Un cadre général pour les mesures de co-similarité

Clément Grimal, Gilles Bisson

Un cadre général pour les mesures de co-similarité.

Clément Grimal *, Gilles Bisson **

*UJF-Grenoble 1/CNRS - Université de Grenoble
**CNRS - Université de Grenoble
LIG UMR 5217 / AMA team
{clement.grimal, gilles.bisson@imag.fr}

Résumé. Nous proposons ici une formulation générale de la notion de co-similarité qui permet d'effectuer des co-classifications conjointes d'objets et de descripteurs – par exemple des matrices documents/termes – en utilisant les algorithmes classiques de classification.

1 Contexte

En classification, lorsque ces caractéristiques décrivant une collection d'instances correspondent à un type de donnée homogène et qu'elles sont sémantiquement comparables, il est possible de les classifier au même titre que les instances. L'objectif de la co-classification (*co-clustering*) est de prendre en compte cette dualité afin de faire émerger automatiquement des regroupements plus pertinents : ainsi, dans le contexte des données textuelles, il est clair que la ressemblance entre documents dépend de la ressemblance entre les termes qui les composent, et non de leur seule identité, et réciproquement pour les termes.

Cette approche est largement étudiée dans le contexte de la *bioinformatique* et de la *fouille de textes*. Dans ce domaine, elle permet de surmonter le double problème de la faible densité des données (*sparsity*) et de la taille élevée du nombre des caractéristiques (*curse of dimensionality*). Parmi les nombreuses approches proposées, l'une des plus connue est l'analyse sémantique latente (LSA) introduite par Deerwester et al. (1990). Ces travaux repose sur le fait qu'un être humain utilise un vocabulaire varié pour décrire un même thème. Par exemple, si l'on considère un premier corpus contenant plusieurs co-occurrences des termes *océan* et *vagues* et un second contenant les termes *mer* et *vagues*, on peut inférer par transitivité que les termes *océan* et *mer* sont possiblement sémantiquement reliés. Cette association correspond à une co-occurrence du second ordre (une seule indirection) et peut être généralisée à des ordres supérieurs.

Nous avons développé une mesure nommée χ-`Sim` (Bisson et Hussain (2008); Hussain et al. (2010)) qui capture ces régularités par l'intermédiaire de la notion de *co-similarité*. L'idée est de construire de manière conjointe les deux matrices de similarité entre documents et mots, chacune prenant en compte durant le calcul les informations fournies par les autres. Ces matrices permettent ensuite de faire de la co-classification en utilisant des algorithme de classification standards tels les k-means, la CAH, ...

2 Notations utilisées

Les matrices (capitales) et les vecteurs (minuscules), sont en gras alors que les variables sont en italique.

- *Matrice de données :* $\mathbf{M}$ représente la matrice de données de n_a lignes et de n_b colonnes avec m_{ij} correspondant à l'intensité du lien entre l'objet représenté par la ligne i (a_i) et l'objet représenté par la colonne j (b_j). Nous utiliserons également une notation vectorielle pour les lignes $\mathbf{m_{i:}} = [m_{i1} \cdots m_{in_b}]$ et pour les colonnes $\mathbf{m_{:j}} = [m_{1j} \cdots m_{n_aj}]$. Dans la suite, nous nous réfèrerons à a_i quand nous nous intéresseront à l'objet i de type A, et nous utiliserons $\mathbf{m_{i:}}$ pour dénoter son vecteur.
- *Matrices de similarité :* $\mathbf{SA}$ et $\mathbf{SB}$ représentent respectivement les matrices (carrées et symétriques) de similarité pour les objets de type A (de taille $n_a \times n_a$) et les objets de type B (de taille $n_b \times n_b$). Ainsi $\forall i, j = 1..n_a, sa_{ij} \in [0, 1]$ et $\forall i, j = 1..n_b, sb_{ij} \in [0, 1]$.
- *Fonction de similarité :* la fonction générique $\mathrm{F_s}(\cdot, \cdot)$ prend en argument deux éléments de $\mathbf{M}$, m_{il} et m_{jn} et retourne une mesure de similarité entre ces deux éléments $\mathrm{F_s}(m_{il}, m_{jn})$.

3 Fondements de la mesure χ-Sim

Classiquement, la mesure de similarité (ou de distance) entre deux objets a_i et a_j est définie comme une fonction – notée ici $\mathrm{Sim}(a_i, a_j)$ – qui ne dépend que des « éléments » communs entre objets. On peut ajouter d'autres éléments, par exemple pour normaliser la valeur, mais fondamentalement l'idée reste la même, la conséquence étant que la similarité entre deux objets ne partageant aucune information est nulle :

$$\mathrm{Sim}(a_i, a_j) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc}) \tag{1}$$

Maintenant, supposons que l'on dispose d'un matrice $\mathbf{SB}$ dont les éléments sont les mesures de similarité entre les objets représentés par les colonnes de la matrice de données (ici les mots des documents). Simultanément, introduisons, par analogie à la norme L_k (distance de Minkowski), la notion de *pseudo-norme k*. Ainsi, si $\mathbf{SB} = \mathbf{I}$ et $k = 1$, l'équation (1) peut être réécrite comme :

$$\mathrm{Sim}^k(a_i, a_j) = \sqrt[k]{\sum_{l=1}^{n_b} \left(\mathrm{F_s}(m_{il}, m_{jl})\right)^k \times sb_{ll}} \tag{2}$$

Maintenant, nous allons généraliser (2) afin de prendre en compte, non plus uniquement les éléments communs aux deux objets, mais également l'ensemble des paires d'éléments. De la sorte, nous devenons capable de « capturer » non seulement la similarité entre les éléments identiques mais aussi celle provenant d'éléments différents : dans le cas de corpus, on devient ainsi potentiellement capable de comparer des documents contenant des termes différents. Bien sûr, pour chaque paire d'éléments b_l et b_n qui ne sont pas directement partagés par a_i et a_j, nous pondérons leur contribution à la similarité $\mathrm{Sim}(a_i, a_j)$ par leur propre similarité normalisée sb_{ln}. Cette nouvelle similarité entre les deux objets a_i et a_j est définie dans l'équation (3) dans laquelle les termes dont $l = n$ sont ceux qui apparaissaient dans (2) :

$$\mathrm{Sim}^k(a_i, a_j) = \sqrt[k]{\sum_{l=1}^{n_b} \sum_{n=1}^{n_b} \left(\mathrm{F_s}\left(m_{il}, m_{jn}\right)\right)^k \times sb_{ln}} \tag{3}$$

En supposant, comme pour le cosinus, que $\mathrm{F_s}(m_{ij}, m_{jn})$ correspond au produit de ses deux arguments, i.e. $\mathrm{F_s}(m_{ij}, m_{jn}) = m_{il} \times m_{jn}$, nous pouvons réécrire (3) sous la forme du produit matriciel suivant :

$$\mathrm{Sim}^k(a_i, a_j) = \mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SB} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k} \tag{4}$$

avec $\left(\mathbf{m}_{i:}\right)^k = \left[\left(m_{ij}\right)^k \cdots \left(m_{ic}\right)^k\right]$ et $\mathbf{m}_{j:}^{\mathrm{T}}$ représentant le vecteur transposé de $\mathbf{m}_{j:}$.

Finalement, nous pouvons introduire un terme de normalisation – noté $\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})$ – afin que toutes les mesures de similarité soient comprises dans l'intervalle $[0, 1]$. Nous obtenons alors l'équation ci-dessous (5), dans laquelle sr_{ij} représente un élément de la matrice $\mathbf{SA}$:

$$sa_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SB} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} \tag{5}$$

On peut remarquer que l'équation (5) généralise différentes mesures de similarité classiques :

- *L'indice de Jaccard* peut être obtenu pour les valeurs de paramètres suivantes : $k = 1$, $\mathbf{SB} = \mathbf{I}$ (la matrice identité), et $\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1 - \mathbf{m}_{i:}\mathbf{m}_{j:}^{\mathrm{T}}$
- *Le coefficient de Dice* correspond à $k = 1$, $\mathbf{SB} = 2\mathbf{I}$, et $\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1$
- *La mesure de similarité du cosinus généralisé (Qamar et Gaussier, 2009)* est obtenu lorsque $\mathbf{SB}$ est une matrice *semi-définie positive (SDP)* notée $\mathbf{A}$. Sous cette hypothèse, on peut donc définir le produit scalaire $< \mathbf{m}_{i:}, \mathbf{m}_{j:} >_{\mathbf{A}} = \mathbf{m}_{i:} \times A \times \mathbf{m}_{j:}^{\mathrm{T}}$, ainsi que la norme associée notée $\|\mathbf{m}_{i:}\|_{\mathbf{A}} = < \mathbf{m}_{i:}, \mathbf{m}_{i:} >_{\mathbf{A}}$. On définit alors $\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt{\|\mathbf{m}_{i:}\|_{\mathbf{A}}} \times \sqrt{\|\mathbf{m}_{j:}\|_{\mathbf{A}}}$.

3.1 Fonction de normalisation de χ-Sim

Nous allons introduire un nouveau schéma de normalisation, que nous nommerons *pseudo-normalisation*, et qui s'inspire de la mesure de similarité du cosinus généralisé, en relaxant la contrainte sur la propriété SDP de la matrice $\mathbf{A}$ et en ajoutant le paramètre k de pseudo-norme. En utilisant le produit matriciel (4) introduit dans la section 3 nous définissons symétriquement les éléments des matrices $\mathbf{SA}$ et $\mathbf{SB}$ ainsi :

$$sa_{ij} = \frac{\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:})}{\sqrt{\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{i:})} \times \sqrt{\mathrm{Sim}^k(\mathbf{m}_{j:}, \mathbf{m}_{j:})}} \tag{6a}$$

$$sb_{ij} = \frac{\mathrm{Sim}^k(\mathbf{m}_{:i}, \mathbf{m}_{:j})}{\sqrt{\mathrm{Sim}^k(\mathbf{m}_{:i}, \mathbf{m}_{:i})} \times \sqrt{\mathrm{Sim}^k(\mathbf{m}_{:j}, \mathbf{m}_{:j})}} \tag{6b}$$

Les équations (6a) et (6b) définissent donc *un système d'équations linéaires*, dont les solutions correspondent aux (co-)similarité entre les deux types d'objets dont la relation est décrite par la matrice de données $\mathbf{M}$. Par conséquent, l'algorithme χ-Sim est basé sur une approche itérative, i.e. un calcul alterné des valeurs des matrices $\mathbf{SA}$ et $\mathbf{SB}$. La normalisation assure que $sa_{ii} = 1$ et que $sb_{ii} = 1$ sans garantir toutefois que les indices de similarité soient toujours $\leqslant 1$. Dans le cas de données textuelles, cela correspond à des problèmes de polysémies de termes. Considérons ainsi un corpus contenant, parmi plusieurs autres documents, les documents d_1 contenant le mot *orange* et d_2 contenant les mots *rouge* et *banane*. Supposons qu'à une itération quelconque la matrice $\mathbf{SB}$ indique que que la valeur de similarité entre *orange* et *rouge* est 1, celle entre *orange* et *banane* est 1 et celle entre *rouge* et *banane* est 0. Dès lors, en appliquant les formules précédentes, $\mathrm{Sim}^1(\mathrm{d}_1, \mathrm{d}_1) = 1$, $\mathrm{Sim}^1(\mathrm{d}_2, \mathrm{d}_2) = 2$ and $\mathrm{Sim}^1(\mathrm{d}_1, \mathrm{d}_2) = 2$ et donc $sr_{12} = \frac{2}{\sqrt{1 \times 2}} > 1$. Ici, la similarité entre ces deux documents est surestimée à cause de la nature polysémique du mot *orange* qui présenter une double analogie avec la couleur *red* et le fruit *banane*. Expérimentalement, on observe que les valeurs sr_{ij} and sc_{ij} reste la plupart du temps inférieures ou égales à 1.

Parallèlement, lorsque l'on affecte à k des valeurs inférieures à 1 comme suggéré par Aggarwal et al. (2001) pour la norme L_k dans le contexte d'espaces de grandes dimensions, on observe une amélioration des résultats sur des test de classification (cf. Hussain et al. (2010)). Il faut cependant noter que nous sommes dans une situation différente de la norme L_k car notre méthode ne définit pas un espace vectoriel normé. Si l'on examine le cas simple où $k = 1$, on a alors $\mathrm{Sim}^1(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathbf{m}_{i:} \times \mathbf{SB} \times \mathbf{m}_{i:}^{\mathrm{T}}$, ce qui correspond à la forme générale d'un produit scalaire, à la condition que $\mathbf{SB}$ soit symétrique et semi-défini positive (SDP). Malheureusement, cette condition n'est pas nécessairement vérifié du fait de notre schéma de normalisation [1]. Aussi, notre mesure de similarité est simplement une forme bilinéaire dans un espace préhilbertien « dégénéré », dans lequel notre mesure correspond au Cosinus.

Pour résoudre complètement ces problèmes, il est possible de projeter les matrices $\mathbf{SB}$ et $\mathbf{SA}$ à chaque itération sur l'espace des matrices SDP comme le propose (Qamar et Gaussier, 2009). Ainsi, nous garantirions que le nouvel espace engendré soit bien un espace préhilbertien. Cependant, on constate expérimentalement que l'ajout d'une telle étape ne permet pas d'améliorer significativement les résultats de notre approche, car les matrices de similarités sont déjà très proches de l'espace des matrices SDP.

3.2 Traiter le 'bruit' dans les matrices de similarité

Si l'on considère le graphe biparti associé à une matrice documents/termes, on peut aisément montrer (cf. Hussain et al. (2010)) que, à l'itération n de l'algorithme, l'élément sa_{ij} de la matrice de similarité des lignes est fonction du nombre de *chemins d'ordre* n qui existent entre les objets i et j. Cependant, dans les jeux de données textuelles, certains termes sont rares et/ou ne sont pas spécifiques d'une classe d'objets. Dans le cas de données textuelles cela peut correspondre à des mots soit mal-orthographié, soit qui apparaissent de manière fortuite dans une classe de documents ; par exemple, le fait de trouver dans un document qu'une personne est « ... une nouvelle étoile au firmament ... » ne rattache a priori en rien ce document à la catégorie « astronomie ». On peut alors considérer ces termes comme une sorte de *bruit*

1. Une condition nécessaire pour que $\mathbf{SB}$ soit SDP serait que $\forall i, j \in 1..c, |sb_{ij}| \leqslant \sqrt{sb_{ii} \times sb_{jj}} = 1$.

dans les données car itérations après itérations, ces termes permettent à l'algorithme d'établir de nouveaux chemins erronés entre les différentes familles d'objets. Ces chemins induisent des similarités très faibles mais ils sont nombreux, et nous pouvons faire l'hypothèse qu'ils peuvent brouiller les « vraies » similarités. En se basant sur cette observation, nous introduisons dans l'algorithme χ-Sim une étape de *seuillage* associée à une paramètre p qui a pour objectif de supprimer ces infomations. Concrètement il va s'agir, à chaque itération, de remettre à 0 les p % des plus petites valeurs des matrices de similarité $\mathbf{SA}$ et $\mathbf{SB}$.

3.3 Un algorithme générique pour χ-$\mathtt{Sim}_p^k$

Les équations (6a) et (6b) nous permettent de calculer les similarités entre deux lignes et entre deux colonnes. L'extension à toutes les paires de lignes et toutes les paires de colonnes peut être formulée sous la forme d'une simple multiplication matricielle. Nous avons besoin de définir ici $\mathbf{M}^{\circ_k} = \left((m_{ij})^k\right)_{i,j}$ qui représentent la mise à la puissance k de la matrice $\mathbf{M}$. L'algorithme générique est le suivant :

1. On initialise les matrices de similarité $\mathbf{SA}$ et $\mathbf{SB}$ avec la matrice identité $\mathbf{I}$. En effet, on considère que l'on a pas de connaissance *a priori*, et que donc seul la similarité entre un objet et lui-même vaut 1. On note ces matrices $\mathbf{SA}^{(0)}$ et $\mathbf{SB}^{(0)}$.
2. À l'itération t, on calcule la nouvelle matrice de similarité $\mathbf{SA}^{(t)}$ en utilisant la matrice $\mathbf{SB}^{(t-1)}$:

$$\mathbf{SA}^{(t)} = \mathbf{M}^{\circ_k} \times \mathbf{SB}^{(t-1)} \times \left(\mathbf{M}^{\circ_k}\right)^{\mathrm{T}} \text{ and } sa_{ij}^{(t)} \leftarrow \frac{\sqrt[k]{sa_{ij}^{(t)}}}{\sqrt[2k]{sa_{ii}^{(t)} \times sa_{jj}^{(t)}}} \tag{7}$$

   On fait la même chose pour la matrice de similarité $\mathbf{SB}^{(t)}$ en utilisant $\mathbf{SA}^{(t-1)}$:

$$\mathbf{SB}^{(t)} = \left(\mathbf{M}^{\circ_k}\right)^{\mathrm{T}} \times \mathbf{SA}^{(t-1)} \times \mathbf{M}^{\circ_k} \text{ and } sb_{ij}^{(t)} \leftarrow \frac{\sqrt[k]{sb_{ij}^{(t)}}}{\sqrt[2k]{sb_{ii}^{(t)} \times sb_{jj}^{(t)}}} \tag{8}$$

3. On fixe à 0 les p % des plus petites valeurs des matrices de similarité $\mathbf{SA}$ et $\mathbf{SB}$.
4. Les étapes 2 et 3 sont répétées t fois (un nombre d'itérations de 4 est une valeur raisonnable pour des données textuelles) pour mettre à jour itérativement $\mathbf{SA}^{(t)}$ et $\mathbf{SB}^{(t)}$.

Il est important de remarquer que même si χ-$\mathtt{Sim}_p^k$ calcule une mesure de similarité entre chaque paire d'objets en utilisant toutes les paires de composantes des vecteurs les représentant, la complexité de l'algorithme reste comparable aux mesures de similarité classiques comme celle du Cosinus. En supposant que, pour une matrice générale de taille $n \times n$, la complexité de la multiplication matricielle est de $\mathcal{O}(n^3)$ et que la complexité pour calculer $\mathbf{M}^{\circ_k}$ est de $\mathcal{O}(n^2)$, la complexité totale de χ-$\mathtt{Sim}_p^k$ est donnée par $\mathcal{O}(tn^3)$.

Références

Aggarwal, C. C., A. Hinneburg, et D. A. Keim (2001). On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science*, pp. 420–434. Springer.

Bisson, G. et F. Hussain (2008). Chi-sim : A new similarity measure for the co-clustering task. In *Proceedings of the Seventh ICMLA*, pp. 211–217. IEEE Computer Society.

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

Hussain, S. F., C. Grimal, et G. Bisson (2010). An improved co-similarity measure for document clustering. In *International Conference on Machine Learning and Applications*.

Qamar, A. M. et E. Gaussier (2009). Online and batch learning of generalized cosine similarities. In *Proceedings of the Ninth IEEE ICDM*, Washington, DC, USA, pp. 926–931. IEEE Computer Society.