Amélioration de la co-similarité pour la classification de documents

Clément Grimal et Gilles Bisson

Laboratoire d'Informatique de Grenoble – Equipe AMA
{prénom.nom}@imag.fr

Conférence Francophone sur l'Apprentissage Automatique
Chambéry, 20 mai 2011

UNIVERSITÉ DE GRENOBLE

Le contexte de la fouille de texte

Document#1 :
Une construction dont la principale fonction est d'abriter une famille.

Document#2 :
Un bâtiment qui sert d'habitation à des êtres humains.

Le contexte de la fouille de texte

Document#1 :
Une construction dont la principale fonction est d'abriter une famille.

Document#2 :
Un bâtiment qui sert d'habitation à des êtres humains.

Avec une approche classique :

Pas de termes en communs entre les deux documents

$\rightarrow \text{Similarity}(Document\#1, Document\#2) = 0$

Le contexte de la fouille de texte

Document#1 :
Une construction dont la principale fonction est d'abriter une famille.

Document#2 :
Un bâtiment qui sert d'habitation à des êtres humains.

Avec une approche classique :

Pas de termes en communs entre les deux documents

$\rightarrow \text{Similarity}(Document\#1, Document\#2) = 0$

En utilisant une approche basée sur la co-similarité :

Classification des termes

$\rightarrow \text{Similarity}(Document\#1, Document\#2) > 0$

Modèle

Représentation vectorielle classique de Salton (1971) :
$\mathbf{M}$: matrice documents/termes de r lignes et de 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

Modèle

Représentation vectorielle classique de Salton (1971) :
$\mathbf{M}$: matrice documents/termes de r lignes et de 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

Ce que nous voulons calculer :

- $\mathbf{SR}$: matrice de similarité pour les documents, avec $sr_{ij} \in [0, 1]$
- $\mathbf{SC}$: matrice de similarité pour les termes, avec $sc_{ij} \in [0, 1]$

Modèle

Représentation vectorielle classique de Salton (1971) :
$\mathbf{M}$: matrice documents/termes de r lignes et de 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

Ce que nous voulons calculer :

- $\mathbf{SR}$: matrice de similarité pour les documents, avec $sr_{ij} \in [0, 1]$
- $\mathbf{SC}$: matrice de similarité pour les termes, avec $sc_{ij} \in [0, 1]$

Idée de base

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

Nous allons contruire conjointement $\mathbf{SR}$ et $\mathbf{SC}$.

Plan

Similarité entre deux documents

- Approche classique : similarité = f(termes communs)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

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

Similarité entre deux documents

- Approche classique : similarité = f(termes communs)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

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

- En utilisant $\mathbf{SC}$ (en supposant que $sc_{ii} = 1$) :

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \mathrm{F_s}(m_{il}, m_{jl}) \times sc_{ll}$$

Similarité entre deux documents

- Approche classique : similarité = f(termes communs)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

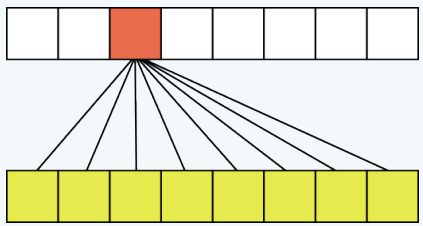
avec $\mathrm{F_s}$ une fonction de similarité (différence absolue, produit, etc.)

- En utilisant $\mathbf{SC}$ (en supposant que $sc_{ii} = 1$) :

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \mathrm{F_s}(m_{il}, m_{jl}) \times sc_{ll}$$

- Maintenant, on compare toutes les paires de termes :

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \sum_{n=1}^{c} \mathrm{F_s}\left(m_{il}, m_{jn}\right) \times sc_{ln}$$



Nouvelle approche - Pseudo-norme k

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

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

Nouvelle approche - Pseudo-norme k

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

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

- Nous introduisons une pseudo-norme k, voir [Aggarwal et al.(2001)] :

$$\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k} = \langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k$$

$\rightarrow$ nous avons $\|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k = \sqrt{\langle \mathbf{m}_{i:}, \mathbf{m}_{i:} \rangle_{\mathbf{SC}}^k}$

Nouvelle approche - Pseudo-norme k

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

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

- Nous introduisons une pseudo-norme k, voir [Aggarwal et al.(2001)] :

$$\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times (\mathbf{m}_{j:}^{\mathrm{T}})^k} = \langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k$$

$\rightarrow$ nous avons $\|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k = \sqrt{\langle \mathbf{m}_{i:}, \mathbf{m}_{i:} \rangle_{\mathbf{SC}}^k}$

- On normalise ensuite cette mesure de similarité :

$$sr_{ij} = \frac{\sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times (\mathbf{m}_{j:}^{\mathrm{T}})^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} \in [0, 1]$$

Forme générique

- Ainsi :

$$sr_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} = \frac{\langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})}$$

- Avec des valeurs particulières pour k, $\mathbf{SC}$ et $\mathcal{N}$, on a :
  - Jaccard : $\mathbf{SC} = \mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1 - \mathbf{m}_{i:}\mathbf{m}_{j:}^{\mathrm{T}}$
  - Dice : $\mathbf{SC} = 2\mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1$
  - χ-Sim “classique” : $k = 1$, $\mathcal{N} = |\mathbf{m}_{i:}| \times |\mathbf{m}_{j:}|$
  - Cosinus généralisé : $\mathbf{SC} > 0$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}} \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}$

Forme générique

- Ainsi :

$$sr_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} = \frac{\langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})}$$

- Avec des valeurs particulières pour k, $\mathbf{SC}$ et $\mathcal{N}$, on a :
  - Jaccard : $\mathbf{SC} = \mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1 - \mathbf{m}_{i:}\mathbf{m}_{j:}^{\mathrm{T}}$
  - Dice : $\mathbf{SC} = 2\mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1$
  - χ-Sim "classique" : $k = 1$, $\mathcal{N} = |\mathbf{m}_{i:}| \times |\mathbf{m}_{j:}|$
  - Cosinus généralisé : $\mathbf{SC} > 0$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}} \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}$
  - χ-Sim^k : $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}^k$

Etape de seuillage

Dans un corpus de texte...

De nombreux termes ne sont pas suffisamment spécifiques, ce qui induit beaucoup de similarités insignifiantes.
Ces similarités peuvent être considérées comme du bruit.

Exemple : Astronomie / Mythologie

Le mot *Hercule* peut apparaître dans une document d'astronomie pour désigner la constellation, et ainsi le "lier" à tous les documents de mythologie traitant du héros grec...

Etape de seuillage

Dans un corpus de texte...

De nombreux termes ne sont pas suffisamment spécifiques, ce qui induit beaucoup de similarités insignifiantes.
Ces similarités peuvent être considérées comme du bruit.

Exemple : Astronomie / Mythologie

Le mot *Hercule* peut apparaître dans une document d'astronomie pour désigner la constellation, et ainsi le "lier" à tous les documents de mythologie traitant du héros grec...

Comment traiter ce problème ?

Hypothèse : ces similarités sont faibles.
$\rightarrow$ on force à zéro les plus petites valeurs des matrices de similarités

Algorithme pour χ-Sim$_p^k$

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

Algorithme pour χ-Sim$_p^k$

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

2. À chaque itération t, on met à jour les deux matrices de similarités :

    3. Mise à jour de $\mathbf{SR}^{(t)}$ à l'aide de $\mathbf{SC}^{(t-1)}$
    4. Seuillage de $\mathbf{SR}^{(t)}$: $p\%$ des plus petites valeurs à 0

    5. Mise à jour de $\mathbf{SC}^{(t)}$ à l'aide de $\mathbf{SR}^{(t-1)}$
    6. Seuillage de $\mathbf{SC}^{(t)}$: $p\%$ des plus petites valeurs à 0

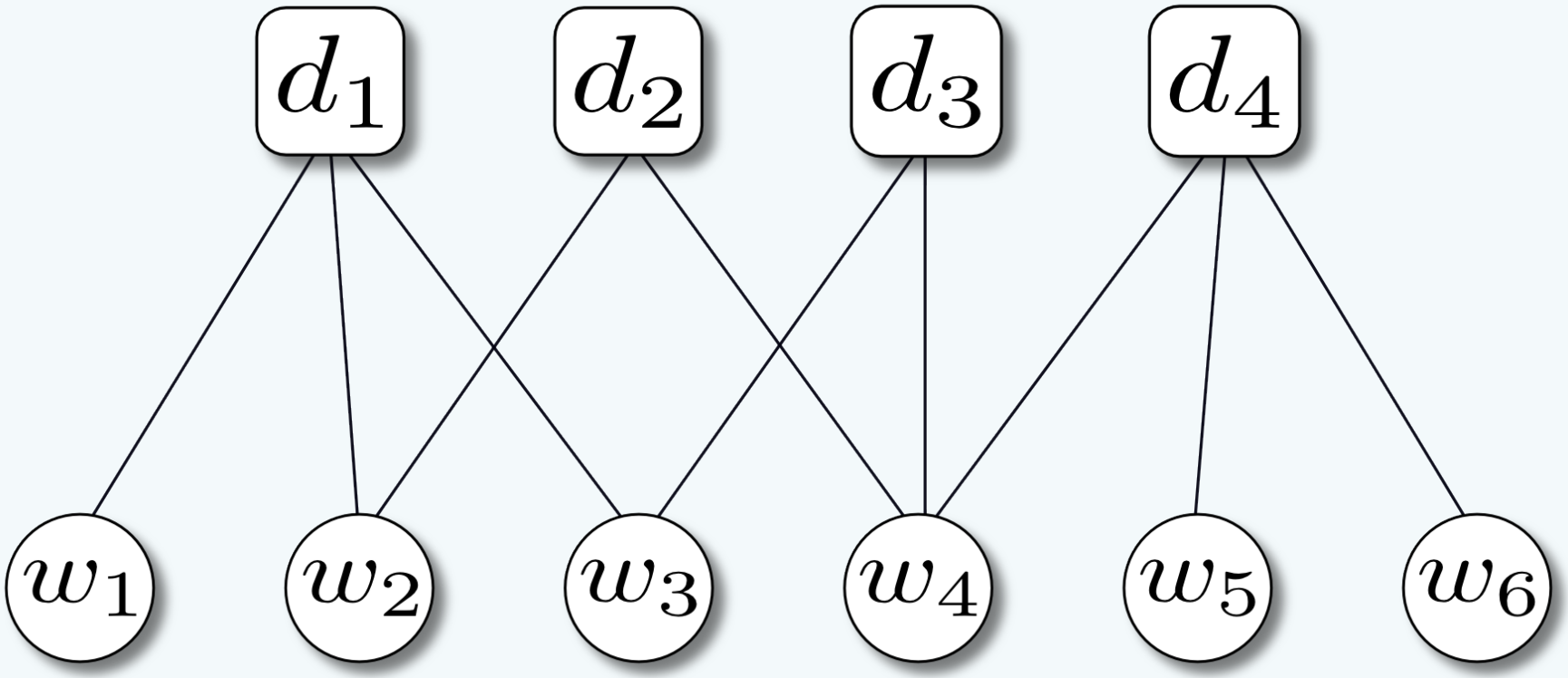
Algorithme pour χ-Sim$_p^k$

1. $\mathbf{SR}^{(0)}$ et $\mathbf{SC}^{(0)}$ sont initialisées par la matrice identité.
2. À chaque itération t, on met à jour les deux matrices de similarités :
   3. Mise à jour de $\mathbf{SR}^{(t)}$ à l'aide de $\mathbf{SC}^{(t-1)}$
   4. Seuillage de $\mathbf{SR}^{(t)}$: $p\%$ des plus petites valeurs à 0
   5. Mise à jour de $\mathbf{SC}^{(t)}$ à l'aide de $\mathbf{SR}^{(t-1)}$
   6. Seuillage de $\mathbf{SC}^{(t)}$: $p\%$ des plus petites valeurs à 0

En pratique, $t = 4$ est satisfaisant.

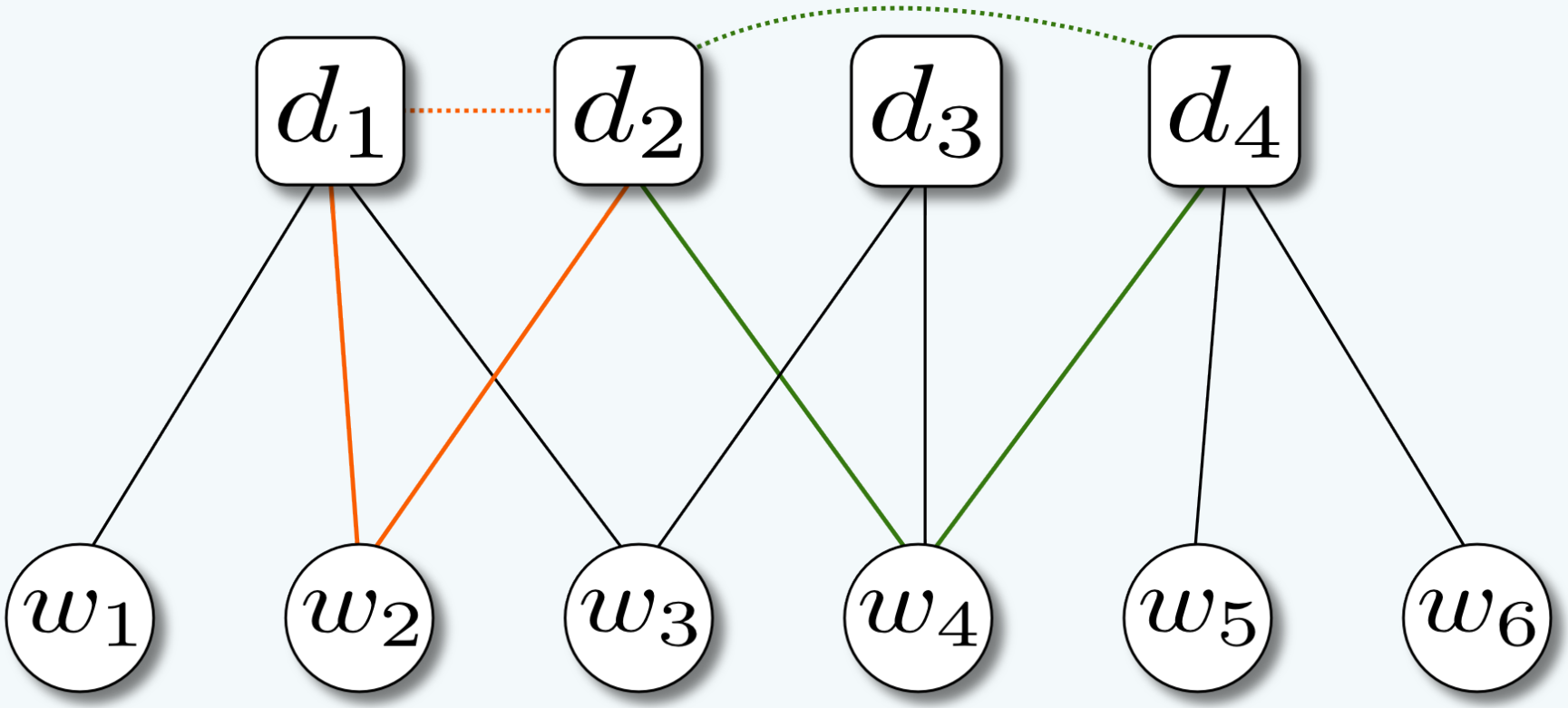
Signification d'une itération

Graphe bi-partie représentant un petit corpus



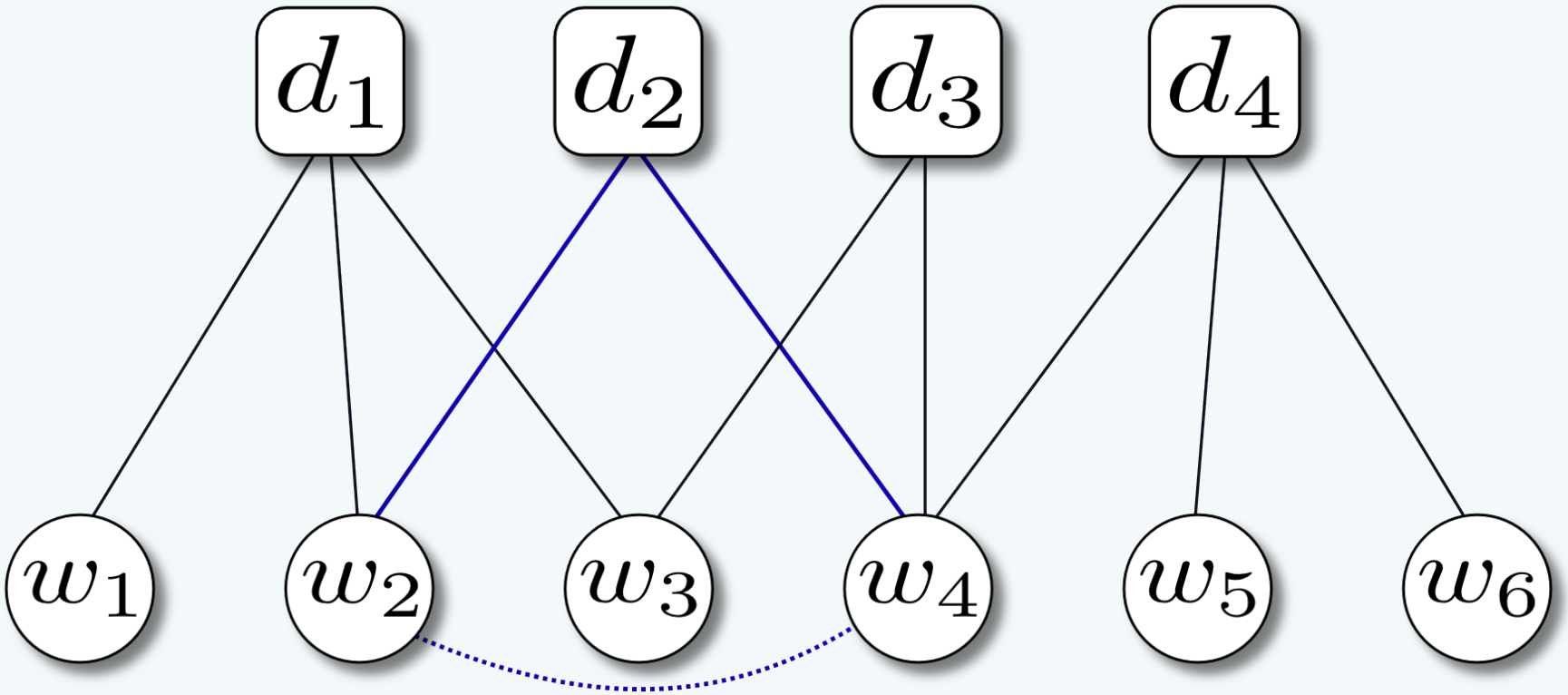
Signification d'une itération

Première itération : $sr_{12} > 0$ et $sr_{24} > 0$, mais $sr_{14} = 0$



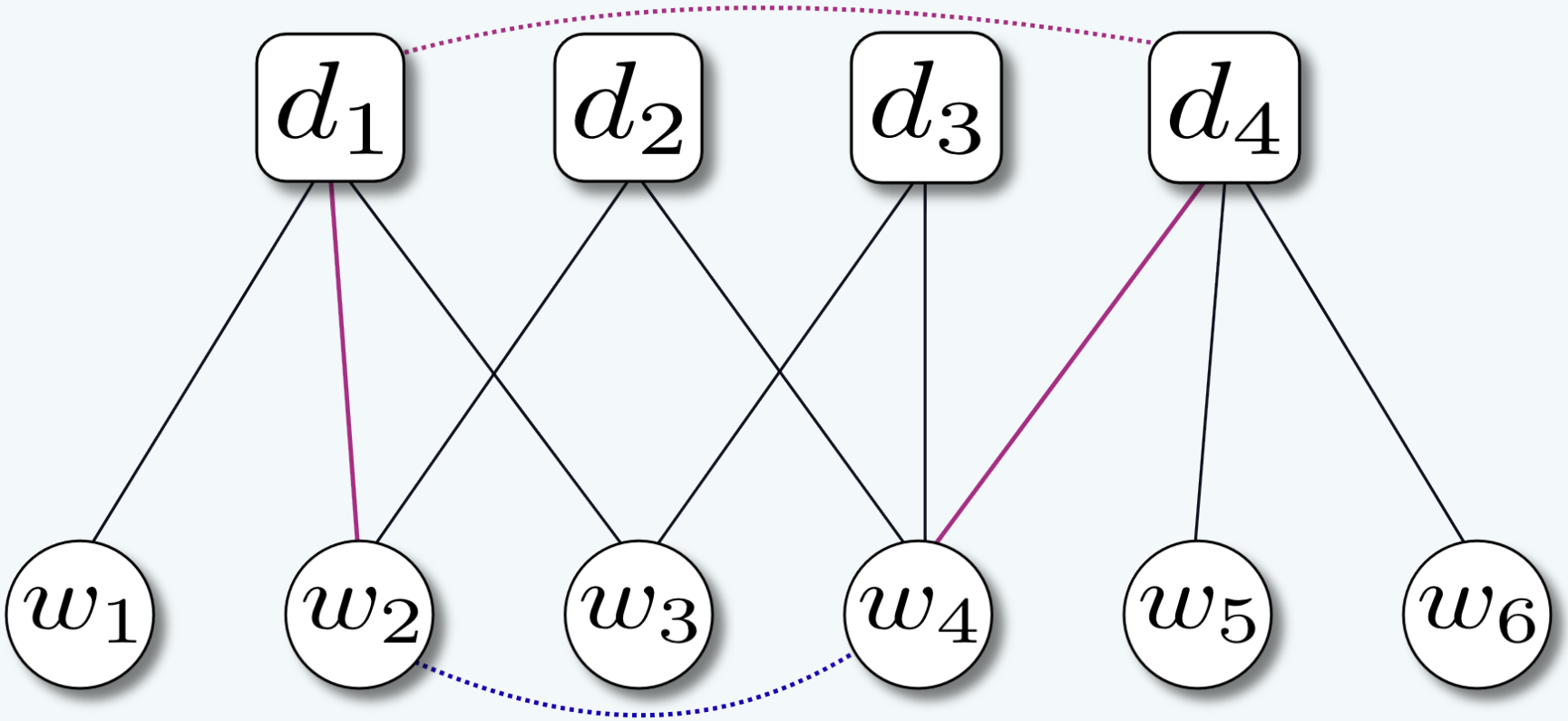
Signification d'une itération

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



Signification d'une itération

Deuxième itération : par sc_{24}, on a $sr_{14} > 0$



Plan

Méthodes comparées

- Cinq mesures de similarité
  - Le Cosinus
  - χ-Sim (avec ou sans k et p) [Hussain et al.(2010)]
  - LSA (Latent Semantic Analysis) [Deerwester et al.(1990)]
  - SNOS (Similarity in Non-Orthogonal Space) [Liu et al.(2004)]
  - CTK (Commute Time Kernel) [Yen et al.(2009)]

  $+$ Classification Ascendante Hiérarchique, avec l'indice de Ward

- Trois méthodes de co-classification
  - ITCC (Information Theoric Co-Clustering) [Dhillon et al.(2003)]
  - BVD (Block Value Decomposition) [Long et al.(2005)]
  - RSN (k-partite graph partioning algorithm) [Long et al.(2006)]

Méthodologie et Données

Méthodologie

- Sélection aléatoire de sous-ensembles de documents déjà étiquetés
- Mesure la qualité des classes à l'aide la précision micro-moyennée

Les sous-ensembles :

Nom	Newsgroups inclus	#classes	#docs.
M2	talk.politics.mideast, talk.politics.misc	2	500
M5	comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast	5	500
M10	alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.gun	10	500
NG1	rec.sports.baseball, rec.sports.hockey	2	400
NG2	comp.os.ms-windows.misc, comp.windows.x, rec.motorcycles, sci.crypt, sci.space	5	1000
NG3	comp.os.ms-windows.misc, comp.windows.x, misc.forsale, rec.motorcycles, sci.crypt, sci.space, talk.politics.mideast, talk.religion.misc	8	1600

Résultats

	M2	M5	M10	NG1	NG2	NG3
Cosinus	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02

Résultats

	M2	M5	M10	NG1	NG2	NG3
Cosinus	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02
χ-Sim	0.58 ± 0.07	0.62 ± 0.12	0.43 ± 0.04	0.54 ± 0.03	0.60 ± 0.12	0.47 ± 0.05
χ-Sim$_p$	0.65 ± 0.09	0.68 ± 0.06	0.47 ± 0.04	0.62 ± 0.12	0.63 ± 0.14	0.57 ± 0.04

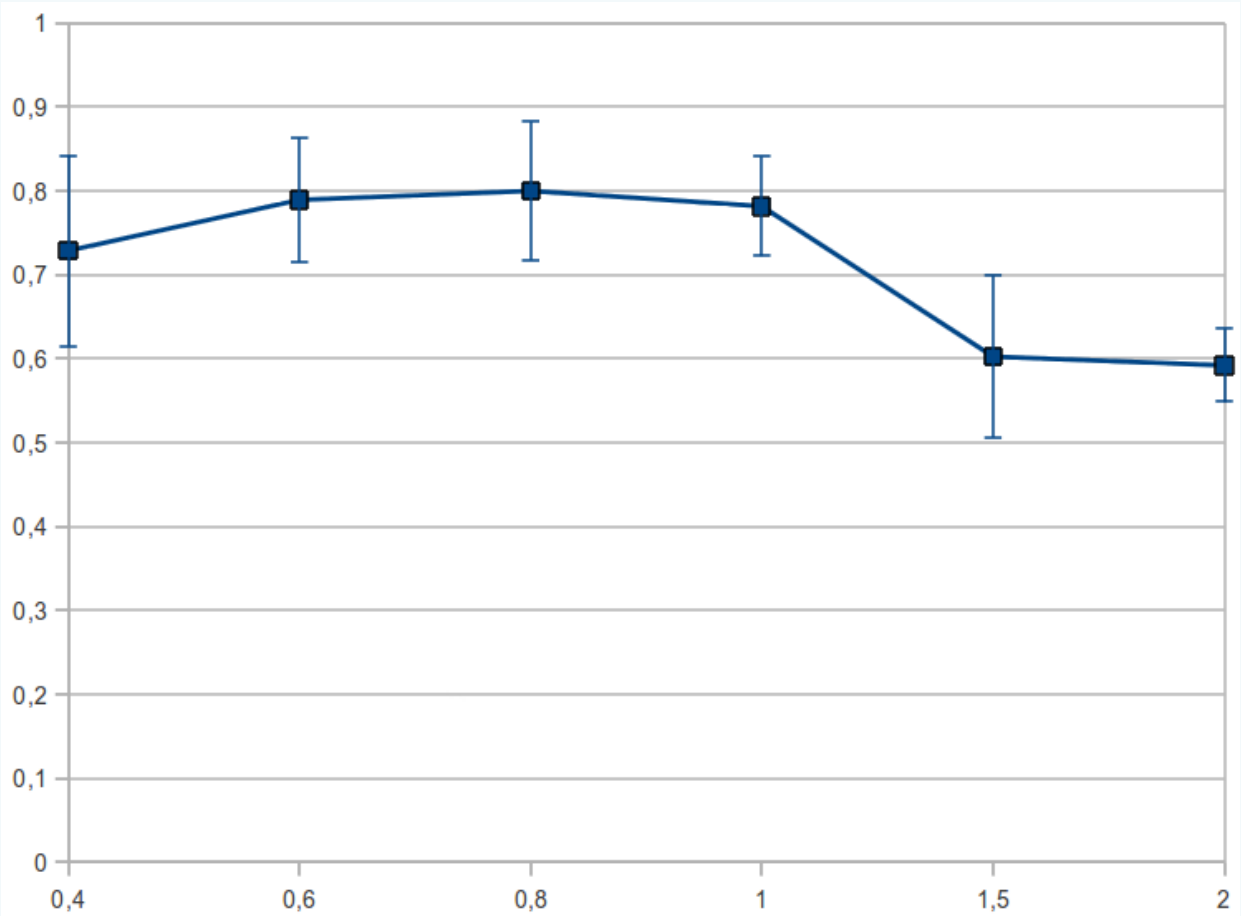
Résultats

	M2	M5	M10	NG1	NG2	NG3
Cosinus	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02
χ-Sim	0.58 ± 0.07	0.62 ± 0.12	0.43 ± 0.04	0.54 ± 0.03	0.60 ± 0.12	0.47 ± 0.05
χ-Sim$_p$	0.65 ± 0.09	0.68 ± 0.06	0.47 ± 0.04	0.62 ± 0.12	0.63 ± 0.14	0.57 ± 0.04
χ-Sim$_p^{0.8}$	$\mathbf{0.81} \pm 0.10$	$\mathbf{0.79} \pm 0.05$	$\mathbf{0.55} \pm 0.04$	$\mathbf{0.81} \pm 0.02$	$\mathbf{0.72} \pm 0.02$	$\mathbf{0.64} \pm 0.04$

Encore plus de résultats dans le papier...

Influence de k sur la précision

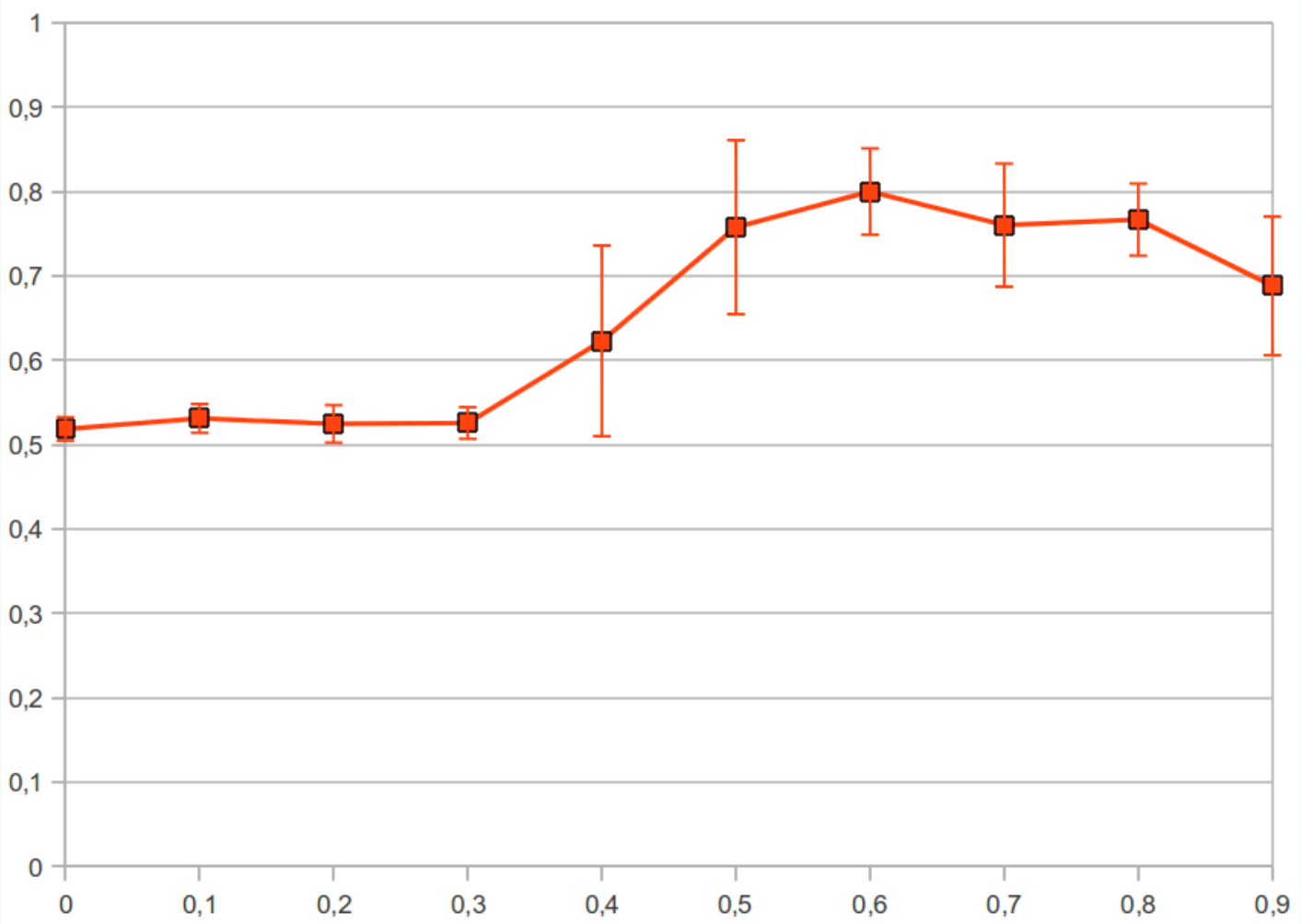
Tests sur NG1 (écart-type représenté par les barres d'erreurs)



Voir [Aggarwal et al.(2001)].

Influence de p sur la précision

Tests sur NG1 (écart-type représenté par les barres d'erreurs)



Conclusion & Perspectives

Améliorations de χ-Sim

- Exploration de différents espaces normés (k)
- Seuillage des matrices de similarités (p)
- Très bons résultats expérimentaux

Conclusion & Perspectives

Améliorations de χ-Sim

- Exploration de différents espaces normés (k)
- Seuillage des matrices de similarités (p)
- Très bons résultats expérimentaux

Perspectives

- Utiliser un facteur d'amortissement pour diminuer l'influence des co-occurrences d'ordre supérieur
- Trouver automatiquement les meilleurs valeurs pour k et p
- Obtenir une meilleure compréhension théorique
- Utiliser les matrices de similarité calculées par χ-Sim comme noyaux

Merci de votre attention !

References

S. Deerwester, S. T. Dumais, G. W. Furnas, Thomas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9^{th} ACM SIGKDD*, pages 89–98, 2003.

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

N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W. ying Ma. Learning similarity measures in non-orthogonal space. In *Proceedings of the 13^{th} ACM CIKM*, pages 334–341. ACM Press, 2004.

B. Long, Z. M. Zhang, and P. S. Yu. Co-clustering by block value decomposition. In *Proceedings of the 11^{th} ACM SIGKDD*, pages 635–640, New York, NY, USA, 2005. ACM.

B. Long, Z. M. Zhang, X. Wú, and P. S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23^{rd} ICML*, pages 585–592, New York, NY, USA, 2006. ACM.

L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering with the sigmoid commute-time kernel: A comparative study. *Data Knowl. Eng.*, 68(3):338–361, 2009.

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.

Paramètre k

χ-Sim généralisé

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

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

Pour $k = 1$, $\mathbf{SR}$ et $\mathbf{SC}$ ne sont pas semi-définies positives...
Nous ne définissons pas de produit scalaire, il ne s'agit donc pas d'une mesure de cosinus généralisée...