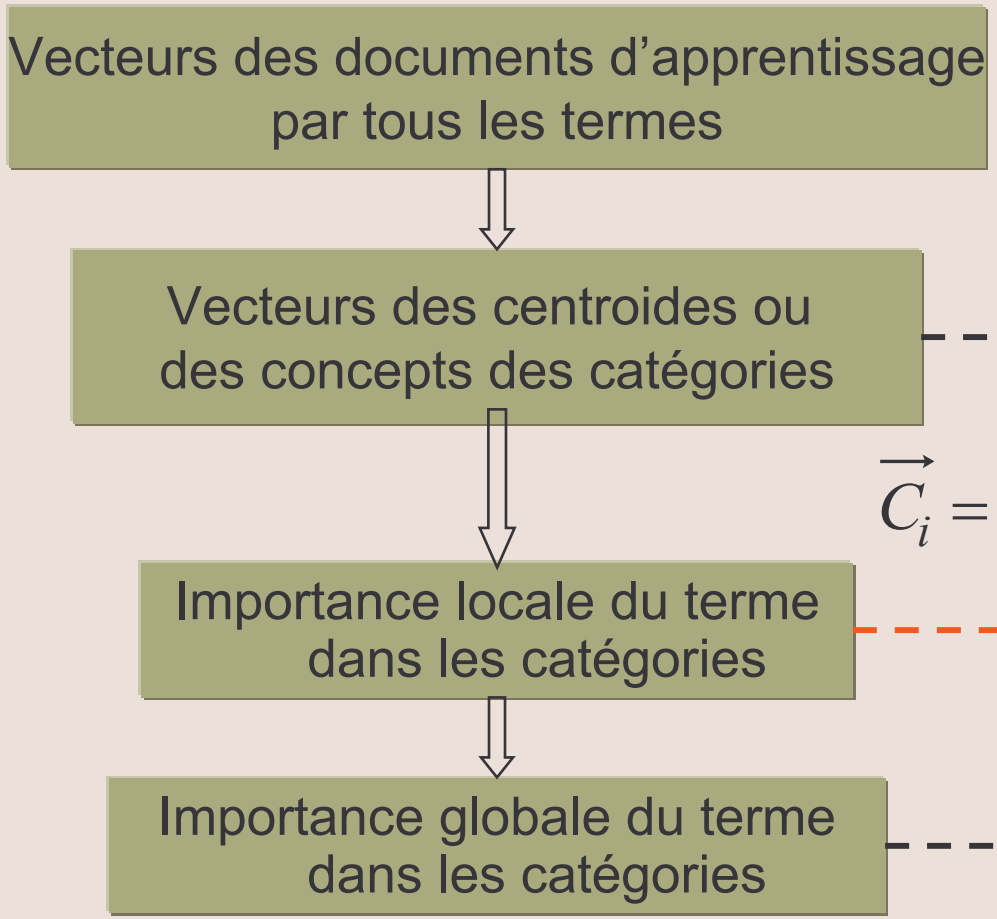
4.4. Approches de sélection de termes

- But: sélectionner un sous-ensemble optimal des termes pour représenter des documents.
  - Trop de termes pour l'apprentissage
  - L'emploie de tous les termes n'améliore pas les performances
- Approches de filtrage
- Contribution:
  - CBA et IBA pour mesurer les importances des termes



- **Gain informationnel (IG)**
- **Information mutuelle (MI)**
- **CHI-test (χ^2): performant**

Algorithme CBA* de sélection de termes



$$\vec{C}_i = (w_{ci1}, w_{ci2}, \ldots, w_{ciL})$$

$$CW(t_l) = \sum_{i=1}^{m} P(C_i)\, w_{cil}$$

*Concept-Based Approach

Algorithme IBA* de sélection de termes

Probabilité conditionnelle $P(t|C_i)$: importance locale du terme t pour C_i

Importance globale de terme dans toutes les catégories

$$CW(t) = \sum_{i=1}^{m} P(C_i)P(t|C_i)$$

Plus $P(t|C_i)$ est grand, plus le terme t est important pour la catégorie C_i

Probabilité de la catégorie C_i

**Independance-Based Approach*

Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - 4.2 Modèles de présentation de documents
   - 4.3. Algorithmes de catégorisation
   - 4.4. Approches de sélection de termes
   - Modèle de similarité et association de termes
   - 4.6. Calcul de score de catégories pour k-NN
5. Conclusion

4.5. Modèle de similarité et association de termes

- Hypothèse d'orthogonalité
  - Orthogonalités des vecteurs des termes
  - Néglige les associations des termes
- Similarité de documents: produit scalaire

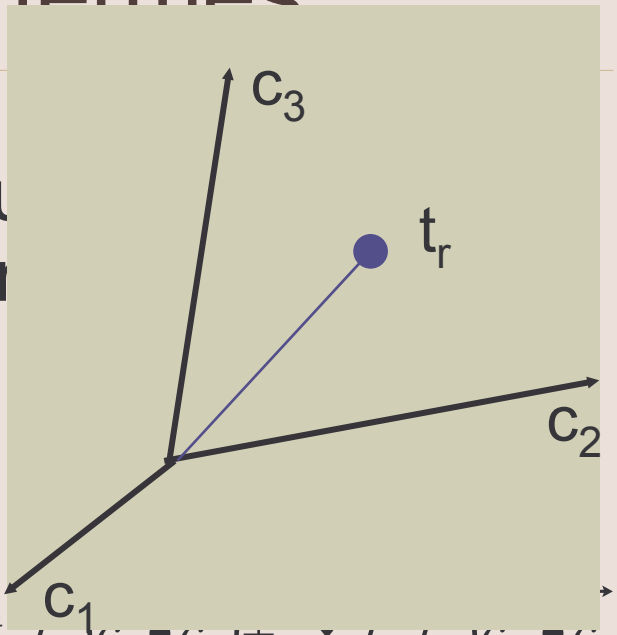
$$simil(d_1,d_2) = d_1 \bullet d_2 = (\sum_{r=1}^{T} w_{1r}\vec{t_r})(\sum_{s=1}^{T} w_{2s}\vec{t_s}) = \sum_{r=1}^{T} w_{1r}w_{2r}(\vec{t_r} \bullet \vec{t_r}) + \sum_{\substack{r,s=1 \\ r\neq s}}^{T} w_{1r}w_{2s}(\vec{t_r} \bullet \vec{t_s})$$

?

- Contribution :
  - Proposer un modèle mathématique pour estimer les associations des termes et les intégrer dans les modèles de similarités

Modèle d'associations de termes

- Principe: utiliser
  - les pertinences des termes pou
  - les appartenances de documer
- Espace de terme-catégorie
  - Définit associations de termes



$$t_r = \sum_{i=1}^{m} t_{ri}\,\vec{c}_{i_m}$$

$$tc\_sim(t_r, t_s) = \vec{t_r} \bullet \vec{t_s} = \sum_{i=1}^{m}\sum_{j=1}^{m} t_{ri}t_{sj}(\vec{c_i}\bullet\vec{c_j}) = \sum_{i=1}^{m} t_{ri}t_{si} + \sum_{\substack{k,l=1 \\ k\neq l}} t_{rk}t_{sl}(\vec{c_k}\bullet\vec{c_l})$$

CHI-test

- Espace booléen de catégorie-document

$$c_r = (c_{r1}, c_{r2}, \ldots, c_{rn}) \quad c_{ri} = 1 \text{ si } d_i \in c_r; 0 \text{ sinon}$$

  - Coefficient de Jaccard

$$cd\_sim(c_r, c_s) = \frac{|c_r \cap c_s|}{|c_r| + |c_s| - |c_r \cap c_s|}$$

Intégration dans les modèles de similarités

Associations de termes

$$ass(t_r, t_s) = \sum_{l=1}^{m} t_{rl} t_{sl} + \sum_{\substack{k,l=1 \\ k \neq l}}^{m} t_{rk} t_{sl} cd\_sim(c_k, c_l)$$

Forme normalisée

$$ass(t_r, t_s) = \frac{ass(t_r, t_s)}{\max_{k,l} \{ass(t_k, t_l)\}}$$

Seuil d'associations de termes

$$\varepsilon\_ass(t_r, t_s) = \begin{cases} 0 & si\ ass(t_r, t_s) < \varepsilon \\ ass(t_r, t_s)\ si\ ass(t_r, t_s) \geq \varepsilon \end{cases} \qquad \varepsilon \in (0,1)$$

Similarité avec associations de termes

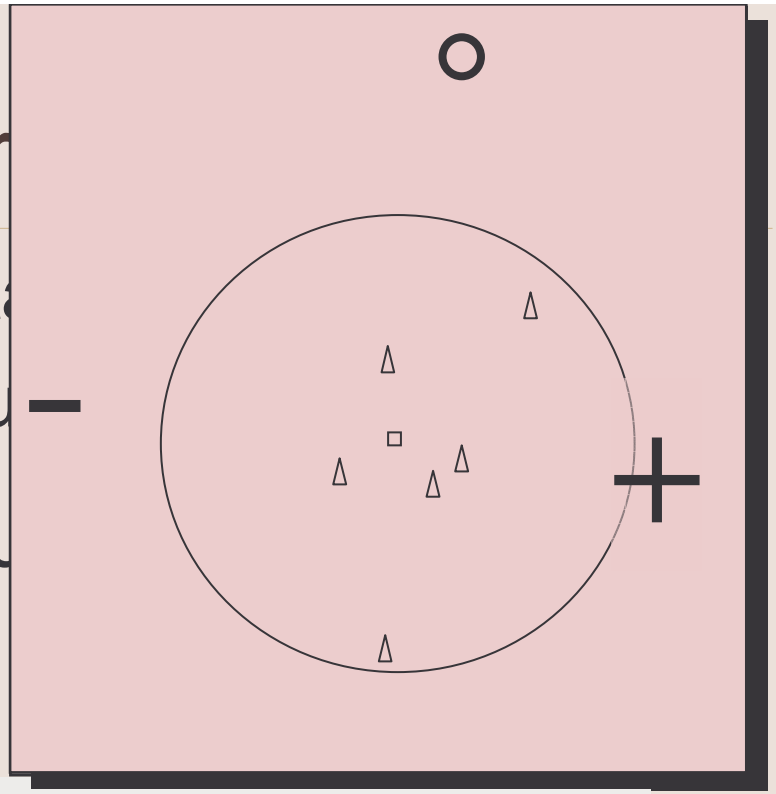
$$\varepsilon\_sim(d_i, d_j) = \sum_{r=1}^{T} w_{ir} w_{jr} + \sum_{\substack{r,s=1 \\ r \neq s}}^{T} w_{ir} w_{js} \varepsilon\_ass(t_r, t_s)$$

Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
   - 4.1. Processus de catégorisation
   - 4.2 Modèles de présentation de documents
   - 4.3. Algorithmes de catégorisation
   - 4.4. Approches de sélection de termes
   - 4.5. Modèle de similarité et association de termes
   - Calcul de score de catégories pour k-NN
5. Conclusion

5.6. Score de catégor

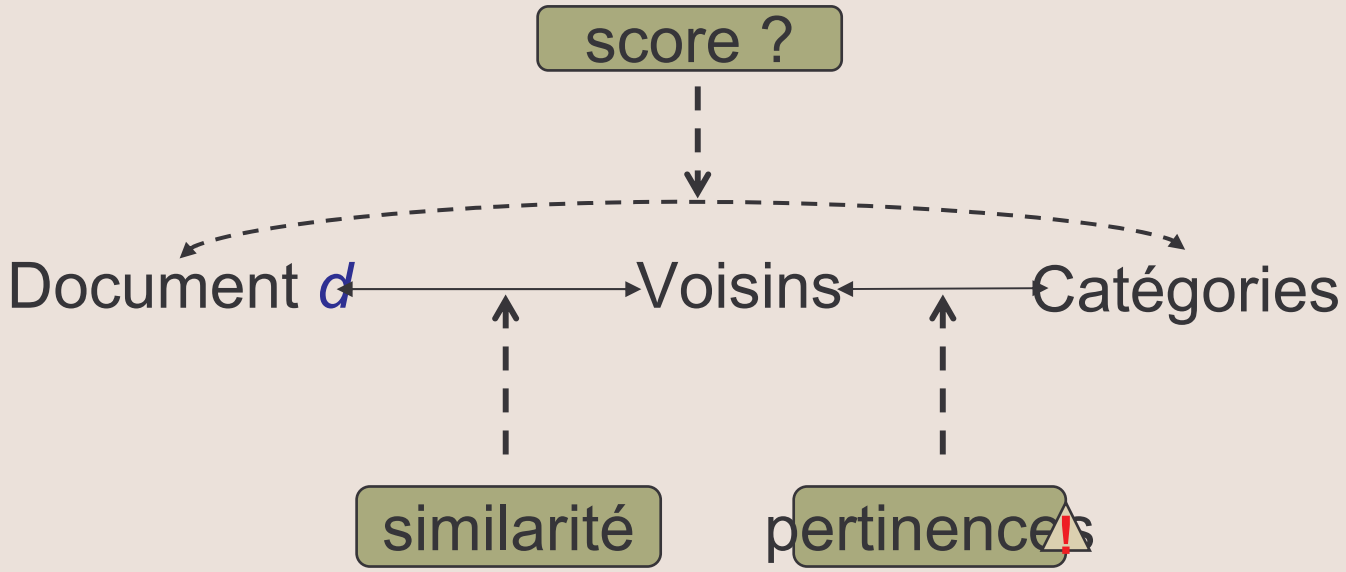
- Score de catégories permett
  pertinences du document pou
- Algorithmes plus répandu
  - Majorité votant (MVA)
  - Somme simple de similarité(SSA)
    - Score.SSA=somme simple de similarité entre *d* et des documents dans les k voisins appartenant à la même catégorie
    - pb: Considérant équitablement des apports de différents documents au score de catégories



- Contribution :
  - Algorithmes pondérés CBW et IBW

Algorithmes pondérés de score de catégories pour k-NN

- Profiter des relations existantes pour calculer des scores

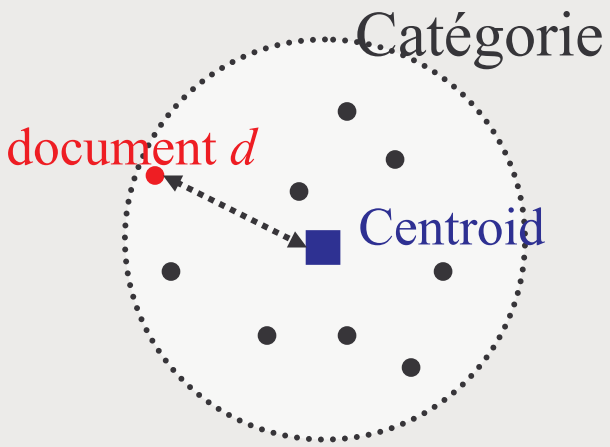


Score = ∑(pertinence•similarité)

Algorithmes pondérés CBW et IBW

- Algorithme pondéré basé sur des vecteurs de concept (CBW: Concept-based weighting)

$$per(d, \text{catégorie}) = simil(d, \text{centroid})$$

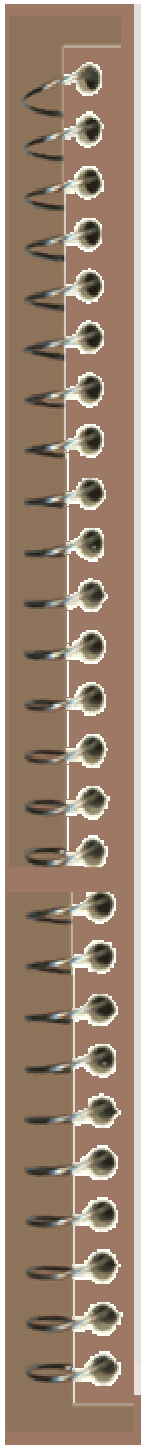


Plan

1. Contexte : architecture de médiation web
2. Objectif
3. SEWISE: adaptateur web sémantique
4. Catégorisation documentaire
5. Conclusion

5.Conclusion

- Contribution
  - Cadre du système SEWISE **(NLDB'03)**
  - Deux approches de sélection de termes **(NLDB'03)**
  - Associations de termes et des modèles de similarité **(DEXA'02)**
  - Approches du calcul de score de catégories employées dans k-NN. **(SIGIR'02)**
  - Prototype DocCat **(ISI'03, AIA'02)**
    - Application au projet CONTEXTE Bourse



Merci