Analyse en composantes locales et graphes de similarité entre textes

Alain Lelu

INRA / unité Mathématiques, Informatique, Génome
et Université de Franche-Comté / LASELDI
alain.lelu@univ-fcomte.fr

Abstract

Cosines between line-vectors (or column-vectors) of a frequency [texts × terms] matrix define 2 valued graphs. Using distributional (or "Hellinger") distance, transition formulas allow one to deduce important properties of a graph from the other's : the first eigenvalue of each cosine adjacency matrix can be interpreted as a centrality index for texts and terms ; the following two factors provide an optimal 2D mapping for the graph, as correspondence analysis does. Our Local Components Analysis method uses these properties 1) globally: the extended concept of K-reciprocal neighbors defines a graph with adaptive density indices for its nodes, allowing for Trémolière's percolation algorithm to cluster it, 2) locally: each cluster is characterized by its first eigenvalues (resulting in texts and terms centrality indices in the cluster, and projections of the whole corpus'texts and terms on the cluster axoïd), and the following two ones (2D mapping of the K-reciprocal links). Illustration is given with Corneille and Molière theater corpus.

Résumé

Les matrices de cosinus entre vecteurs-ligne et colonne d'un tableau de fréquences [textes × termes] définissent des graphes valués sur ces deux ensembles. L'utilisation de la distance distributionnelle (dite aussi de Hellinger) fournit des relations de transition entre propriétés importantes calculées sur ces deux graphes : le 1er vecteur propre de chaque matrice d'adjacence fournit les indices de centralité des textes (resp. mots), les 2 suivants, proches des 2 premiers facteurs non triviaux d'AFC, permettent une représentation optimale plane des graphes. L'Analyse en Composantes Locales utilise ces propriétés : 1) au niveau global, la notion de voisinage K-réciproque détermine un graphe de contiguïté où chaque nœud est doté d'une valeur de densité *adaptative*, utilisée par l'algorithme de percolation de Trémolières pour déterminer des classes denses, 2) au niveau local chaque classe est caractérisée par ses 1ers vecteurs propres (définis comme ci-dessus : centralités des textes et des termes dans chaque classe) et les projections de tous les termes et textes du corpus sur l'axe des classes. Et par ses deux suivants (en tant que fond de carte pour représenter ses graphes de liens K-réciproques). Ces points sont illustrés à partir des pièces de Corneille et Molière indexées par D. Labbé.

Mots-clés : graphes, voisins réciproques, distance de Hellinger, densité, centralité, classification automatique, percolation, analyse en composantes locales.

1. Introduction

Les méthodes de fouille de textes, objets d'importants enjeux (génomique fonctionnelle, veille scientifique…) se heurtent à deux difficultés : 1) celle de la sélection des descripteurs (ici : termes lemmatisés ou non, composés ou non) pertinents, 2) celle de l'unicité de la représentation obtenue : seules des méthodes convergeant vers un maximum absolu d'un critère de qualité peuvent rendre compte de la variation dans le temps des représentations ; et il est toujours souhaitable pour l'expert qui les interprète de disposer d'une représentation stable et unique pour une collection donnée de textes.

De nombreux travaux récents sur les graphes à invariance d'échelle (scale-free) et dont la signature est spécifique (répartitions à lois de puissance, comme la « loi de Zipf » dans le domaine textuel), renouvellent le point de vue sur le domaine (Watts et Strogatz, 1998). Nous présentons ici une formulation en termes de graphes de nos travaux sur la représentation dynamique d'un flux de textes (Lelu et Ferhan, 1998 ; Lelu et François, 2003).

2. Constituer les graphes de similarité entre textes :

2.1. *Sélectionner les descripteurs pertinents :*

Disposer d'un bon extracteur de lemmes et termes composés n'est plus aujourd'hui une gageure. Mais la quantité de candidats termes obtenue, autant que le caractère « Zipfien » de la répartition des mots (il existe un continuum allant des mots courants non contextuels aux nombreux mots de fréquence faible, en passant par les mots de fréquence moyenne caractérisant tel ou tel contexte sémantique du corpus) exigent un processus d'élimination des deux types extrêmes de termes. Aujourd'hui nécessairement supervisé par un expert, ce processus doit être intégré dans une boucle itérative incluant une évaluation (automatique dans la mesure du possible) de la qualité de la représentation obtenue. Ce qui constitue un problème en soi, dont nous ne parlerons pas ici ; c'est sans aucune sélection de mots que nous illustrerons notre propos, sur le corpus des pièces de Corneille et Molière indexé par Dominique Labbé (2001) et mis à disposition de la communauté scientifique par celui-ci (67 œuvres caractérisées par 9946 termes grammaticalement typés, dont 3076 hapax, pour un total de 928 000 occurrences).

2.2. *Choix de l'indice de similarité*

Dans notre domaine d'application, plusieurs critères doivent caractériser cet indice :

– Ne pas dépendre de la taille des textes : les vecteurs caractérisant les textes doivent être normalisés.

– Ne pas dépendre de l'arbitraire du choix des mots par les auteurs quand ils expriment un certain contenu sémantique – souvent en référence à des critères esthétiques comme éviter de répéter le même mot à peu de distance : ce qui se traduit par le critère d'« équivalence distributionnelle » (stabilité des similarités par rapport à la fusion ou à l'éclatement de descripteurs de répartitions voisines), satisfait, entre autres, par la transformation de l'espace des données dans lequel opère l'analyse factorielle des correspondances (AFC).

Notre réponse à cette double exigence est remplie par la distance distributionnelle d(t,t') (dite aussi de Hellinger (Domengès et Volle, 1979)), évaluée dans (Lelu, 2003), et utilisée par ailleurs (Bacelar-Nicolau, 1987) : $\mathrm{Sim}(t, t') = 2\text{-}d(t, t')^2 = \cos(x_t, x_{t'}) = < \underline{x}_t, \underline{x}_{t'} >$

où $\underline{x}_t = \{\ldots \sqrt{(x_{it}/x_{.t})} \ldots\}$ est un vecteur-texte normalisé, x_{it} est la fréquence du mot i dans le texte t, $x_{.t}$ est la fréquence totale des mots du texte t, et <..., …> symbolise le produit scalaire.

2.3. *Du tableau des similarités au graphe de similarité*

Le tableau complet des similarités entre les textes du corpus peut être considéré comme la matrice d'adjacence d'un graphe valué porteur de la totalité de l'information disponible sur les similarités. Cependant, l'examen direct de celui-ci, à diverses valeurs de seuil de similarité, par des outils d'exploration de graphe (par ex. BioLayout (Enright et Ouzounis, 2001)) ne fait ressortir le plus souvent qu'une grande classe connexe et touffue entourée par quelques classes de faibles effectifs et beaucoup de points isolés, du fait de la répartition très inégalitaire, Zipfienne, des mots.

3. Exploiter les graphes de similarité entre textes :

Il est donc impératif de styliser ce graphe en « augmentant le contraste » de façon pertinente, si l'on veut isoler des cliques de documents thématiquement homogènes. Plusieurs méthodes ont été proposées pour ce problème de « graph partitioning » (pour une revue, cf. (Brandes *et al.*, 2003)), qui font le plus souvent appel à des processus itératifs.

Pour notre part nous avons opté pour une méthode utilisant une mesure de densité locale des nœuds d'un graphe, inspirée de l'algorithme de percolation de (Trémolières, 1979) : à paramètre de finesse d'analyse fixé, l'énumération des maxima locaux de densité constitue l'optimum global recherché. On obtient ainsi 1) un ensemble de noyaux de classes, 2) un ensemble de nœuds ambivalents, communs à plusieurs noyaux, 3) les nœuds isolés restants peuvent être projetés sur les axes caractérisant les noyaux (cf. §3.1) autorisant ainsi une notion de classes floues et recouvrantes, plus « naturelle » dans notre domaine d'application que celle de partition stricte. À noter que (Colombo *et al.* ,2003) utilisent le même principe de recherche de maxima de densité, avec un principe d'extension de ceux-ci moins riche que dans la méthode de Trémolières.

Reste le choix de la fonction densité ; après des essais décevants avec un noyau de taille fixe, paramétrable (Lelu, 1994), dus aux disparités considérables de densité des classes dans un espace à grand nombre de dimensions, nous avons adopté la mesure *adaptative* de densité suivante : 1) parmi les K plus proches voisins de chaque texte, définis (sans garantie de symétrie) sur la matrice de similarité, nous ne retenons que les voisins réciproques, 2) la densité de chaque nœud est une version modifiée du « clustering coefficient » (Albert et Barabasi, 2001) : 1 (si le noeud a un ou plusieurs 1-voisins) + (nombre d'arêtes de son 1-voisinage non rattachées au nœud considéré / nombre maximum possible de ces arêtes) ; ainsi un point isolé a une densité de 0, le centre d'une « étoile », ou chaque élément d'un couple isolé ont une densité de 1, et toute autre configuration une densité comprise entre 1 et 2. Le graphe correspondant traduit l'« essence » du graphe complet, le stylise de façon robuste : les voisins réciproques marquent l'extremum d'un gradient de densité croissante, le long duquel les points se rapprochent de plus en plus, ou des « grumeaux » accidentels intéressants en soi à repérer, au même titre que les points isolés. Cette notion de K-réciprocité étend celle de voisins réciproques (1-réciprocité) déjà utilisée en classification automatique (Benzécri, 1982).

La figure 1 montre les voisinages 3-réciproques des 67 œuvres de Corneille et Molière dans le plan des facteurs 2 et 3 d'analyse factorielle sphérique du corpus présenté (où les positions relatives des œuvres sont les mêmes, à quelque % près, que celles obtenues par AFC). Les deux croix dans la partie négative de F1 sont *Le Menteur* et *La suite du Menteur* , comédies tardives de Corneille ; le rond dans la partie positive de F1 est *Dom Garcie de Navarre* (seule tentative de comédie héroïque de Molière).

4. Caractériser chaque noyau de textes

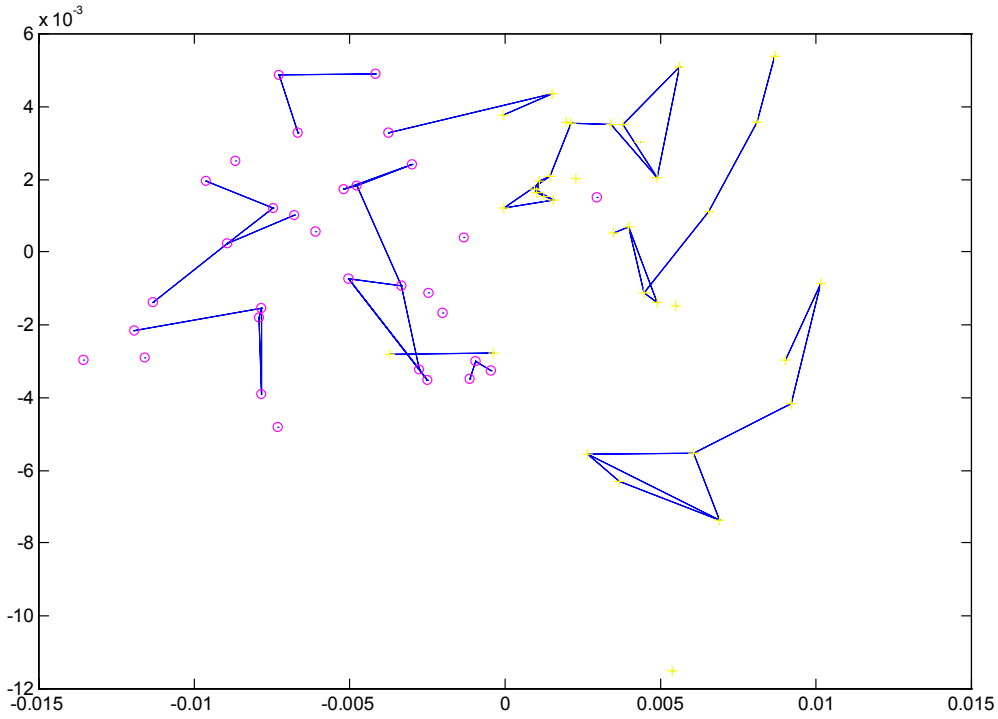
Le choix de l'espace distributionnel est également pertinent au niveau local de chaque noyau :

4.1. *Centralités des textes, des mots, et de variables explicatives extérieures*

Le premier vecteur propre de la matrice de similarité entre textes du même noyau (pondérés par leur nombre total de mots) constitue un indicateur de centralité de chaque texte au sein de son noyau (cf. Brandes, 2003) pour une revue sur l'*eigenvector centrality*) ; la pondération choisie permet d'en déduire l'indicateur de centralité de chaque mot au sein du graphe

« dual » des similarités locales entre mots de ce noyau (Domengès et Volle, 1979). La projection de l'ensemble des textes et des mots du corpus sur l'axe du noyau considéré illustre la « zone d'influence » de ce noyau : certains textes, isolés ou appartenant à d'autres noyaux, peuvent se projeter « haut », s'ils comportent plus d'un thème abordé. Nous verrons dans la section suivante l'illustration de ces principes pour caractériser des sous-ensembles homogènes de pièces dégagés par notre méthode.

Cette même procédure permet également de situer l'importance explicative de variables extérieures à l'analyse comme l'auteur du texte, la source, l'année de parution, etc. en les projetant à titre illustratif sur l'axe du noyau.



Graphe des voisins 3-réciproques sur fond (plan F2 × F3) d'Analyse Factorielle Sphérique (Corneille : + ; Molière : o).

4.2. *Visualisation des graphes locaux*

Les mêmes auteurs ont montré que les projections des textes sur les vecteurs propres suivants constituent une approximation des facteurs non triviaux extraits par AFC de la matrice des fréquences brutes. Or des travaux récents (Brandes, 2003), qui rejoignent une lignée plus ancienne (Lebart, 1984), montrent que les avant-derniers vecteurs propres de la matrice « laplacienne » issue de la matrice d'adjacence, autrement dit les premiers facteurs non triviaux d'AFC sur cette matrice d'adjacence, permettent une visualisation graphique optimale dans deux (ou trois) dimensions, optimale au sens d'un *spring-embedder algorithm* particulier : celui d'un système d'anneaux (nœuds) reliés par des ressorts (arcs) de longueur nulle au repos dans un espace à 2 (ou 3) dimensions. On peut également en déduire la représentation dans 2 ou 3 dimensions du graphe « dual » des mots (cf. §3).

5. Résultats et perspectives

L'ACL de notre corpus extrait 23 pièces isolées (en majorité des comédies en prose de Molière) et 9 noyaux de classes :

– 3 noyaux : comédies en prose de Molière
– 1 noyau : comédies en vers de Molière
– 1 noyau : comédies et tragi-comédies de Corneille (en vers)
– 3 noyaux : tragédies de Corneille (en vers)
– 1 noyau : les 2 versions de *Psyché* (co-écrite par Corneille et Molière) présentes dans le corpus.

Aucun de ces noyaux ne mélange les œuvres des deux auteurs. Les deux seules proximités notables indiquées par les projections sont celle du *Menteur* et de la *Suite du Menteur* avec les grandes comédies en vers de Molière, et de *Psyché* avec le noyau *Amants magnifiques* et *Princesse d'Elide*.

Chaque noyau est caractérisé par ses textes et mots les plus centraux (cf. ci-dessous le noyau le plus dense : les grandes comédies en prose de Molière).

Noyau N°1 :

Densité :	*Centralité :*	*N° int. :*	*Titre (v : vers ; C : comédie) :*
2.00	(.947)	64 =	Les_Fourberies_de_Scapin___C
2.00	(.957)	51 =	Dom_Juan___C
1.50	(.945)	59 =	George_Dandin__C
1.40	(.965)	60 =	L'Avare___C+
1.00	(.943)	63 =	Le_Bourgeois_gentilhomme__C
Individus ambivalents :			
1.40	(.950)	45 =	L'Ecole des femmes_v_C
1.30	(.953)	50 =	Le_Tartuffe_v_C
1.00	(.942)	67 =	Le_Malade imaginaire___C
Individus extérieurs les plus proches :			
1.30	(.935)	39 =	Dépit_amoureux_v_C
1.33	(.932)	38 =	L'Etourdi_v_C
2.00	(.932)	53 =	Le_Misanthrope_v_C

Mots :

Centralité :	*Freq.locale :*	*Freq.relative (‰) :*	*Libelle/code grammatical:*
0.9997	6077	52.5	de/8_____
0.9993	2731	23.6	que/8_____
0.9992	2453	21.1	ne/6_____
0.9992	5895	50.9	le/7_____
0.9991	2317	20.0	il/5_____
0.9991	1955	16.8	un/7_____
0.9990	95	0.82	foi/2_____
0.9990	2407	20.8	à/8_____
0.9990	693	5.98	voir/1_____
0.9990	1403	12.1	faire/1_____
0.9989	3631	31.3	être/1_____
0.9987	1734	14.9	que/5_____

Les mots les plus centraux du noyau sont ceux qui présentent les fréquences relatives les plus constantes dans les trois œuvres. Pour interpréter les thèmes en matière de style et de contenu, ceci peut fournir un angle de vue complémentaire de ceux donnés par d'autres indicateurs, comme l'écart réduit à la fréquence moyenne dans le corpus – ainsi *peu* et *fort* sont les mots les plus centraux de trois grandes comédies en vers de Molière, alors qu'*avoir*, *à*, *faire* le sont dans trois tragédies de Corneille. Une grande finesse d'analyse est ainsi possible à partir de notre extraction d'œuvres isolées et de noyaux denses : on peut les placer sur un graphe d'ensemble, les caractériser par plusieurs indices et listes ordonnées de mots, les représenter

localement sur un axe de centralité, aussi bien que sur un plan factoriel local si ce noyau est complexe. Ces possibilités nous paraissent intéressantes si elles sont appropriées par des experts du domaine qui peuvent les mettre en relation avec leur connaissance fine des textes et des éléments d'analyse extérieurs à ceux-ci ; moins intéressantes en tant que « tranchoir » pour un usage décisionnel d'attribution d'auteur…

Au delà d'une application « grandeur réelle » à 1400 enregistrements bibliographiques médicaux (Lelu et François, 2003), qui a donné lieu à un graphe de 232 noyaux consultable sur une interface Web graphique, notre horizon est celui de l'analyse dynamique de corpus : cette méthode produit une séquence réticulée de classes quand on introduit de nouveaux textes dans l'analyse ; chaque texte, ou paquet de textes, nouveau entraîne des créations /suppressions/ fusions /éclatements localisés de noyaux existants. La difficulté est alors de concevoir et réaliser une interface conviviale pour parcourir dans ses dimensions tant spatiales que temporelle ce réticulogramme de construction des classes.

Bibliographie

Bacelar-Nicolau H. (1987). On the distribution equivalence in cluster analysis. In Devijver PA and Kittler J. (Eds), *Patt. Rec. Theory and Appl.*, NATO ASI series F, vol. (30), Springer-Verlag : 73-79

Benzécri J.P. (1982). Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, vol. (2) : 208-218.

Brandes U., Gaertler M. et Wagner D. (2003). Experiments on Graph Clustering Algorithms. In Proceedings of the *11th Europ. Symp. Algorithms* (ESA '03). Springer LNCS.

Colombo T., Quentin Y. et Guénoche A. (2003). Recherche de zones denses dans un graphe : application aux gènes orthologues. In San Juan E. (Ed.), *JIM'2003*, sept. 3-6, 2003, Metz.

Domengès D. et Volle M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, vol. (35) : pp. 3-84.

Enright AJ et Ouzounis C. (2001). BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, vol. (17/9) : 853-4.

Labbé D. et Labbé C. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*, vol (8/3) : 213-231.

Lebart L. (1984). Correspondence Analysis of graph structures. *Bull. Techn. du CESIA*, vol. (2/1-2) : 5-19.

Lelu A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In Diday E., Lechevallier Y. *et al.* (Eds), *New Approaches in Classification and Data Analysis*. Springer-Verlag : 241-248.

Lelu A. et Ferhan S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In Proceedings of *IFCS'98* (International Federation of Classification Societies) : 206-209.

Lelu A. (2003). Évaluation de trois mesures de similarité utilisées en sciences de l'information. Information Sciences for Decision Making, vol. (6).

Lelu A. et François C. (2003). Un algorithme de détection de maximas de densité basé sur la distance distributionnelle : application à la classification optimale fine d'un corpus documentaire. In Dodge Y. et Melfi G. (Eds), *10èmes Rencontres SFC*, Pr. Acad. Neufchâtel.

Trémolières R.C. (1979). The percolation method for an efficient grouping of data. *Pattern Recognition*, vol. (11/4).

Watts D. et Strogatz S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, vol. (393) : 440-442.