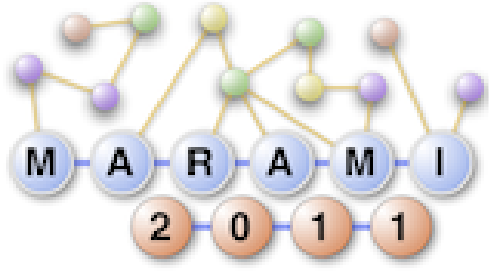
Calcul de co-similarité pour la classification de documents à grande échelle

Clément Grimal et Gilles Bisson

Université de Grenoble, France



L I G

UNIVERSITÉ DE GRENOBLE

Plan

Plan

Le challenge Pascal LSHTC

Classification de documents textuels, basée sur leur contenu

Hiérarchie de catégories

Exemple de hierarchies sur le web : Wikipedia, DMOZ...
Comment classifier automatiquement un nouveau document dans cette hiérarchie ?

Les classes sont les feuilles terminales de la hiérarchie.
On appelle catégories tous les nœuds de la hiérarchie, et classes les nœuds terminaux.

Restriction au cas mono-classe, hiérarchie stricte.

Le challenge Pascal LSHTC

Classification de documents textuels, basée sur leur contenu

Hiérarchie de catégories

Exemple de hierarchies sur le web : Wikipedia, DMOZ...
Comment classifier automatiquement un nouveau document dans cette hiérarchie ?

Les classes sont les feuilles terminales de la hiérarchie.
On appelle catégories tous les nœuds de la hiérarchie, et classes les nœuds terminaux.

Restriction au cas mono-classe, hiérarchie stricte.

Problèmes

- Taille des données
- Beaucoup de documents mais surtout **beaucoup de classes** !
- Nombreux auteurs $\Rightarrow$ grande variabilité terminologique

Propositions

Variabilité terminologique

Utilisation de l'algorithme χ-SIM_p^h

Trouver des liens entre des documents ne partageant **directement** de termes

Propositions

Variabilité terminologique

Utilisation de l'algorithme χ-SIM_p^h
Trouver des liens entre des documents ne partageant **directement** de termes

Nombre de classes et taille des données

Découpage de la hiérarchie,
i.e. clustering préalable sur les catégories
i.e. **Diviser pour mieux régner**

Propositions

Variabilité terminologique

Utilisation de l'algorithme χ-SIM_p^h
Trouver des liens entre des documents ne partageant **directement** de termes

Nombre de classes et taille des données

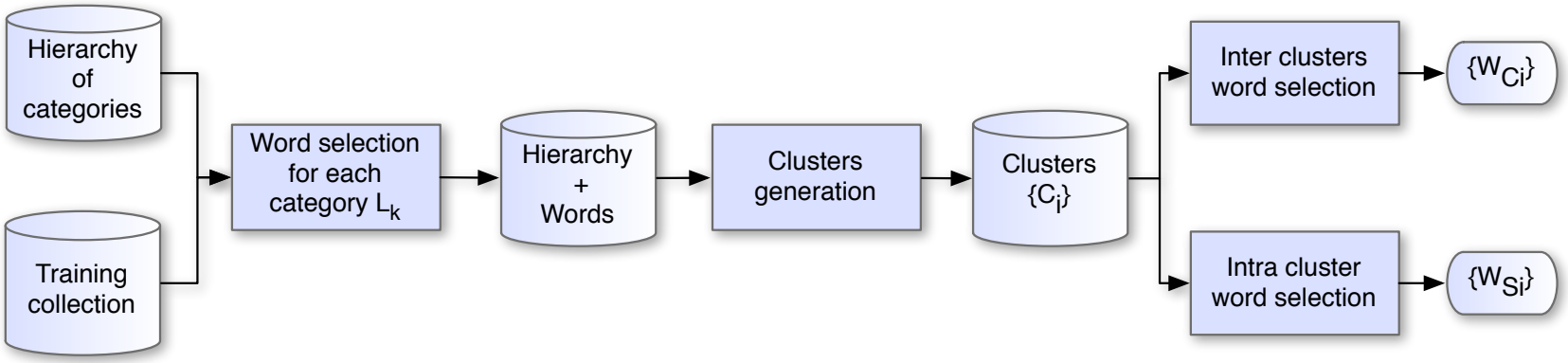
Découpage de la hiérarchie,
i.e. clustering préalable sur les catégories
i.e. **Diviser pour mieux régner**

Classification en 2 étapes

1. Affectation du document test à un cluster
2. k plus proches voisins **au sein** du cluster

Plan

Le processus de pré-traitement



Trois étapes de sélection de mots :

1. pour le clustering des catégories
2. pour discriminer entre les clusters
3. pour discriminer entre les classes au sein d'un cluster

Création des clusters

Pourquoi ?

- Beaucoup de classes
- Certaines classes contiennent très peu de documents
- Calcul des similarités entre tous les mots de la base impossible

Création des clusters

Pourquoi ?

- Beaucoup de classes
- Certaines classes contiennent très peu de documents
- Calcul des similarités entre tous les mots de la base impossible

Comment ?

Approche agrégative :

- sélection des mots les plus représentatifs des catégories
- fusion des catégories “sœurs” les plus similaires

Création des clusters – Choix des mots

On veut un vecteur de mots pour chaque catégorie...

Score pour les mots

Différents scores possibles : Information Mutuelle, BNS (Bi-Normal Separation), etc.
On utilise la Double Probabilité Conditionnelle :

$$\mathrm{DPC}(\mathrm{mot}, \mathrm{categorie}) = P(\mathrm{mot}|\mathrm{categorie}) \times P(\mathrm{categorie}|\mathrm{mot})$$

Création des clusters – Choix des mots

On veut un vecteur de mots pour chaque catégorie...

Score pour les mots

Différents scores possibles : Information Mutuelle, BNS (Bi-Normal Separation), etc.
On utilise la Double Probabilité Conditionnelle :

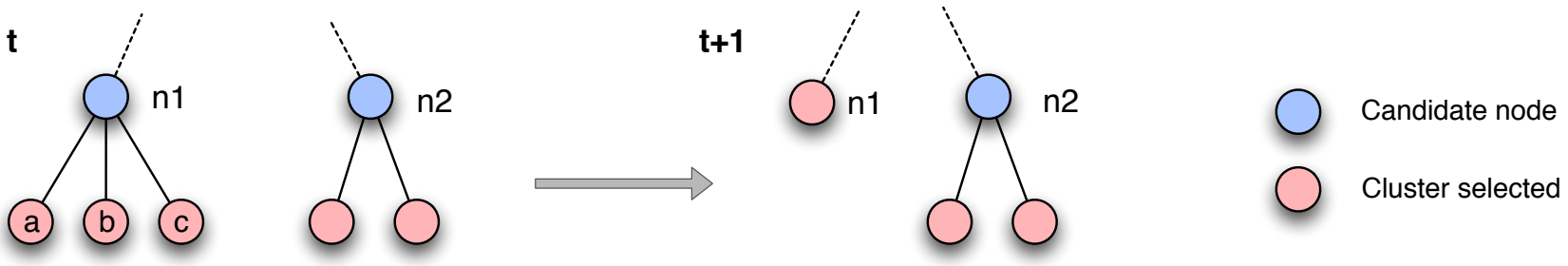
$$\mathrm{DPC}(\mathrm{mot}, \mathrm{categorie}) = P(\mathrm{mot}|\mathrm{categorie}) \times P(\mathrm{categorie}|\mathrm{mot})$$

Différentes stratégies de choix :

- meilleurs mots individuellement (BIF)
- prise en compte des dépendances entre mots (séquentiel)

Notre choix :
BIF + seuil, i.e. seuil sur les scores pour choisir un nombre variable de mots en fonction de la difficulté à décrire une catégorie

Création des clusters – Fusion des catégories



score(n_1) $>$ score(n_2) $\Rightarrow$ fusion des catégories filles de n_1, et n_1 devient un cluster

score(n) $\equiv$ moyenne pondérée de la similarité entre n et ses catégories filles

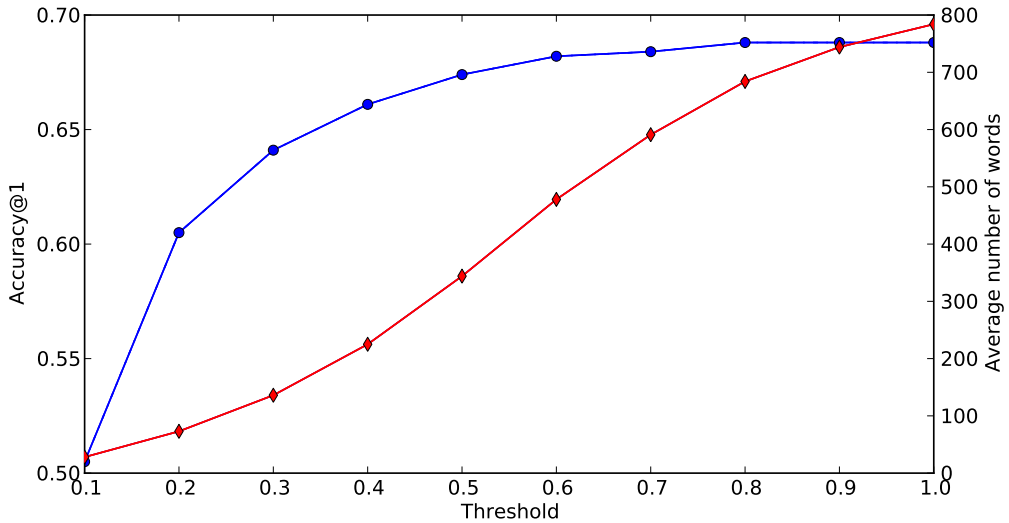
$$\frac{\sum_{n' \in \text{children}(n)} \text{size}(n') \times \text{Cosine}(n, n')}{\sum_{n' \in \text{children}(n)} \text{size}(n')}$$

Sélection des mots inter-clusters

Pour chaque cluster i, on cherche W_{C_i}, les mots inter-clusters, qui nous permettront de trouver les clusters les plus proches d'un document test.

On ne cherche plus à discriminer qu'entres les clusters (plus entre toutes les catégories).

Sélection des mots inter-clusters (2)



Pourcentage de documents (apprentissage) affectés au bon cluster : Accuracy@1

Variation du seuil sur le score des mots, on observe la variation du nombre de mots sélectionnés et de la précision Accuracy@1 quand on augmente ce seuil :

- le nombre de mots augmente (linéairement)
- la précision Acc@1 augmente (asymptote)

On choisit donc de fixer le seuil à 50%

Sélection des mots intra-clusters

Pour chaque cluster i, on cherche W_{S_i}, les mots intra-clusters, qui nous permettront de trouver les documents d'apprentissage les plus proches du document test au sein d'un cluster.

Ces mots doivent être discriminants entre les classes représentées dans le cluster.

Plan

Apprentissage – Objectif

L'objectif est de calculer les indices de similarités entre toutes les paires de mots intra-clusters, afin d'améliorer la classification.
$\Rightarrow$ Pour chaque cluster i, on veut $\mathbf{SC}_i$, la matrice de similarité entre les mots W_{S_i}.

On pourrait aussi par extension, calculer la similarité entre les mots inter-clusters, afin d'améliorer la prédiction du cluster.

On va utiliser χ-$\textsc{Sim}_p^h$, algorithme de calcul de co-similarité.

Apprentissage – Présentation de χ-SIM_p^h

Modèle

$\mathbf{M}$: matrice documents/mots de r lignes et c colonnes

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: vecteur ligne décrivant le document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: vecteur colonne décrivant le mot j

Apprentissage – Présentation de χ-SIM$_p^h$

Modèle

$\mathbf{M}$: matrice documents/mots de r lignes et c colonnes

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: vecteur ligne décrivant le document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: vecteur colonne décrivant le mot j

Output de l'algorithme :

- $\mathbf{SR}$: matrice de similarité (carrée et symétrique) des documents de taille $r \times r$, avec $sr_{ij} \in [0, 1]$
- **$\mathbf{SC}$: matrice de similarité (carrée et symétrique) des mots de taille $c \times c$, avec $sc_{ij} \in [0, 1]$**

Apprentissage – Présentation de χ-SIM$_p^h$

Modèle

$\mathbf{M}$: matrice documents/mots de r lignes et c colonnes

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: vecteur ligne décrivant le document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: vecteur colonne décrivant le mot j

Output de l'algorithme :

- $\mathbf{SR}$: matrice de similarité (carrée et symétrique) des documents de taille $r \times r$, avec $sr_{ij} \in [0, 1]$
- **$\mathbf{SC}$: matrice de similarité (carrée et symétrique) des mots de taille $c \times c$, avec $sc_{ij} \in [0, 1]$**

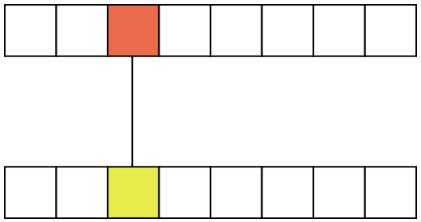
Idée de base

- Deux documents sont similaires s'ils contiennent des mots similaires.
- Deux mots sont similaires s'ils apparaissent dans des documents similaires.

Apprentissage – Mesures classiques

- Approche classique : similarité = f(caractéristiques communes)

$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \sum_{l=1}^{r} \mathrm{F_s}(m_{li}, m_{lj})$$

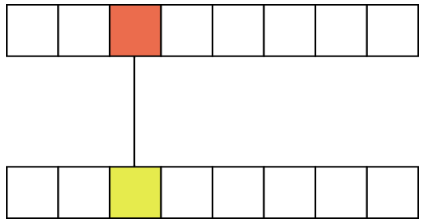


avec $\mathrm{F_s}$ une fonction de similarité (différence, produit, etc.)

Apprentissage – Mesures classiques

- Approche classique : similarité = f(caractéristiques communes)

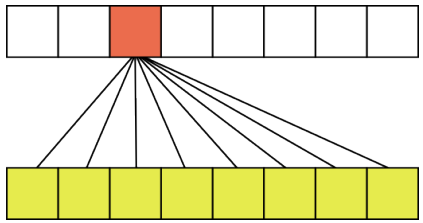
$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \sum_{l=1}^{r} \mathrm{F_s}(m_{li}, m_{lj})$$



avec $\mathrm{F_s}$ une fonction de similarité (différence, produit, etc.)

- Et si l'on compare maintenant toutes les paires de caractéristiques :

$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \sum_{l=1}^{r} \sum_{n=1}^{r} \mathrm{F_s}\left(m_{li}, m_{nj}\right) \times sr_{ln}$$



avec une pondération par la similarité entre ces caractéristiques sr_{ln}

Apprentissage – Pseudo-norme h

- Si $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \mathbf{m}_{:i}^{\mathrm{T}} \times \mathbf{SR} \times \mathbf{m}_{:j}$$

Apprentissage – Pseudo-norme h

- Si $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \mathbf{m}_{:i}^{\mathrm{T}} \times \mathbf{SR} \times \mathbf{m}_{:j}$$

- On introduit une pseudo-norme h (voir [Aggarwal et al.(2001)]) :

$$\mathrm{Sim}^h(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \left(\left(\mathbf{m}_{:i}^{\mathrm{T}} \right)^h \times \mathbf{SR} \times (\mathbf{m}_{:j})^h \right)^{1/h} = \langle \mathbf{m}_{:i}, \mathbf{m}_{:j} \rangle_{\mathbf{SR}}^h$$

$\rightarrow$ On a $\|\mathbf{m}_{:i}\|_{\mathbf{SR}}^h = \sqrt{\langle \mathbf{m}_{:i}, \mathbf{m}_{:i} \rangle_{\mathbf{SR}}^h}$

Apprentissage – Pseudo-norme h

- Si $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \mathbf{m}_{:i}^{\mathrm{T}} \times \mathbf{SR} \times \mathbf{m}_{:j}$$

- On introduit une pseudo-norme h (voir [Aggarwal et al.(2001)]) :

$$\mathrm{Sim}^h(\mathbf{m}_{:i}, \mathbf{m}_{:j}) = \left(\left(\mathbf{m}_{:i}^{\mathrm{T}} \right)^h \times \mathbf{SR} \times (\mathbf{m}_{:j})^h \right)^{1/h} = \langle \mathbf{m}_{:i}, \mathbf{m}_{:j} \rangle_{\mathbf{SR}}^h$$

$\rightarrow$ On a $\|\mathbf{m}_{:i}\|_{\mathbf{SR}}^h = \sqrt{\langle \mathbf{m}_{:i}, \mathbf{m}_{:i} \rangle_{\mathbf{SR}}^h}$

- On doit finalement normaliser cette similarité :

$$sc_{ij} = \frac{\left(\left(\mathbf{m}_{:i}^{\mathrm{T}} \right)^h \times \mathbf{SR} \times (\mathbf{m}_{:j})^h \right)^{1/h}}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})} \in [0, 1]$$

Apprentissage – Forme générique

- Finalement :

$$sc_{ij} = \frac{\left(\left(\mathbf{m}_{:i}^{\mathrm{T}}\right)^h \times \mathbf{SR} \times \left(\mathbf{m}_{:j}\right)^h\right)^{1/h}}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})} = \frac{\langle \mathbf{m}_{:i}, \mathbf{m}_{:j} \rangle_{\mathbf{SR}}^h}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})}$$

Apprentissage – Forme générique

- Finalement :

$$sc_{ij} = \frac{\left(\left(\mathbf{m}_{:i}^{\mathrm{T}}\right)^h \times \mathbf{SR} \times \left(\mathbf{m}_{:j}\right)^h\right)^{1/h}}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})} = \frac{\langle \mathbf{m}_{:i}, \mathbf{m}_{:j} \rangle_{\mathbf{SR}}^h}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})}$$

- Pour des valeurs particulières de h, $\mathbf{SR}$ et $\mathcal{N}$, on a :
  - Jaccard: $\mathbf{SR} = \mathbf{I}$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_1 + \|\mathbf{m}_{:j}\|_1 - \mathbf{m}_{:j}\mathbf{m}_{:j}^{\mathrm{T}}$
  - Dice: $\mathbf{SR} = 2\mathbf{I}$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_1 + \|\mathbf{m}_{:j}\|_1$
  - Ancien χ-Sim: $h = 1$, $\mathcal{N} = |\mathbf{m}_{:i}| \times |\mathbf{m}_{:j}|$
  - Cosinus généralisé: $\mathbf{SR} > 0$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_{\mathbf{SR}} \times \|\mathbf{m}_{:j}\|_{\mathbf{SR}}$

Apprentissage – Forme générique

- Finalement :

$$sc_{ij} = \frac{\left(\left(\mathbf{m}_{:i}^{\mathrm{T}}\right)^h \times \mathbf{SR} \times \left(\mathbf{m}_{:j}\right)^h\right)^{1/h}}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})} = \frac{\langle \mathbf{m}_{:i}, \mathbf{m}_{:j} \rangle_{\mathbf{SR}}^h}{\mathcal{N}(\mathbf{m}_{:i}, \mathbf{m}_{:j})}$$

- Pour des valeurs particulières de h, $\mathbf{SR}$ et $\mathcal{N}$, on a :
  - Jaccard: $\mathbf{SR} = \mathbf{I}$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_1 + \|\mathbf{m}_{:j}\|_1 - \mathbf{m}_{:j}\mathbf{m}_{:j}^{\mathrm{T}}$
  - Dice: $\mathbf{SR} = 2\mathbf{I}$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_1 + \|\mathbf{m}_{:j}\|_1$
  - Ancien χ-Sim: $h = 1$, $\mathcal{N} = |\mathbf{m}_{:i}| \times |\mathbf{m}_{:j}|$
  - Cosinus généralisé: $\mathbf{SR} > 0$, $h = 1$, $\mathcal{N} = \|\mathbf{m}_{:i}\|_{\mathbf{SR}} \times \|\mathbf{m}_{:j}\|_{\mathbf{SR}}$
  - χ-Sim$_p^h$: $\mathcal{N} = \|\mathbf{m}_{:i}\|_{\mathbf{SR}}^h \times \|\mathbf{m}_{:j}\|_{\mathbf{SR}}^h$

Apprentissage – Paramètre de seuillage p

Dans un tel corpus...

De nombreux mots ne sont pas suffisamment spécifiques, et créent des similarités non pertinentes. On considère ces similarités comme étant du bruit.

Exemple: Astronomie / Mythologie

Le mot *Hercules* peut apparaître dans un document d'astronomie (constellation Hercule de l'hémisphère nord) et dans les documents de mythologie traitant des exploits du héros grec du même nom...

Apprentissage – Paramètre de seuillage p

Dans un tel corpus...

De nombreux mots ne sont pas suffisamment spécifiques, et créent des similarités non pertinentes. On considère ces similarités comme étant du bruit.

Exemple: Astronomie / Mythologie

Le mot *Hercules* peut apparaître dans un document d'astronomie (constellation Hercule de l'hémisphère nord) et dans les documents de mythologie traitant des exploits du héros grec du même nom...

Comment traiter ce problème ?

Hypothèse : ces similarités non pertinentes sont de faible valeur.
$\rightarrow$ on supprime les $p\%$ des valeurs de similarité les plus faibles.

Apprentissage – Algorithme pour χ-SIM_p^h

1. $\mathbf{SR}^{(0)}$ et $\mathbf{SC}^{(0)}$ sont initialisées avec la matrice identité.

Apprentissage – Algorithme pour χ-SIM_p^h

1. $\mathbf{SR}^{(0)}$ et $\mathbf{SC}^{(0)}$ sont initialisées avec la matrice identité.
2. À chaque itération t, on met à jour les deux matrices de similarité :
    3. Màj de $\mathbf{SR}^{(t)}$ en utilisant $\mathbf{SC}^{(t-1)}$
    5. Màj de $\mathbf{SC}^{(t)}$ en utilisant $\mathbf{SR}^{(t-1)}$

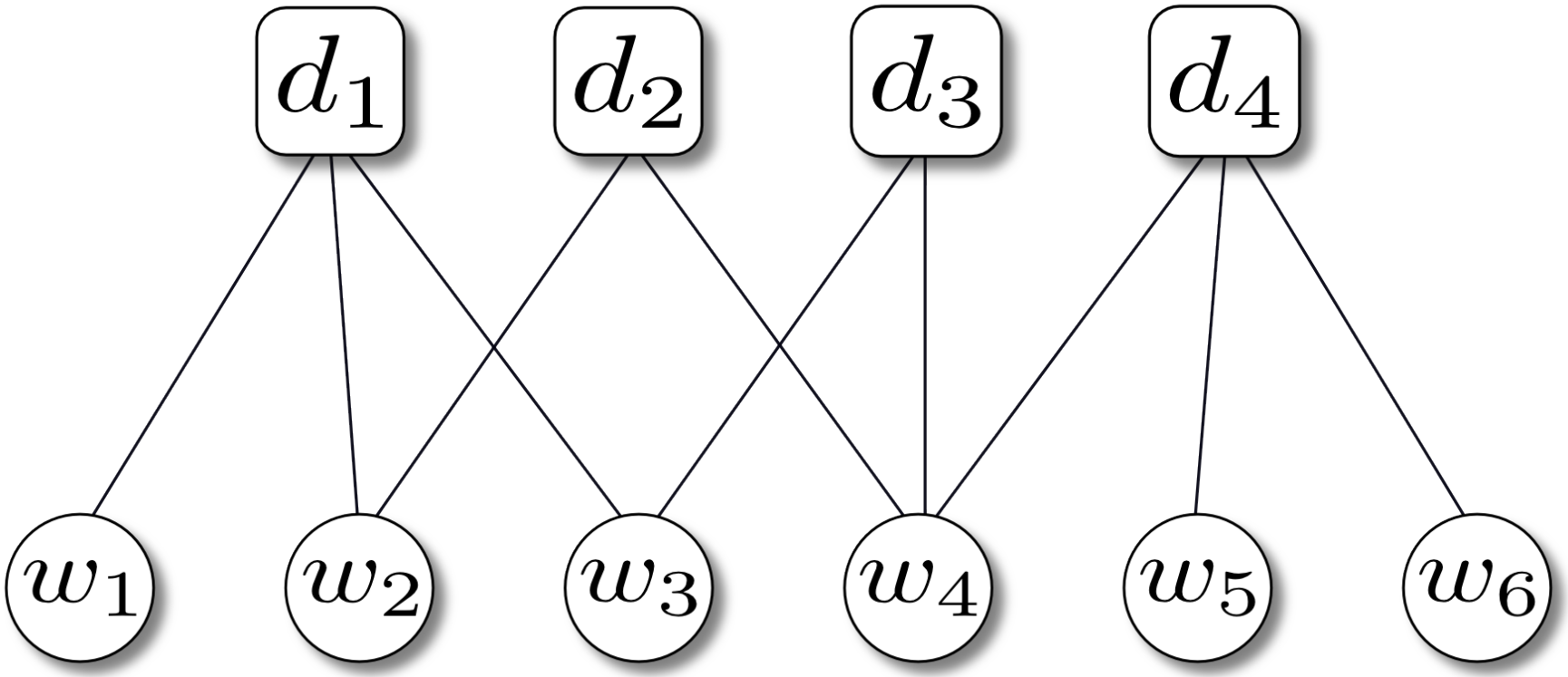
Apprentissage – Algorithme pour χ-SIM_p^h

1. $\mathbf{SR}^{(0)}$ et $\mathbf{SC}^{(0)}$ sont initialisées avec la matrice identité.
2. À chaque itération t, on met à jour les deux matrices de similarité :
    3. Màj de $\mathbf{SR}^{(t)}$ en utilisant $\mathbf{SC}^{(t-1)}$
    5. Màj de $\mathbf{SC}^{(t)}$ en utilisant $\mathbf{SR}^{(t-1)}$

Classiquement, $t = 4$ est suffisant.

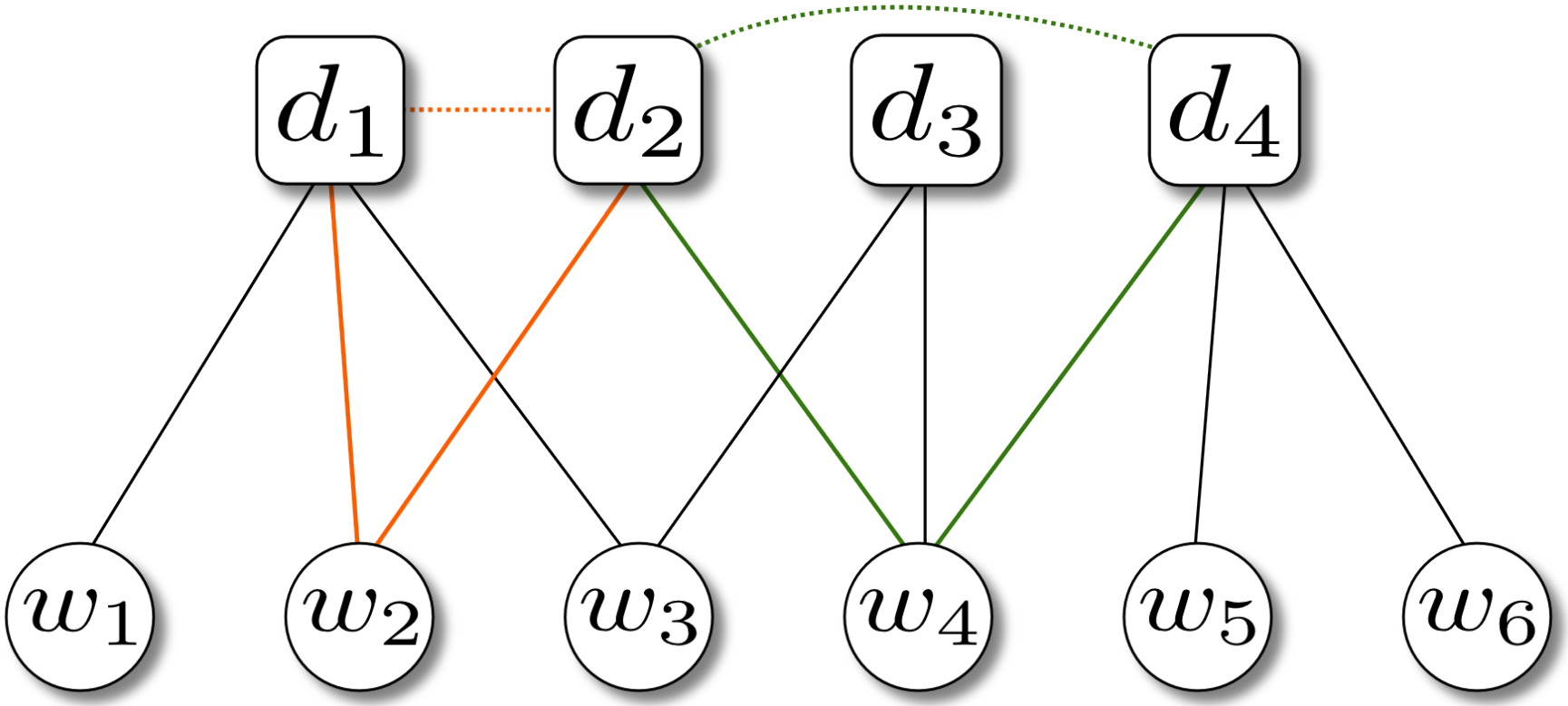
Apprentissage – Signification d'une itération de χ-SIM_p^h

Graphe bi-partie représentant un corpus



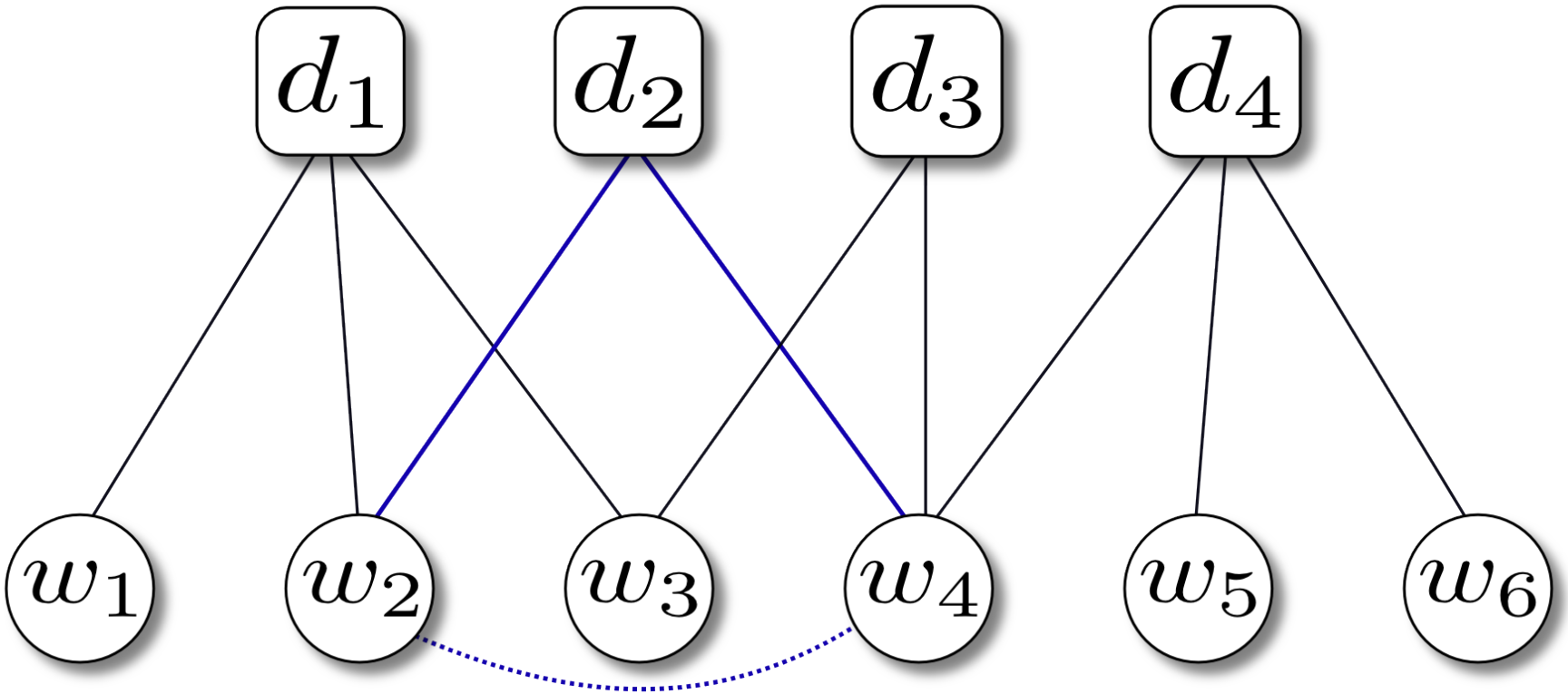
Apprentissage – Signification d'une itération de χ-SIM_p^h

Première itération : $sr_{12} > 0$ and $sr_{24} > 0$, but $sr_{14} = 0$



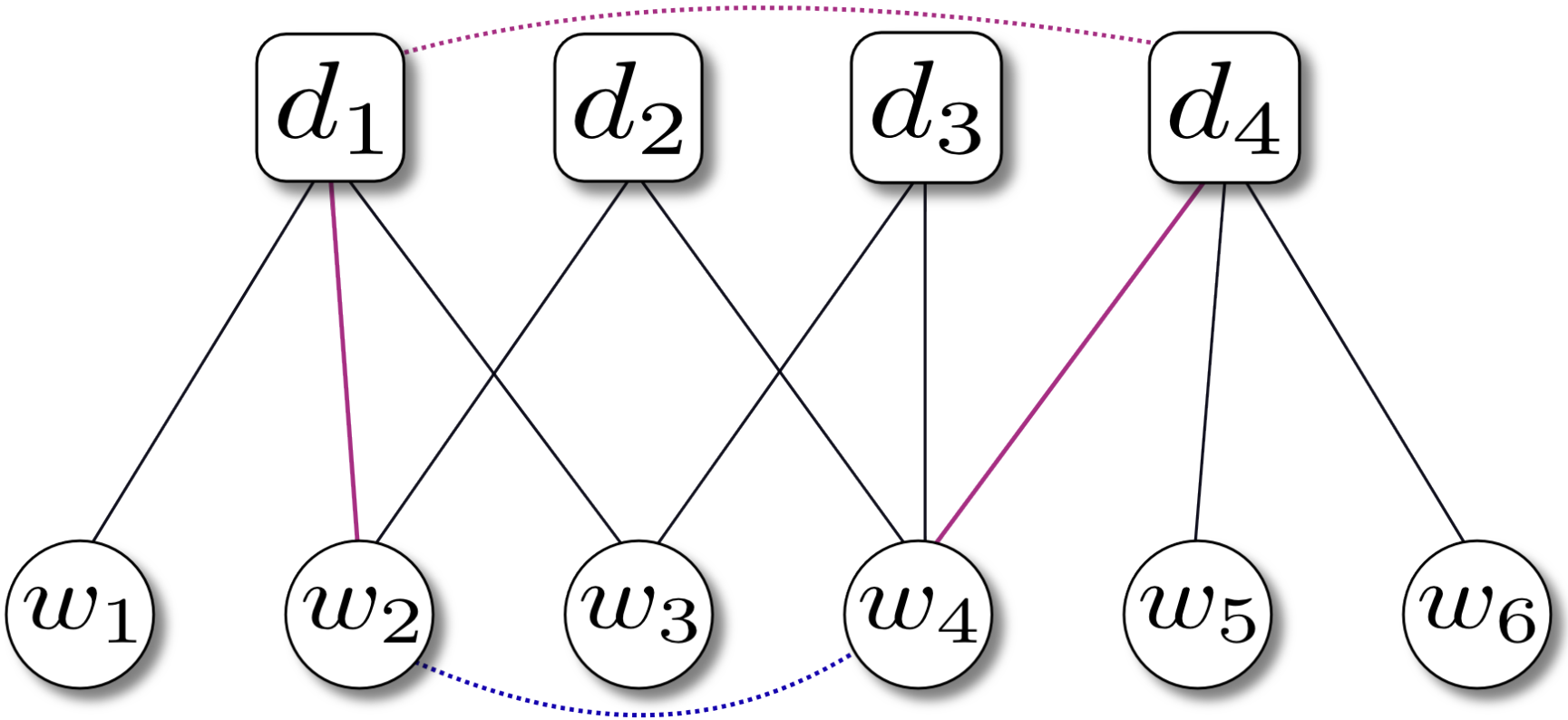
Apprentissage – Signification d'une itération de χ-SIM_p^h

Seconde partie de la première itération : $sc_{24} > 0$



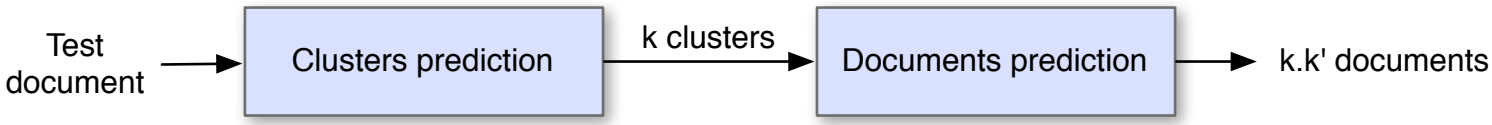
Apprentissage – Signification d'une itération de χ-SIM_p^h

Deuxième itération : à travers sc_{24}, now $sr_{14} > 0$



Plan

Classification des documents



Classification du document d_i test en deux étapes :

- Prédiction des k clusters les plus proches du document :
  $$\text{Sim}(d_i, C_j) = d_i \times W_{C_j}^T$$
- Prédiction des k' documents (de chacun des k clusters) les plus proches :
  $$\text{Sim}(d_i, Classe_l \ \ de \ \ C_j) = d_i \times \mathbf{SC}_j \times Classe_l^T$$

Mécanisme de vote : les $k\,k'$ documents votent pour leur classe et on choisit la classe ayant reçu le plus grand nombre de vote.

Plan

Expérimentation

Base DMOZ du challenge Pascal LSHTC de 2010

- 2387 catégories dont 1139 classes
- 4463 documents dans la base d'apprentissage
- 1859 documents dans la base de validation
- 1857 documents dans la base de test

Expérimentation

Base DMOZ du challenge Pascal LSHTC de 2010

- 2387 catégories dont 1139 classes
- 4463 documents dans la base d'apprentissage
- 1859 documents dans la base de validation
- 1857 documents dans la base de test

Meilleur résultat publié : 46.8% de précision

Notre meilleur résultat :

- 91 clusters
- 66.5% de précision Accuracy@1
- 66.3% de précision sur l'ensemble d'apprentissage
- 28.3% de précision sur l'ensemble de validation
- 28.3% de précision sur l'ensemble de test

Conclusion et perspectives

Conclusion

- Développement d'un cadre pour la catégorisation à grande échelle de documents par approche de calcul de co-similarité
- Résultats expérimentaux encore insuffisants mais encourageants en vues des améliorations prévues
- Utilisation de χ-SIM_p^h pour traiter la variabilité terminologique

Conclusion et perspectives

Conclusion

- Développement d'un cadre pour la catégorisation à grande échelle de documents par approche de calcul de co-similarité
- Résultats expérimentaux encore insuffisants mais encourageants en vues des améliorations prévues
- Utilisation de χ-SIM_p^h pour traiter la variabilité terminologique

Perspectives

- Tester différentes méthodes de sélection de mots
- Intégrer la co-similarité au niveau inter-cluster
- Généralisation de cette approche à plus de deux niveaux
- Tester sur d'autres bases de données (plus grandes)

Merci beaucoup pour votre attention.

References


S. F. Hussain, C. Grimal, and G. Bisson. An improved co-similarity measure for document clustering. In *Proceedings of the 9^{th} ICMLA*, 2010.


S. F. Hussain and G. Bisson. A supervised Approach to Text Categorization using Higher Order Co-Occurrences. In *SDM 2010*, Columbus, Ohio, 2010.


C. C. Aggarwal and A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Lecture Notes in Computer Science*, 420–434, 2001.


Bottou L.and Bousquet O. Learning using large datasets. *Mining Massive DataSets for security*, 2008, Citeseer.


Yang Y. and Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization, *ICML*, 1997, 412-420.

Parameter h

The generalized χ-SIM

$$\forall i,j \in 1..r,\ sr_{ij} = \frac{\mathrm{Sim}^h(\mathbf{m}_{i:}, \mathbf{m}_{j:})}{\sqrt{\mathrm{Sim}^h(\mathbf{m}_{i:}, \mathbf{m}_{i:})} \times \sqrt{\mathrm{Sim}^h(\mathbf{m}_{j:}, \mathbf{m}_{j:})}}$$

$$\forall i,j \in 1..c,\ sc_{ij} = \frac{\mathrm{Sim}^h(\mathbf{m}_{:i}, \mathbf{m}_{:j})}{\sqrt{\mathrm{Sim}^h(\mathbf{m}_{:i}, \mathbf{m}_{:i})} \times \sqrt{\mathrm{Sim}^h(\mathbf{m}_{:j}, \mathbf{m}_{:j})}}$$

For $h = 1$, $\mathbf{SR}$ and $\mathbf{SC}$ are not positive semi-definite...
We are not defining an inner product so it is not a generalized cosine measure...