Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel : test de randomisation TourneBool sur le corpus Reuters

Alain Lelu* **, Martine Cadot** ***

*Université de Franche-Comté
30, rue Mégevand
25030 Besancon Cedex
Alain.Lelu@univ-fcomte.fr
**LORIA, Bât. C
Campus scientifique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Alain.Lelu@loria.fr
** Université Henri Poincaré – Nancy1
Département informatique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Martine.Cadot@loria.fr
http://www.loria.fr/~cadot/

Résumé. La définition du voisinage est un élément central en fouille de données, et de nombreuses définitions ont été avancées. Nous en proposons ici une version statistique issue de notre test de randomisation TourneBool, qui permet, à partir d'un tableau de relations binaires objets décrits / descripteurs, d'établir quelles relations entre descripteurs sont dues au hasard, et lesquelles ne le sont pas, sans faire d'hypothèse sur les lois de répartitions sous-jacentes, c'est-à-dire en tenant compte de lois de tous types sans avoir besoin de les spécifier. Ce test est basé sur la génération et l'exploitation d'un ensemble de matrices randomisées ayant les mêmes sommes marginales en lignes et colonnes que la matrice d'origine. Après une première application encourageante à un corpus textuel réduit, nous avons opéré le passage à l'échelle adéquat pour traiter des corpus textuels de taille réelle, comme celui des dépêches Reuters. Nous caractérisons le graphe des mots de ce corpus au moyen d'indicateurs classiques comme le coefficient de clustering, la distribution des degrés et de la taille des « communautés », etc. Une autre caractéristique de TourneBool est qu'il permet aussi de dégager les "anti-liens" entre mots, à savoir les mots qui « s'évitent » plus qu'attendu du fait du hasard. Le graphe des liens et celui des anti-liens seront caractérisés de la même façon.

1 Introduction et problématique

La définition du voisinage est un élément central en fouille de données : elle est à la base de méthodes supervisées, comme l'apprentissage à partir des K plus proches voisins, ou non-supervisées, comme celles basées sur les graphes. De nombreuses définitions de la notion de voisinage ont été proposées :

- voisinage des K plus proches voisins, des voisins réciproques (Benzécri 1982), des voisins K-réciproques (Lelu 2004),
- voisinage de la lunule (ou : des voisins relatifs) (Toussaint 1980) : deux points sont voisins si la lunule qu'ils définissent (intersection des deux sphères dont ils sont les centres, de rayon la distance qui les sépare) est vide,
- graphe de Gabriel (deux points sont voisins si la sphère de diamètre défini par ces deux points est vide) (Gabriel & Sokal 1969),
- triangulation de Delaunay (Goodman & O'Rourke 2004) : telle qu'aucun point n'est dans la sphère circonscrite à tout simplexe (équivalent n-dimensionnel du triangle) dans l'ensemble des points.

Basées sur des notions géométriques et topologiques (cf. Scuturici et al., 2005), elles ont pour avantage leur adaptativité à des effets de distances paradoxaux bien connus dans les espaces fortement multidimensionnels, mais leur caractère systématique les expose à des effets de bords indésirables : par exemple le voisinage des K plus proches voisins peut relier un élément isolé à d'autres sans rapports aucuns.

Nous explorons ici une autre voie : définir les liens de voisinage entre éléments (lignes ou colonnes) d'un tableau de données à partir de considérations statistiques. Nous nous limiterons ici à des tableaux binaires de présence / absence d'attributs pour un ensemble d'objets décrits. Ainsi deux attributs seront considérés comme significativement liés si leur présence simultanée dans les objets décrits est plus grande qu'attendue statistiquement. Nous discuterons plus loin cette notion, centrale ici, d'attente statistique.

A noter qu'une des originalités de notre démarche est qu'elle permet également de définir et d'opérationnaliser la notion d'*anti-lien* : on détecte de la même façon les couples d'éléments dont la co-présence est significativement moindre qu'attendu.

2 Le test de randomisation TourneBool

Une façon classique de définir la significativité du lien entre deux attributs X et Y est de considérer les seules quatre cases du tableau croisé de leurs présences / absences, habituellement notées a, b, c, d (cf. tableau 1), et de les comparer aux valeurs qu'elles auraient théoriquement en cas d'absence de lien. La comparaison se fait au moyen d'un test statistique.

	Y	non Y
X	**a**	**b**
non X	**c**	**d**

TAB. 1 – *Les quatre valeurs de base des indices locaux d'association entre X et Y.*

Le test du khi-deux d'indépendance est le plus connu, mais il n'est pas utilisé systématiquement car il est mal adapté aux tableaux de données particuliers comme par exemple les données textuelles. Ce test ne peut en effet pas s'appliquer en cas d'effectifs déséquilibrés (il faut qu'aucune des 4 cases n'aie de valeur théorique trop faible, cf Morineau, 1996), ou en cas de données trop nombreuses (il a tendance à devenir alors toujours significatif) qui sont la caractéristique des données textuelles. Parmi les autre tests utilisés pour établir la significativité d'un lien entre deux variables à partir des seuls effectifs de la table 1, citons deux tests plus adaptés à des données déséquilibrées : 1) la « vraisemblance du lien » de Lerman, 2003 qui utilise une modélisation probabiliste du déséquilibre, 2) le test exact de Fisher, 1936, bien antérieur au test du khi-deux, qui procède par comptage de toutes les configurations possibles du tableau obtenues en changeant les valeurs des propriétés pour les sujets sous la contrainte de conservation des marges du tableau, donc de leur éventuel déséquilibre.

A notre connaissance, tous les tests basés sur les seules quatre cases du tableau 1 ont la même particularité que le test du khi-deux d'indépendance : les liens recherchés par ces tests expriment des associations locales entre paires de variables, c'est-à-dire sans prendre en compte les autres variables. Plus généralement, ce type de test, et son utilisation répétée pour toutes les paires d'attributs des données, n'a pas été conçu pour la découverte de liaisons significatives dans les données typiques de la fouille de données, qui présentent des variables en grand nombre d'une part, avec des distributions très hétérogènes d'autre part ; leur utilisation sur de telles données pose des problèmes de plus en plus régulièrement pointés (problème des comparaisons multiples cf. Jensen, 1998, des échantillons exhaustifs cf. Press, 2004 , de l'indépendance relative cf. Bavaud, 1998, …).

Une autre voie, rendue possible par les capacités croissantes des ordinateurs modernes et par leur mise en réseau, est celle de la comparaison à de l'aléatoire simulé, et non théorique, que ce soit par l'introduction de bruit sur le tableau originel (bootstrap, Jacknife) ou par génération de tableaux aléatoires sans rapport avec celui-ci, mais sous les mêmes contraintes structurelles.

C'est cette dernière voie que nous avons choisie, en concevant la méthode TourneBool de génération de matrices aléatoires de mêmes sommes marginales en lignes et colonnes que la matrice d'origine (cf Cadot 2005, Cadot 2006), et le test qui en découle. Cette méthode, comme toutes les méthodes des tests de randomisation (cf Manly 1997) dont elle fait partie, est inspirée du test de Fisher, mais elle s'applique aux variables prises dans leur ensemble et non par paires. Elle procède par une suite d'échanges rectangulaires élémentaires qui ne modifient pas les sommes en lignes et en colonnes. Ces échanges permettent de casser tous les liens entre variables qui peuvent l'être sans modifier la structure des données. Par exemple, dans le cas d'un tableau d'incidence entre les textes et les mots, si certains mots sont dans presque tous les textes, ils auront presque tous les textes en commun dans la matrice d'origine comme dans les matrices simulées. Mais si dans la matrice d'origine ces mots très fréquents sont presque tous absents de certains textes contenant suffisamment de mots, dans les matrices simulées ces textes contiendront souvent ces mots. Ainsi la simulation permet de mettre au jour la partie structurelle d'une liaison et d'éliminer la partie non structurelle, celle qui nous intéresse, qui est alors extraite par comparaison avec la matrice d'origine. La partie structurelle est spécifique au type des données et liée aux distributions de probabilités des marges de la matrice. Elle correspond à notre « attente statistique » d'absence de liens entre les variables conditionnellement au corpus. L'éliminer par des simulations rend ainsi notre méthode capable de traiter tout type de données, 1) en

tenant compte des lois marginales *et* 2) sans qu'il soit besoin de spécifier celles-ci, qui peuvent être de nature quelconque.

Le principe d'examen des liens est le suivant :

Soient deux mots m_i et m_j, présents simultanément dans p_0 (p_0>=0) textes du corpus d'origine, et dans p_k (p_k>=0) textes de la k-ième simulation du corpus. Le lien entre m_i et m_j peut correspondre à un des 3 cas distincts suivants:

- Si p_0 est supérieur à la quasi-totalité des p_k, alors lien (m_i,m_j) > 0 (attirance).
- Si p_0 est inférieur à la quasi-totalité des p_k, alors lien (m_i,m_j) < 0 (répulsion).
- Sinon lien (m_i,m_j) = 0 (indépendance sachant le corpus).

Dans le premier cas, on parlera de lien significativement positif, ou tout simplement de lien, dans le deuxième de lien significativement négatif, ou plus simplement d'anti-lien, et dans le dernier cas de lien nul, ou non significatif, ou d'absence de lien.

Par exemple, dans la figure 1 est représentée la suite ordonnée des valeurs de p_k d'un lien entre deux mots pour 100 simulations (k varie de 1 à 100). Si on décide de prendre un seuil de risque alpha[1] de 10%, on choisit les deux bornes indiquées par un triangle sur le graphique : la valeur 2 correspondant à k=6, et la valeur 22 à k=95. Si on note par p la valeur théorique du nombre de co-occurrences dans le corpus en cas d'absence de lien entre les mots m_i et m_j, son intervalle bilatéral à 90% de confiance est estimé par l'intervalle [2 ; 22], qui contient les 90 valeurs les moins extrêmes de la suite des p_k.

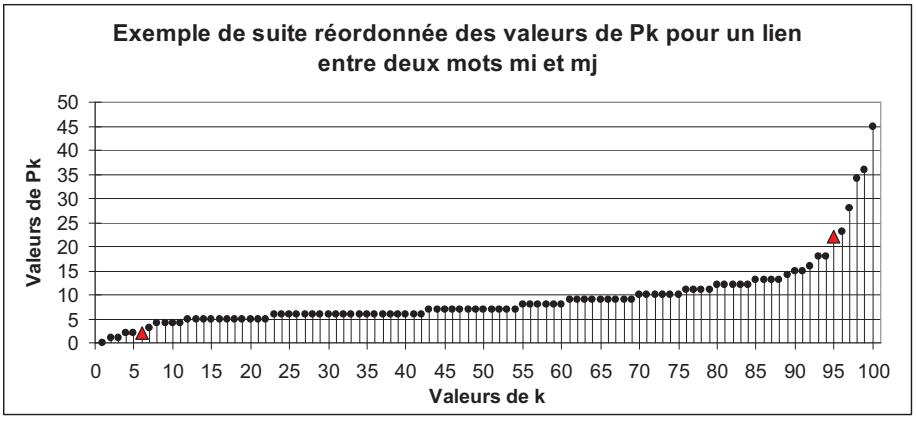

FIG. 1 – *Le nombre p_k de co-occurrences d'un lien pour la k-ième simulation (les 100 simulations sont rangées par valeurs de p_k croissantes).*

Selon que la valeur de p_0 (nombre de co-occurrences de ces deux mots dans le corpus d'origine) est en dehors ou non de cet intervalle, on décidera que le lien considéré est significatif ou non. Notamment, si la valeur de p_0 est 0 ou 1, on dira que le lien entre les 2 mots est significativement négatif, c'est-à-dire que la rareté de l'apparition simultanée ce ces

[1] Risque alpha : risque de se tromper en jugeant significative une valeur de p qui est due au hasard. Il correspond à la notion de « faux positif » des tests cliniques. La notion de « faux négatif » correspond quant à elle au risque bêta : risque de se tromper en jugeant due au hasard une valeur de p qui est significative.

deux mots dans les mêmes textes du corpus correspond à une répulsion entre eux qui doit avoir une signification ; si la valeur de p_0 est 23, 24, ou plus, on dira que le lien entre ces deux mots est significativement positif, c'est-à-dire que la valeur élevée du nombre de leurs co-occurrences correspond à une attraction entre ces mots qui est certainement chargée de sens. Par contre si la valeur de p_0 est 2, 3, …, 22, on l'attribue au hasard et on conclut que les deux mots ne sont pas liés par le sens, mais par la structure du corpus. L'utilisation de ce test permet de mettre au jour automatiquement les associations de deux mots pouvant avoir un sens.

Dans l'encadré 1 figure l'algorithme TourneBool qui permet de trouver les liens significatifs. Dans cette version de l'algorithme, les données sont sous forme d'un tableau rectangulaire booléen M, les colonnes correspondant aux variables dont on désire extraire les liens significatifs, par exemple les mots d'un corpus de textes, les lignes figurant alors les textes. Un 1 présent à l'intersection d'une colonne et d'une ligne signifie que le mot correspondant figure au moins une fois dans le texte, et un 0 qu'il n'y figure pas.

TourneBool

• 1. générer *q* versions randomisées de *M*
• 2. pour chaque paire de colonnes (*i, j*) de *M*
Calculer la valeur *p0* de *p*, nombre de co-occurrences de *(i,j),* pour *M*
Comparer *p0* aux bornes de l'intervalle de confiance de *p* obtenu d'après les *q* versions randomisées de *M*
Choisir parmi ces 3 alternatives celle qui convient :
- si *p0* est entre ces bornes, le lien entre i et *j* est déclaré non significatif et est éliminé.
- si *p0* est inférieur à la borne inférieure, le lien est déclaré significativement négatif et est stocké.
- si *p0* est supérieur à la borne supérieure, le lien est déclaré significativement positif et est stocké.

Génération de *q* versions randomisées de *M*

• 0. Choisir un nombre *r* d'échanges rectangulaires à exécuter
• 1. Répéter *q* fois :
1.1. Faire une copie *Mc* de *M*
1.2. Répéter *r* fois:
- Choisir au hasard une paire de lignes et une paire de colonnes de *Mc*.
- Examiner si les coins du rectangle ainsi formé contiennent une alternance de 0 et de 1
si oui, les remplacer par leur complément à 1
si non, ne rien faire
1.3. stocker *Mc* :

Construction de l'intervalle de confiance, au risque alpha, du nombre *p*

• 1. Pour chaque version *Mk* randomisée de *M*, calculer le nombre *pk* de co-occurrences des deux colonnes *i, j* (produit scalaire des deux colonnes). Stocker tous les nombres *pk* dans une liste.
• 2. Trier la liste des *pk* dans l'ordre croissant. La borne inférieure de l'intervalle de confiance est l'élément de rang *q.alpha/2* et sa borne supérieure l'élément de rang *q(1-alpha)/2*.

Encadré 1 *: Algorithme TourneBool*

L'utilisation de cet algorithme nécessite de fixer la valeur de 3 paramètres : le nombre d'échanges rectangulaires, le nombre de matrices randomisées, et le risque alpha. Les deux derniers paramètres se fixent selon les compromis habituels. Compromis qualité/rapidité de l'informatique : plus on a de matrices simulées, plus la qualité de l'estimation est grande, mais plus les calculs demandent de temps. Notre habitude est de faire 100 ou 200 simulations. Compromis risque alpha/risque bêta de la statistique : plus le risque alpha est petit, moins on risque d'extraire des liens dus au hasard, mais dans ce cas le risque beta

augmente, et on risque d'extraire moins de liens porteurs de sens. Pour alpha, nous avons donné en exemple la valeur 10%, car elle est facilite la description des exemples, mais nous avons utilisé des valeurs de 5% ou 1%.

Le nombre r d'échanges rectangulaires est contrôlé par deux mesures de distance entre matrices : 1) entre la matrice générée et la matrice d'origine, 2) entre la matrice générée et la matrice générée précédemment. On augmente le nombre d'échanges tant que les deux suites de distances sont croissantes. La valeur de ce paramètre est jugée optimale dès qu'elles se stabilisent. La distance entre deux matrices est ici la distance de Hamming (nombre de cases de valeurs différentes).

Après une première application encourageante à un corpus textuel réduit (Cadot et al., 2007), nous avons opéré le passage à l'échelle adéquat pour traiter des corpus textuels de taille réelle. La génération des matrices randomisées s'effectue en temps O(n m v) et espace O(v), où n et m sont les nombres de lignes et de colonnes de la matrice à tester, v son nombre de valeurs « 1 ». Composée de processus indépendants, cette phase se parallélise de façon naturelle. Pour éviter les problèmes de mémoire vive dans la phase suivante d'exploitation des N matrices (stocker 100 ou 200 fois v nombres…), on la fragmente en processus parallèles indépendants, chacun en charge d'un fragment du tableau (symétrique) des co-occurrences. L'application au corpus Reuters décrite plus bas (23 000 dépêches, 28 000 mots) a pris de l'ordre de deux journées et demi de calcul pour l'ensemble des phases, sous forme de 3 processus parallèles lancés sur un PC quadricoeur standard.

3 Caractériser les graphes

L'étude de ce que l'on appelle aujourd'hui les grands graphes de terrain (Latapy, 2007) est en plein essor. Outre notre domaine de l'analyse de corpus textuels, ces méthodes concernent des domaines aussi différents que l'analyse des réseaux sociaux, la circulation des données sur Internet, ou les réseaux d'interaction entre protéines. La validation statistique d'un lien entre deux variables étant un processus binaire, en tout ou rien, la représentation d'un grand ensemble de tels liens se prête bien au formalisme des graphes. Le nombre d'unités textuelles et de mots en jeu – plusieurs dizaines de milliers au bas mot – appelle le vocable *grand graphe de terrain*. Nous utiliserons ici quelques indicateurs habituels permettant de caractériser globalement ou localement un tel graphe.

- nombre de nœuds et de liens ;
- densité du graphe : nombre de liens / nombre de liens possibles ;
- degré moyen (moyenne du nombre de degrés par nœud) ;
- distribution des degrés ;
- corrélation de Pearson entre degrés des nœuds voisins ;
- distribution des coefficients de clustering des nœuds (ou indicateurs de « cliquicité » : entre 1 si tous les voisins d'un nœud sont liés entre eux, et 0 s'ils ne le sont pas ; plus précisément : $C_i = n_i / (k_i (k_i-1) / 2)$, rapport du nombre n_i de liens entre voisins du noeud i, de degré k_i, au nombre de liens possibles)
- coefficient de clustering moyen (moyenne des coefficients de clustering de tous les nœuds) ;

- clusters (ou « communautés ») de mots proches, et distribution de la taille de ces clusters. Nous avons utilisé pour les constituer le programme d'accès libre WalkTrap2 (Pons, Latapy, 2006).

4 Application au corpus Reuters

L'agence de presse Reuters et D. Lewis (2004) ont mis à la disposition des chercheurs en fouille de données plusieurs corpus de dépêches, sous diverses formes, afin de permettre des comparaisons entre travaux et méthodes sur une base commune et publique. Nous avons utilisé le corpus[3] de 23 149 dépêches lemmatisées par l'outil Bril, afin de nous concentrer sur l'exploitation d'un vocabulaire publié, sans bruiter nos résultats par un processus d'indexation spécifique, fût-il de meilleure qualité (les lemmes consistent ici en de simples troncatures de mots, limitant le vocabulaire à quelques dizaines de milliers de termes, mais créant aussi beaucoup d'ambiguïtés sémantiques pour les mots anglais, et des formes peu compréhensibles ; pas question non plus de trouver des formes composées, de signification plus univoque…).

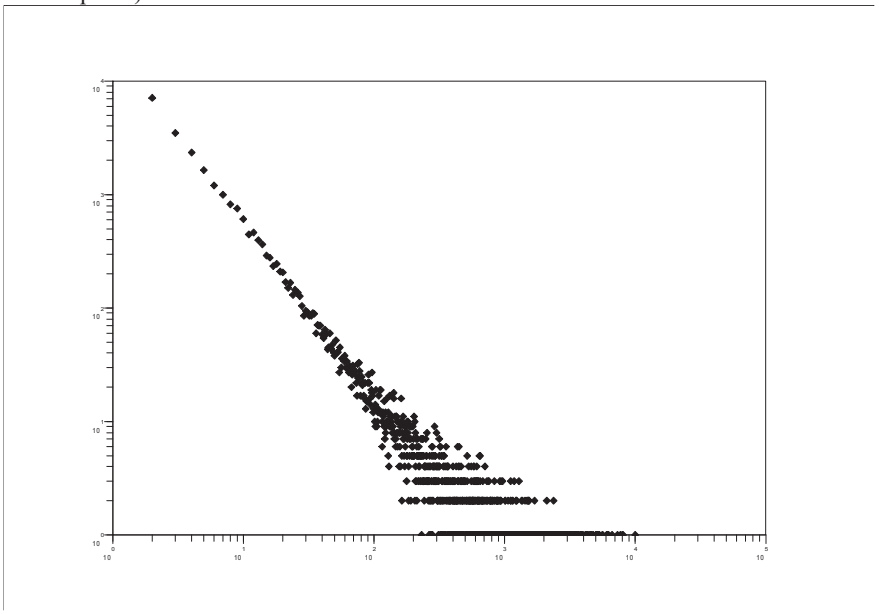

FIG. 2 – *Corpus Reuters RCV1 : en abscisse, les occurrences de chaque mot dans le corpus ; en ordonnée, le nombre de répétitions de ces occurrences (coordonnées log-log). Ex. : il y a 7100 mots d'occurrences = 2.*

[2] WalkTrap <www-rp.lip6.fr/~latapy/PP/walktrap.html>.

[3] Lewis, D. RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (12-Apr-2004 Version).
http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.

Après suppression des hapax, le vocabulaire s'établit à 28 450 lemmes, de *a0* à *zywnociowej*, à raison d'environ 75 lemmes uniques par dépêche. La distribution des occurrences des mots présente l'allure Zipfienne habituelle (en coordonnées log-log : décroissance linéaire), en loi de puissance d'exposant -1,5 environ (cf. fig. 2). Celle du nombre de mots uniques pour les dépêches a un mode prononcé, autour de 26 mots, et une forte dissymétrie (cf. fig. 3).

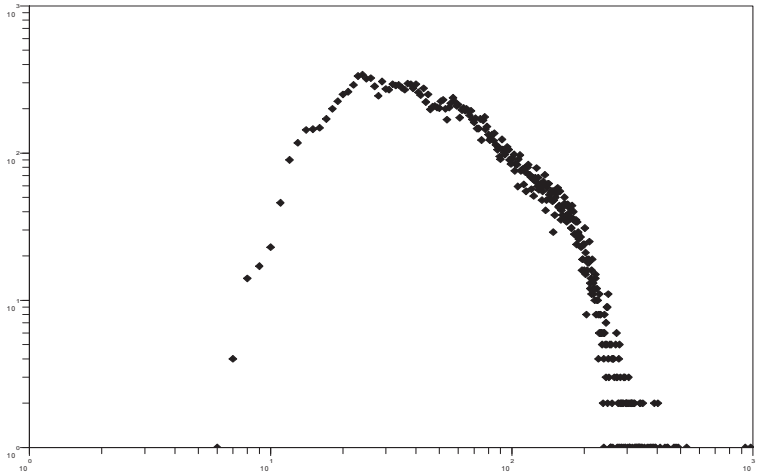

FIG. 3 – *Corpus Reuters RCV1 : en abscisse, les occurrences de mots dans chaque dépêche ; en ordonnée, le nombre de répétitions de ces occurrences (coordonnées log-log). Ex. : il y a 324 dépêches de 26 mots.*

Nous avons extrait par notre méthode TourneBool les matrices d'incidence du graphe des liens (resp. anti-liens) au seuil de confiance de 99%. Celui des liens en comporte 2,8 millions, sa densité est de 0.0071 ; par comparaison, la matrice des co-occurrences brute des mots a une densité bien supérieure (0.0406). La matrice des anti-liens ne comporte que 490 000 « liens » (densité ~ 0.0012). Le degré moyen de ces matrices est de 201,4 (resp 34,5), contre 1156 pour la matrice des co-occurrences brutes. La corrélation entre degrés de part et d'autre des liens a la valeur négligeable de +0.017 pour la matrice des liens (réseau « non-assortatif »), contre un assez fort -0.39 pour les co-occurrences brutes (réseau « anti-assortatif » : les nœuds avec un nombre très différent de liens ont tendance à s'agréger ensemble) ; ce qui montre à quel point l'opération de validation statistique des liens a changé la nature profonde de ce réseau de référence. Le graphe des anti-liens est lui aussi anti-assortatif (-0,37). Dans la distribution des degrés des liens, on retrouve la même allure que dans la distribution des nombres de mots par dépêche (cf. fig. 4), alors que celle des anti-liens présente une loi de puissance bien nette (cf. fig. 5).

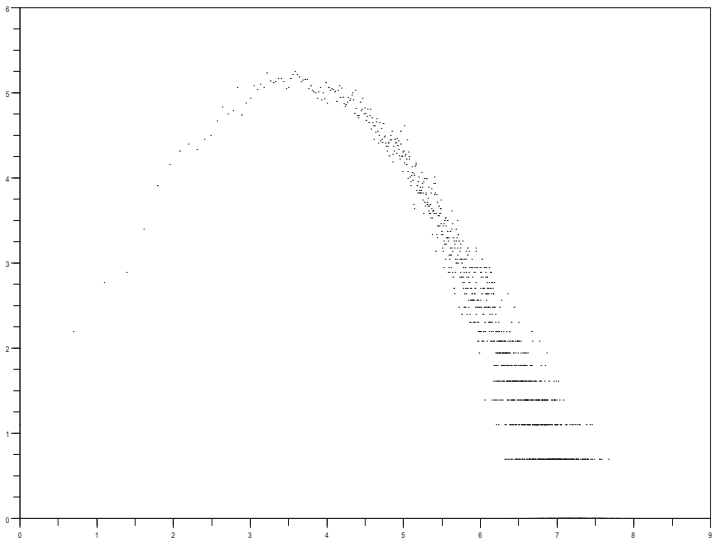

FIG. 4 – *Pour le graphe des liaisons valides entre mots du corpus Reuters RCV1, en abscisse, les degrés des mots ; en ordonnée, le nombre de répétitions de ces degrés (coordonnées log-log).*

Les cœfficients de clustering moyens sont plutôt élevés, comme dans la plupart des graphes de terrain : 0.305 pour les liens, 0.294 pour les anti-liens, contre un très fort 0.808 pour le graphe des co-occurrences.

L'application du logiciel de clustering de graphe WalkTrap fournit un arbre de classification ascendante hiérarchique que nous avons coupé pour la valeur maximale de l'indicateur de qualité de la partition (« modularité ») : pour les liens on obtient 21 clusters, dont 13 d'effectifs supérieurs à 55, et deux d'effectifs supérieurs à 7000 – répartition en loi de puissance typique. Leur contenu est surprenant : l'un rassemble des prénoms, l'autre des villes du Royaume Uni… mais dans tous les cas ils présentent un mélange entre des noms de lieux, de personnes ou d'entreprises et des éléments de contenu – ici l'aéronautique, là la chimie ou l'informatique…

Pour les anti-liens on obtient 21 000 éléments isolés, 143 clusters de 2 à 10 éléments, 31 entre 11 et 50, 2 d'environ 500 éléments, et 2 d'environ 2300 ; cette dispersion se traduit par une très mauvaise valeur de modularité maximale (0,011). L'interprétation de ces clusters semble encore plus délicate, bien qu'on puisse dire qu'on y retrouve les lemmes de l'anglais courant, mélangés à d'autres éléments.

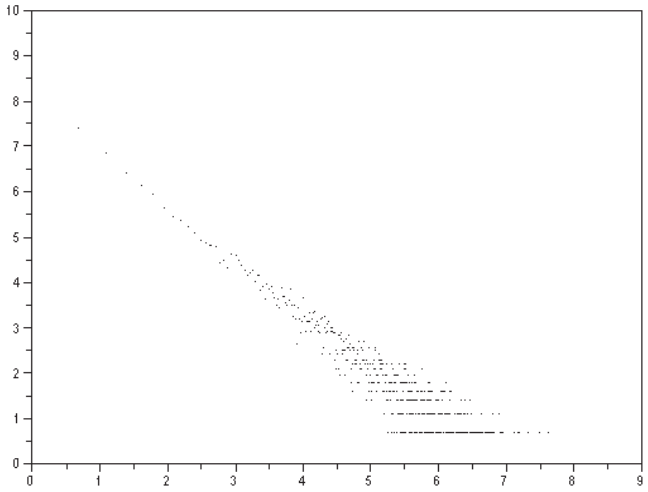

FIG. 5 – *Pour le graphe des mots anti-liés du corpus Reuters RCV1, en abscisse, les degrés des mots ; en ordonnée, le nombre de répétitions de ces degrés (coordonnées log-log).*

Le tableau 2 récapitule les résultats obtenus.

	Graphe des :		
	co-occurrences brutes	liens valides	anti-liens valides
Nombre de noeuds	28 450	28 450	28 450
Nombre de liens	16,0 M	2,8 M	0,49 M
Densité	0,0406	0,0071	0,0012
Degré moyen	1156	201,4	34,5
Corrélation de degrés entre 1-voisins	-0,39	+0,017	-0,37
Coefficient de cliquicité	0,808	0,305	0,294
Nombre de clusters	n.a.	21	~21 200
dont effectif=1	n.a.	3	~20 000
.. pour modularité max =	n.a.	0,276	0,011

TAB. 2 – *Comparaison des 3 graphes des mots, dont les matrices d'adjacence sont 1) la matrice des co-occurrences brutes, 2) celle des liens valides, celle des anti-liens valides.*

5 Conclusion, perspectives

Le présent article n'avait pour but que de constituer une première approche d'une voie d'avenir en fouille de données, autorisée par les performances croissantes du matériel informatique : appliquer à grande échelle les tests de randomisation, qui permettent des validations statistiques tenant compte des lois de distributions sous-jacentes sans qu'il soit nécessaire de les spécifier. Les perspectives sont vastes, et nous nous contenterons d'en énumérer quelques unes à court et moyen terme :

- caractériser les graphes de voisinage obtenus à partir des méthodes géométriques et topologiques, et les comparer à notre approche ;
- explorer l'influence du seuil de confiance statistique (95%, 99%, …) sur la taille et les autres caractéristiques des graphes obtenus par notre méthode ;
- paralléliser plus avant notre chaîne de traitements.

Il reste aussi à travailler avec les experts des domaines d'application concernés pour approfondir l'interprétation des éléments mis en avant par l'analyse, comme les clusters de mots anti-liés.

Références

Bavaud F. (1998). *Modèles et données : une introduction à la Statistique uni-, bi- et trivariée*. Paris ; Montréal (Qc) : L'Harmattan.

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

Cadot M., (2005). A Simulation Technique for extracting Robust Association Rules, *CSDA 2005* (Chypre).

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Thèse de doctorat en informatique. Université de Franche-Comté, 2006.

Cadot M., Lelu A. (2007) Simuler et épurer pour extraire des motifs pertinents. *Atelier QDC, EGC2007*. Namur, 22 janvier 2007

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics,* 7 : 179-188

Gabriel, K. R.; Sokal, R. R. (1969). A new statistical approach to geographic variation analysis, *Systematic Zoology* 18: 259–270

Goodman, J. et J. O'Rourke, éditeurs. (2004). *Handbook of Discrete and Computational Geometry*. CRC Press, USA.

James S. Press. (2004). The role of Bayesian and frequentist multivariate modeling in statistical Data Mining, dans "*Statistical Data Mining and Knowledge Discovery*", H. Bozdogan, Chapman & Hall/CRC, Boca Raton, US

Jensen D. (1998). Multiples comparisons in induction algorithms. *Kluwer Academic Publishers*, Boston p1-33

Latapy, M. (2007). *Grands graphes de terrain - mesure et métrologie, analyse, modélisation, algorithmique*, Habilitation à Diriger des Recherches, LIP6, Université Pierre et Marie Curie. <http://www-rp.lip6.fr/~latapy/HDR/>

Lelu A. (2004). Analyse en composantes locales et graphes de similarité entre textes. *Actes de JADT 2004*, G. Purnelle ed., Université catholique de Louvain., 10-12 mars 2004.

Lerman, I-C., Peter, P. (2003). Indice probabiliste de vraisemblance du lien entre objets quelconques. Analyse comparative entre deux approches. *Revue de Statistique Appliquée*, vol. 51, 1:5-35.

Lewis, D. D.; Yang, Y.; Rose, T.; and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 5:361-397, <http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf>.

Morineau, A., Nakache, J.-P., Krzyzanowski, C. (1996). *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris.

Pons, P, Latapy, M (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, vol.10, 2:191-218.

Scuturici M., Clech J., Scuturici V., Zighed D. (2005). Topological representation model for image databases query, *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 17, No. 1-2, 145-160.

Summary

Neighborhood is a central concept in datamining, and a bunch of definitions have been implemented, mainly rooted in geometrical or topological considerations. We propose here a statistical definition of neighborhood: our TourneBool randomization test processes an objects vs. attributes binary table in order to establish which inter-attribute relation is fortuitous, and which one is meaningful, out of any hypotheses on the underlying statistical distributions, but taking into account these empirical distributions. It ensues a robust and statistically validated graph. A previous encouraging small-scale test led us to scale up the different phases of the process, making it possible to test it on one of the public access Reuters test corpus. We then characterized the resulting word graph with a series of well-known indicators, such as clustering coefficients, degree distribution and correlation, cluster modularity and size distribution. Another graph structure stems from this process: the one conveying the negative « counter-relations » between words, i.e. words which « steer clear » one from another. We characterize in the same way the counter-relations graph.

Keywords

Neighborhood graph; randomization test; graph characterization statistics; data mining; text mining; given-marginals random matrix; statistically significant relation,