

EVALUATION DE TROIS MESURES DE SIMILARITE UTILISEES EN SCIENCES DE L'INFORMATION

Lelu Alain

INRA /unité Mathématiques, Informatique et Génomique
domaine de Viltert, 78352 Jouy en Josas

Université de Franche-Comté / LASELDI
UFR SLHS, 25030 Besançon cedex
alain.lelu@univ-fcomte.fr

Résumé : sous-jacents à la diversité des méthodes et outils des sciences de l'information et de l'analyse des données textuelles, on trouve un petit nombre d'indicateurs de proximité entre textes ou entre descripteurs de ceux-ci . Nous comparons les deux plus répandus, le cosinus de Salton et la distance du khi-deux, à celui que nous avons choisi pour nos traitements : le cosinus dans l'espace distributionnel. Après un exposé et une comparaison théorique, nous procédons à une évaluation empirique par la méthode "rappel / précision" sur le corpus réel multilingue ELDA, fourni dans le cadre de la campagne d'évaluation Amaryllis 99.

Abstract : A small number of similarity indicators are underlying the rich diversity of methods and tools in the overlapping information sciences and text analysis fields. We test the major two of them, Salton's TF-IDF cosine and khi-square distance, against the one we have chosen in our text analysis tools : the cosine measure in the "distributional" vector space. After a theoretical presentation and comparison, we apply the "recall / precision" methodology to their empirical evaluation based on a real text database, provided by the ELDA multilingual corpus (Amaryllis'99 evaluation campaign).

Mots-clés : analyse de données textuelles, sciences de l'information, recherche d'information, évaluation, mesure de similarité, cosinus de Salton, distance du khi-deux, distance distributionnelle, gain d'information de Renyi.

Keywords : textual data analysis, information science, information retrieval, evaluation, similarity indicator, Salton's cosine, khi-square distance, distributional distance, Renyi's information measure.

Evaluation de trois mesures de similarité utilisées en sciences de l'information

INTRODUCTION ET GRILLE D'ANALYSE

Confrontés aujourd'hui à des quantités croissantes de textes bruts (pages Web, dépêches d'agences, articles en ligne, ...) les méthodes et outils de la documentation automatique et des sciences de l'information commencent à chevaucher les frontières de ceux des autres analyses de corpus : analyses de discours, d'œuvres littéraires, d'interviews, études psychologiques ou psycholinguistiques, ... Jusqu'à présent, chaque méthode utilisée dans ces domaines a souvent été présentée comme un tout en soi, le plus souvent "à prendre ou à laisser". De fait chacune est caractérisée par une série de choix situés sur des registres très différents :

A - Découpage des unités textuelles : Ces unités peuvent être "logiques" (articles ou chapitres, sections, paragraphes, phrases, ...), ou "naturelles" (pages, résumés bibliographiques), ou encore être issues d'un découpage à fenêtre glissante autour de chaque mot, de chaque phrase, etc.

B - Choix de codage des textes : A chaque texte on fait correspondre un ensemble de descripteurs codés – se pose donc la question du choix de ces descripteurs. Faut-il opter pour un codage par mots, ou bien par N-grammes (Lelu et al., 1998) ? Au sein du codage par mots : faut-il opter pour une indexation manuelle, automatique, assistée ? Au sein des indexations automatiques et assistées : pour une indexation par simples chaînes de caractères ("full text"), ou par lemmes (mots normalisés) issus d'un traitement morpho-syntaxique, avec ou sans filtrage de certaines catégories grammaticales, et/ou filtrage statistique ? Tous ces choix de codage conduisent *in fine* à la représentation de chaque unité textuelle par un vecteur – logique (valeurs binaires) ou de fréquences - des descripteurs codés.

C - Choix des opérations offertes aux usagers : Au-delà des classiques requêtes booléennes, les systèmes documentaires évolués permettent à l'utilisateur, depuis les travaux pionniers de Gerald Salton et Karen Sparck-Jones dans les années 1960, de trouver les documents les plus proches d'un document désigné comme pertinent (*relevance feed-back* : rétroaction de pertinence) ou proches d'un document "idéal" défini par une série de mots, voire une requête en langage naturel (requête de similarité vectorielle) ; ou encore trouver les mots les plus proches d'un mot donné (*query expansion*), proches au sens de leur co-occurrence dans les unités textuelles. Beaucoup de logiciels d'analyse de données textuelles proposent des fonction voisines, par exemple : extraire les termes environnant un un terme donné, avec des indicateurs statistiques pour valider la significativité de ces liaisons.

Plus récemment, certains systèmes documentaires en ligne ont mis à la disposition de l'utilisateur des algorithmes de classification automatique sur les documents (Zamir et Oztioni, 1999) ou sur les mots (Bourdoncle, 1997), ou même une représentation cartographique d'ensemble des classes ainsi créées (Kohonen et al., 1995) (Lelu et al., 1997). Si les logiciels d'analyse de données textuelles comportent souvent de telles classifications (par ex. classification descendante hiérarchique dans le cas du logiciel Alceste [ALCESTE ; Reinert 1993]), ils proposent également la possibilité d'effectuer une analyse factorielle des correspondances (AFC : Benzecri et al., 1981) sur un sous-ensemble d'unités textuelles et de mots. Dans le même ordre d'idées et sur des domaines d'applications psycho-pédagogiques et documentaires, la méthode Latent Semantic Analysis (LSA ; Dumais 1994) réalise une opération de réduction de dimensions et filtrage des données par décomposition aux valeurs singulières de la matrice complète des données - les composantes issues de cette décomposition ne sont pas interprétées, mais servent à réaliser des opérations de rétroaction de pertinence ou d'expansion de requête dans l'espace des 200 à 300 premières composantes.

Mais toutes ces opérations et algorithmes portent rarement sur les vecteurs bruts issus des choix précédents : ils peuvent utiliser ces vecteurs sous leur forme normalisée en ligne ou en colonne (calculs de cosinus), ou réaliser des transformations plus complexes sur ces éléments.

D - Transformations des vecteurs-données : De façon générale, chaque opération d'analyse des données peut être définie par une primitive standard - comme un calcul de cosinus, une décomposition aux valeurs singulières, une classification à centres mobiles, ... - dans un espace de données transformé : chaque vecteur est déduit du vecteur brut par une certaine opération (normalisation, intersection avec l'hyperplan simplexe, ...) ; il est doté

d'un poids (unitaire, ou déduit des marges de la matrice des données), et ce dans une métrique particulière, en donnant au mot métrique son sens mathématique de matrice \mathbf{M} carrée généralement symétrique, définie positive, intervenant dans la définition générale du produit scalaire : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \mathbf{x}' \mathbf{M} \mathbf{y}$

Cette métrique peut être euclidienne standard (\mathbf{M} est alors une matrice identité), ou du khi-deux, etc.

Les combinaisons de ces divers registres de choix réalisés par chaque méthode comportent une grande part d'arbitraire, et peu d'auteurs se préoccupent de leur justification. Notre exposé se propose de comparer quelques transformations de données sur un plan théorique, puis nous présenterons une tentative d'évaluation pratique à partir d'une méthode reconnue dans le domaine documentaire : l'utilisation des courbes "rappel/précision" relatives à un ensemble de questions pour lesquelles des experts ont déterminé exhaustivement les réponses correctes. Dans ce but, seule est concernée l'opération d'expansion d'un document ou d'une requête vers les documents proches : à ce jour l'évaluation de la qualité d'une classification ou d'une cartographie textuelle reste un problème ouvert, sur lequel peu de propositions existent, et encore moins d'accord.

Notre travail se limitera donc à une comparaison/évaluation de 3 indicateurs de similarité ; il se rapproche de l'exposé (Lebart et Rajman 2000) qui compare de façon théorique et en détail une variante différente du cosinus de Salton à la distance du khi-deux et à la dissimilarité informationnelle de Kullback-Liebler, au sein d'un exposé très large sur l'ensemble des méthodes statistiques utilisées pour traiter la "matière textuelle" dans les multiples domaines qui la concernent (découpage d'unités textuelles, mesures de proximité, enrichies ou non par des connaissances *a priori*, fonction d'indexation, synthèses de corpus, phrases ou documents caractéristiques, classement de textes...).

1. TROIS MESURES DE SIMILARITE DANS L'ESPACE DES UNITES TEXTUELLES ET DES MOTS

1.1. Le cosinus de Salton

Notations utilisées :

x_{it} : fréquence du mot $N^\circ i$ dans le texte $N^\circ t$;

Sommes en ligne, en colonne, totale : $x_{i.} = \sum_t x_{it}$; $x_{.t} = \sum_i x_{it}$; $x_{..} = \sum_i \sum_t x_{it}$

Vecteurs (et matrices) : en minuscules (resp. majuscules) grasses. Ex. :

\mathbf{x}_i : vecteur-mot i ; \mathbf{x}_t : vecteur-texte t ; **Diag** [a] : matrice diagonale d'éléments a

{ } : ensemble des éléments d'un vecteur ou d'une matrice. Ex. :

{ $x_{i.}$ } : vecteur "sommes en ligne" ; { \mathbf{x}_i } : matrice \mathbf{X} des données

Produit scalaire des vecteurs-colonnes \mathbf{x} et \mathbf{y} dans la métrique \mathbf{D} : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}} = \mathbf{x}' \mathbf{D} \mathbf{y}$

(la norme de \mathbf{x} dans la métrique \mathbf{D} est définie par : $\|\mathbf{x}\|_{\mathbf{D}}^2 = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{D}}$)

Parmi les nombreux indices de similarité entre documents proposés et validés par G. Salton et son équipe (Salton, 1983), le plus répandu et apprécié est nommé par ses auteurs *best fully weighted system* (ou encore TF-IDF : term frequency, inverse document frequency) et défini par le produit scalaire usuel (métrique identité) entre vecteurs-documents dont les composantes sont définies ("pondérées", selon la terminologie de Salton) par le terme $x_{it} \log(N/n_i)$, puis normalisés :

$$\mathbf{x}_t : \{ x_{it} \} \rightarrow \{ x_{it} \log(N/n_i) / \sqrt{(\sum_i (x_{it} \log(N/n_i))^2)}\}$$

Où N est le nombre total de documents, et n_i le nombre de documents contenant le mot i , quelle que soit sa fréquence x_{it} dans chaque document.

Il revient à calculer un cosinus, c'est à dire un produit scalaire entre vecteurs normalisés de fréquences de mots dans un espace des documents pourvu de la métrique :

$$\mathbf{Ds} = \mathbf{Diag} [(\log(N/n_i))^2]$$

Pour les besoins de notre comparaison, on peut remarquer que $\log(N/n_i)$ est égal à $\log(x_{.t}/x_{i.})$ si les mots sont codés en présence/absence, ou en diffère peu, à des constantes additives près, pour des corpus volumineux et des documents longs : dans ce cas fréquent la métrique de l'espace des données peut donc être assimilée à :

$$\mathbf{Ds} = \mathbf{Diag} [(\log(x_{.t}/x_{i.}))^2]$$

En résumé, $\text{Cos Salton}(t_1, t_2) \cong \langle \mathbf{x}_{t_1}, \mathbf{x}_{t_2} \rangle_{\mathbf{Ds}} / \|\mathbf{x}_{t_1}\|_{\mathbf{Ds}} \times \|\mathbf{x}_{t_2}\|_{\mathbf{Ds}}$

1.2. Distance du khi-deux

Il est bien connu (Lebart et al., 1977) que l'AFC est équivalente à une analyse en composantes principales (ACP) dans l'espace des données transformé comme suit :

- coordonnées de \mathbf{x}_t : $\{x_{it}\} \rightarrow \{x_{it}/x_{.t}\}$
-

Les points sont alors sur le simplexe, hyperplan lieu des points \mathbf{z} tels que $\sum_i z_i = 1$ On compare donc des profils relatifs, cas bien adapté aux données textuelles où le nombre absolu de mots dans un texte est indifférent, et seule leur répartition relative compte.

- métrique : **Diag** $[x_{.}/x_{.i}]$
- masses des points t : $\{x_{.t}/x_{.}\}$

La propriété la plus intéressante de cet espace est qu'il est doté de la propriété d'*équivalence distributionnelle* : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, cette propriété assure la stabilité du système des distances au regard de l'éclatement ou du regroupement de catégories de descripteurs, tant que les unités textuelles s'y répartissent de façon un tant soit peu identique.

L'ordre des distances croissantes nous donne celui des similarités décroissantes par rapport à un document réel ou "idéal" posé en requête.

Résumons les similarités et différences par rapport à l'espace du cosinus de Salton :

. Dans l'espace de Salton, les points sont représentés sur une hypersphère unité : on compare donc des profils normalisés, et non des profils relatifs, comme en AFC. Toutefois, dans les deux cas la métrique est de la famille *inverse document frequency* (IDF) :

Diag $[x_{.}/x_{.i}]$ pour l'AFC, **Diag** $[(\log(N/n_i))^2]$, souvent équivalent à **Diag** $[(\log(x_{.}/x_{.i}))^2]$, pour le cosinus TF-IDF de Salton

. L'AFC fait un usage intensif de la symétrie formelle entre lignes et colonnes de la matrice des données, alors que peu d'auteurs suggèrent d'utiliser le cosinus de Salton pour l'expansion d'un mot vers les autres mots, bien que rien ne s'y oppose.

1.3. Le cosinus dans l'espace distributionnel

Une lignée ancienne de travaux (Matusita 1955) (Escofier 1978) (Domengès et Volle 1979) (Fichet et Gbegan 1985) se sont intéressés à ce que ces derniers auteurs appellent *distance distributionnelle* : si l'on transforme les données brutes comme suit :

- . Coordonnées : - des vecteurs-colonnes $\mathbf{x}_t : \{x_{it}\} \rightarrow \mathbf{y}_t : \{\sqrt{x_{it}}\}$
- des vecteurs-lignes $\mathbf{x}_i : \{x_{it}\} \rightarrow \mathbf{y}_i : \{\sqrt{x_{it}}\}$
- . Poids de ces vecteurs : unité
- . Métrique : euclidienne standard

les cosinus entre vecteurs-colonnes \mathbf{y}_t (resp. entre vecteurs-ligne \mathbf{y}_i) possèdent dans cet espace des propriétés intéressantes. Ils sont liés en effet à la notion de distance distributionnelle Dd par la relation :

$$Dd(t_1, t_2)^2 = 2(1 - \cos(t_1, t_2))$$

(resp. $Dd(i_1, i_2)^2 = 2(1 - \cos(i_1, i_2))$)

La distance distributionnelle est la distance entre les intersections de 2 vecteurs \mathbf{y}_{t1} et \mathbf{y}_{t2} avec l'hypersphère unité, c'est à dire la longueur de la corde correspondant à l'angle $(\mathbf{y}_{t1}, \mathbf{y}_{t2})$ - égale au plus à 2 quand ces 2 vecteurs sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi partir du tableau des racines carrées plutôt que celui des fréquences brutes ?), mais elle jouit de propriétés intéressantes :

- Escofier et Volle ont montré qu'elle satisfaisait à la propriété d'équivalence distributionnelle décrite plus haut, au même titre que la distance du khi-deux utilisée en AFC.
- Contrairement à celle-ci, elle peut prendre en compte des vecteurs ayant des composantes négatives, propriété utile pour certains types de codage "symétriques" (comme *Oui, Non, Ne sait pas*), ou pour des tableaux de flux divers, économiques ou autres.
- Elle est liée à la mesure du gain d'information de Renyi d'ordre $\frac{1}{2}$ (Renyi 1955) apporté par une distribution \mathbf{x}_q quand on connaît la distribution \mathbf{x}_p :

$$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = -2 \log_2 (\cos(\mathbf{y}_p, \mathbf{y}_q)) = -2 \log_2 (1 - Dd^2/2)$$
- Elle est rapide à calculer dans le cas des données textuelles, où les vecteurs \mathbf{y}_i sont très "creux".

Si on calcule les directions propres $\mathbf{U} = \{ \mathbf{u}^{(k)} \}$ (resp $\mathbf{W} = \{ \mathbf{w}^{(k)} \}$) du nuage des points-colonnes \mathbf{y}_t (resp., des points-lignes \mathbf{y}_i) défini plus haut (k est l'ordre de ces directions propres), on démontre 1) que les valeurs propres $\lambda^{(k)}$ sont communes aux deux nuages analysés, 2) que les cosinus ci-après se déduisent des directions propres :

$$\text{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} \sqrt{(\lambda^{(k)} / x_t)} \quad (1)$$

$$\text{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} \sqrt{(\lambda^{(k)} / x_i)} \quad (2)$$

Ces cosinus peuvent être considérés comme les facteurs d'un cas particulier et simple d'analyse factorielle sphérique, pour reprendre la terminologie de M. Volle, dite centrée sur le "tableau nul". Nous les nommerons désormais respectivement $F_t^{(k)}$ et $G_i^{(k)}$. Ils sont liés entre eux et avec les composantes $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :

$$F_t^{(k)} = \sum_i u_i^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$

$$G_i^{(k)} = \sum_t w_t^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$

Ce qui permet de les déduire à partir de l'extraction des éléments propres du plus petit des deux nuages de points.

Nos algorithmes K-means axiaux (Lelu, 1994) et Analyse en composantes locales (Lelu et Ferhan, 1998) réalisent la partition en classes des unités textuelles dans l'espace distributionnel, et extraient pour chaque classe les facteurs mots et documents définis ci-dessus. Chaque facteur est alors un indicateur de *centralité* (ou "typicité") du document ou du mot dans sa classe.

Contrairement à la distance du khi-deux et au cosinus de Salton, les calculs de similarité dans cet espace ne font pas intervenir une métrique de type *inverse document frequency*. Ceci dit, il est à noter que le passage d'un ensemble de documents à l'ensemble des mots qui en sont les plus caractéristiques se fait en deux étapes :

- 1) Calcul du 1^{er} vecteur propre (facteurs-documents, s'il y a moins de documents que de mots) et de sa valeur propre associée issus du tableau des racines carrées des fréquences brutes,
- 2) Calcul des facteurs-mots via la formule (2), qui fait intervenir une correction de type IDF.

On peut donc dire que la pondération de type IDF intervient aussi dans l'espace distributionnel, mais de façon moins évidente que dans le cosinus de Salton ou la distance du khi-deux. Nous comptons approfondir plus tard ce problème d'attribution de mots caractéristiques à un bloc de texte, dit de "fonction d'indexation", en comparant la solution ci-dessus à d'autres, en particulier celle basée sur un modèle hypergéométrique présentée dans (Lebart et Rajman 2000).

Nos algorithmes ont fait la preuve depuis un bon nombre d'années de résultats "sensés et interprétables" sur de nombreux corpus de toutes tailles, de toutes provenances, et de tous types de codage.

Pour convaincre de leur bien-fondé d'autres personnes que les utilisateurs directs de nos méthodes, la nécessité d'une validation moins subjective se fait sentir, qui permette en particulier une comparaison avec les mesures utilisées de façon courante par ailleurs.

2. COMPARAISONS EMPIRIQUES

2.1. Le corpus utilisé

Parmi les corpus mis à disposition par la campagne d'évaluation de systèmes d'information Amaryllis 99 (ref. : Amaryllis 1999) figure celui de l'ELDA (ref. : Elda), constitué de 3511 questions écrites des députés européens à la Commission des Communautés européennes, avec les réponses de la Commission ; ce qui représente en français 8 Mo de textes bruts, sans mise en page, soit en moyenne autour de 2300 signes par texte. Ces textes comportent également des versions dans les principales langues européennes, rassemblées pour alimenter la tâche multilingue de cette évaluation.

2.2. Les traitements effectués

Nous sommes partis pour la présente expérience de l'ensemble des mots lemmatisés et des candidats termes composés proposés par le logiciel de traitement morpho-syntaxique Nomino (Plante et al., 1997) : 82 572 termes hors les mots grammaticaux, les verbes opérateur comme *avoir* et *être* ; nous avons éliminé par seuillage les hapax et les mots de fréquence supérieure à 1500 : restent 20 034 formes lemmatisées différentes. Nous n'avons pas éliminé les verbes et les adjectifs, contrairement à notre pratique habituelle sur les corpus documentaires, afin de rester les plus fidèles possible au contenu d'origine. Afin d'enlever un maximum de bruit dû aux polysémies, nous avons retiré de l'indexation de chaque texte les mots simples dont la composition formait les mots composés - nous verrons plus loin que cette option a des répercussions importantes en matière de stratégie de recherche d'information.

2.3. Les requêtes de test et leurs expansions

Les organisateurs d'Amaryllis 99 ont fourni au total 30 requêtes, associées chacune avec l'ensemble de documents pertinents validé par des experts. Par manque de temps, nous n'avons exploité que les 15 premières. Nous avons traité comme un seul texte l'ensemble des champs de chaque requête - titre, sous-titre, explications, mots-clés - par le logiciel Nomino. Ainsi par exemple :

Q13. Pollution atmosphérique.
Pollution de l'air et lutte antipollution.
Tous les documents sur la pollution atmosphérique et intérieure par dioxyde et monoxyde de carbone, oxydes de soufre et d'azote, ozone, plomb, particules ... (réglementation, lutte antipollution, aspect économique).
Pollution air.
Lutte antipollution.
Pollution intérieure.
Emission particule.
Carbone dioxyde.
Carbone monoxyde.
Effet serre.
Azote oxyde.
Soufre oxyde.

Après importation des termes trouvés (tous n'étaient pas présents dans le corpus...) nous avons utilisé notre procédure d'expansion lexicale Proxilex (Lelu et al., 1998) sur chacun de ses composants, puis réalisé une expansion "sémantique" par similarité {mots} → mots) dans notre environnement de contrôle de vocabulaire NeuroNav (NeuroNav) où nous avons sélectionné manuellement les termes pertinents parmi les 250 termes les plus associés aux termes de requête.

Ainsi avons-nous obtenu pour la requête N°13 une liste de 81 mots reformulant cette requête initiale, auxquels nous n'avions le plus souvent pas pensé à l'origine, et dont voici un extrait :

<i>énergie</i>	<i>limiter</i>	<i>émission</i>
<i>teneur</i>	<i>réduire</i>	<i>CO2</i>
<i>protection</i>	<i>substance</i>	<i>gaz</i>
<i>qualité</i>	<i>réglementation</i>	<i>combustion</i>
<i>fixer</i>	<i>risque</i>	<i>serre</i>
<i>installation</i>	<i>énergies_renouvelables</i>	<i>climatique</i>
<i>environnement</i>	<i>taxe</i>	<i>moteur</i>
<i>air</i>	<i>lutte</i>	<i>rejet</i>
<i>pollution</i>	<i>eau</i>	<i>prévention</i>
<i>concentration</i>	<i>charbon</i>	<i>station_de_mesure</i>
<i>SO2</i>	<i>gaz_à_effet_de_serre</i>	<i>pollution_marine</i>
<i>NO2</i>	<i>émissions_de_dioxyde_de_carb</i>	<i>pollution_atmosphérique</i>
<i>polluer</i>	<i>one</i>	<i>sol</i>
<i>effet</i>	<i>émissions_de_CO2</i>	<i>soufre</i>
<i>seuil</i>	<i>dioxyde</i>	<i>azote</i>
<i>valeur</i>	<i>carbone</i>	<i>particule</i>

Nous avons vérifié sur un exemple que le choix de termes effectué parmi une longue liste des 250 premiers éléments produit les mêmes termes pertinents, mais avec un ordre et un rang différent, quelle que soit la mesure de similarité utilisée pour l'expansion.

2.4. Etablissement des courbes Rappel / précision

La méthodologie d'évaluation des systèmes documentaires classique depuis (Salton 1968) définit les notions de taux de rappel et taux de précision :

Rappel = nombre de documents pertinents retrouvés / nombre de documents retrouvés

Précision = nombre de documents pertinents retrouvés / nombre de documents pertinents

Nos trois mesures de similarité (cosinus de Salton, distance du khi-deux changée de signe, et cosinus dans l'espace distributionnel) donnent pour notre liste de mots en requête trois listes de titres les plus proches. Pour chaque liste, et pour chaque titre pertinent trouvé par ordre de similarité décroissante, nous calculons les indices de rappel et précision. Pour un exemple de requête nous avons reporté ces valeurs sur la figure 1, où nous avons dessiné également, après lissage, les trois courbes Rappel / Précision respectives. Au total, la figure 2 montre les trois courbes Rappel / Précision pour l'ensemble des requêtes.

2.5. Résultat et discussion

Plus une courbe est haute et plus la concentration de documents pertinents est élevée en début de liste de documents restituée, ce qui est le but recherché : le cosinus distributionnel sort vainqueur de cette comparaison, sauf pour les valeurs moyennes du taux de rappel (0.4 à 0.6), où le cosinus de Salton fait jeu égal avec lui. La distance du khi-deux est systématiquement moins bonne, ce que confirme une expérience antérieure menée sur un corpus-jouet de 8 documents et 7 mots : après expansions exhaustives sur chaque document, la moyenne des corrélations de rang de Spearman entre cos. distributionnel et khi-deux est négligeable (pas de relation) alors qu'elle est de 0,25 entre cos. de Salton et khi-deux, de 0,45 entre cos. de Salton et cos. distributionnel.

Il est possible que le type d'indexation "pointue" adopté, principalement à base de mots composés lemmatisés et de mots simples non redondants avec ceux-ci, donc avec peu de termes génériques, défavorise les similarités avec pondération des mots de type IDF, comme le cosinus de Salton, et plus encore la distance du khi-deux. Des comparaisons dans le cadre de stratégies d'indexation différentes seraient à réaliser pour approfondir cette question, que nous n'avons pas trouvé mentionnée à notre connaissance dans la littérature concernant la recherche d'information.

Mais il faut aussi noter que ce type d'indexation donne d'excellents taux de rappel, 100% en général, en tête des listes de réponses, quelle que soit la mesure de similarité utilisée, du fait de l'absence d'ambiguïté et de la précision conceptuelle des termes composés.

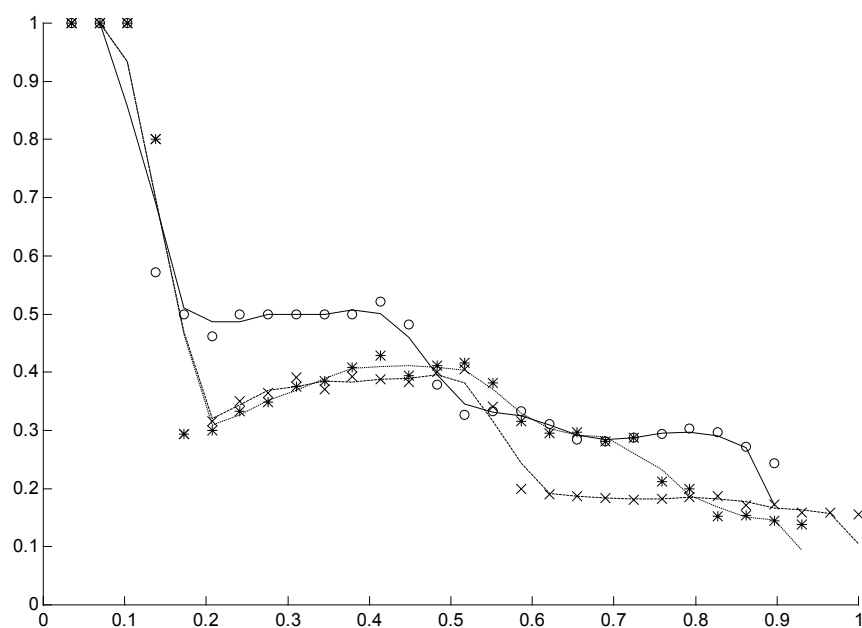


Figure 1 - Courbes Rappel / précision de la question 1 pour les 3 mesures de similarité.

En abscisse : taux de précision ; en ordonnée : taux de rappel

o : *cosinus dans l'espace distributionnel*

* : *cosinus de Salton*

x : *distance du khi-deux*

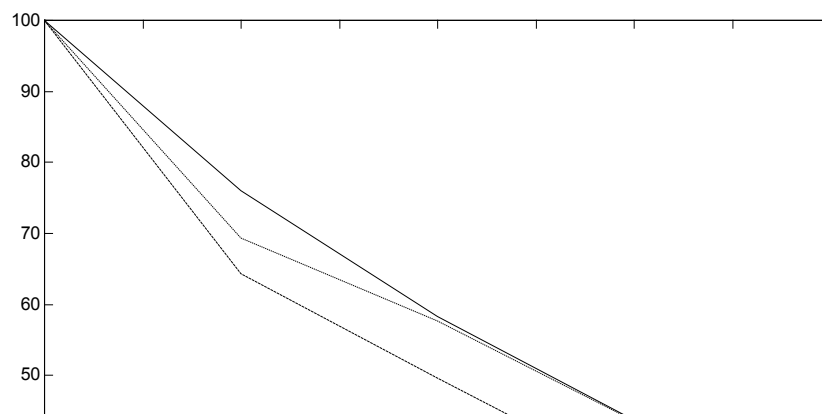


Figure 2 - Courbes Rappel / précision globales pour les 3 mesures de similarité.

En abscisse : taux de précision ; en ordonnée : taux de rappel

o : *cosinus dans l'espace distributionnel*

* : *cosinus de Salton*

x : *distance du khi-deux*

CONCLUSIONS ET PERSPECTIVES

Sur le strict plan de l'étude présentée ici, il serait intéressant de la prolonger par des comparaisons 1) par rapport à d'autres indicateurs de liens, comme ceux issus des théories de l'information, 2) par rapport à d'autres choix d'indexation – filtrer certaines catégories grammaticales ? Ne pas supprimer les composants des mots composés dans une même unité textuelle ? Voire “ indexer ” par des N-grammes, au lieu de mots (Lelu et al. 1998) ?

Un autre axe intéressant à explorer serait la comparaison des résultats de similarités, sur un même corpus, et à transformation égale de l'espace des données, entre l'indexation sémantique latente (LSI : cf. LSA) et l'indexation sémantique *patente* que réalisent nos algorithmes K-means axiaux et Analyse en composantes locales – ils extraient des axes obliques et interprétables, au lieu d'axes orthogonaux et en majeure partie ininterprétables, sur lesquels les mots et les documents n'ont que des composantes positives ou nulles ; composantes qui traduisent une décomposition *additive* des données, plausible sur le plan cognitif (un texte, ou un mot, se rattache, avec plus ou moins de force, à un ou plusieurs thèmes sémantiques parmi de nombreux répertoriés dans le corpus). Cette comparaison devrait être éclairante quant au nombre de dimensions pertinentes sous-jacentes à un corpus textuel typique de quelques méga-octets : moins d'une dizaine de dimensions, d'après la pratique recommandée pour les logiciels SPAD/T (Lebart et Salem, 1994) et Alceste ? De 200 à 300, d'après l'équipe LSA ?

Nous espérons avoir contribué à faire avancer la question de l'évaluation comparée des méthodes en vigueur dans le domaine de l'exploitation des corpus textuels, au sens large, en montrant que pour au moins une stratégie d'indexation automatique le cosinus distributionnel est plus pertinent pour les calculs de similarité que le cosinus “ TF-IDF ” de Salton, lui-même meilleur que la distance du khi-deux. C'est seulement sur ce socle constitué par la transformation de l'espace des données (espérons que nos résultats seront complétés, nuancés ou, pourquoi pas, contestés par des tests ultérieurs !) que les diverses représentations synthétiques des données pourraient être comparées toutes choses égales par ailleurs – que ce soient les nombreuses variantes des analyses factorielles orthogonales ou obliques (Analyse en Composantes Principales, Latent Semantic Analysis, Independent Component Analysis, Analyse en Composantes Locales, Non-negative Matrix Factorization...), et celles des classifications supervisées ou non (pour les non-supervisées : K-means, K-means axiaux, classifications ascendantes ou descendantes hiérarchiques, réseaux de Kohonen, ...) ; en ce qui concerne les méthodes supervisées (analyses discriminantes, perceptrons multi-couches, réseaux RBF, Support Vector Machines, ...) cette distinction permettrait de faire la part de l'algorithme proprement dit dans la qualité des résultats obtenus, par rapport à celle du pré-traitement des données.

REMERCIEMENTS :

Aux organisateurs de la campagne d'évaluation Amaryllis 1999.

A l'agence européenne ELDA (Evaluations and Language resources Distribution Agency) pour avoir fourni le corpus de test utilisé et les couples questions-réponses permettant les tests.

RÉFÉRENCES

ALCESTE : cf. www.image.cict.fr/

Amaryllis 1999 : cf. <http://amaryllis.inist.fr/>

Benzécri J.P. et coll. (1981). Pratique de l'Analyse des Données : Linguistique et Lexicologie. Dunod, Paris.

Bourdoncle F. (1997). LiveTopics : recherche visuelle d'information sur l'Internet. In C. Jacquemin éditeur, Proc. of RIAO'97 (Recherche d'Information Assistée par Ordinateur), CID, Paris.

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

Dumais S.T. (1994). Latent Semantic Indexing (LSI) and TREC-2. NIST special publication, N°500-215, pages 105-115, NIST.

ELDA : cf. www.elda.fr/

Escofier B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionnelle. *Revue de Stat. Appliquée*, 26(4):29-37, Paris

Fichet B. et Gbegan A. (1985). Analyse factorielle des correspondances sur signes de présence-absence. In Diday et al. editors, 4^e Journées Analyse des données et Informatique. INRIA, Rocquencourt.

Matusita, Kameo (1955). Decision rules, based on the distance for problems of fit, two examples, and estimation. *Annals of Statistical Mathematics*, pages 631-640, Tokyo.

Kohonen T., Kaski S., Lagus K., Honkela T. (1995). Very large two-level SOM for the browsing of newsgroups. Proc. of WWW'95 (5th International World Wide Web Conference), Paris.

<http://websom.hut.fi/websom>

Lebart L., Morineau A., Tabard N. (1977). *Techniques de la description statistique*. Dunod, Paris.

Lebart L., Rajman M. (2000). Computing Similarities. In Dale R., Moisl H., Somers H. editors : *Handbook of Natural Language Processing*, Marcel Dekker, pages 477-505, New York.

Lebart L., Salem A. (1994). *Statistique textuelle*. Dunod, Paris.

Lelu A., Tisseau-Pirot A.G. (1993). Emergence de catégories sémantiques à partir d'une base de résumés d'articles. In Anastex S.J. editor, Proc. of JADT'98 (2emes Journées Internationales d'Analyse Statistique des données Textuelles), pages 227-242, ENST, Paris.

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

Lelu A., Tisseau-Pirot A.G., Adnani A. (1997). Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation. *Hypertextes et Hypermédiats*, 1(1):23-55, Hermès, Paris,

Lelu A., Ferhan S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In Rizzi A. et al. editors, Proc. of IFCS'98 (6th Conference of the International Federation of Classification Societies), pages 206-209, Università La Sapienza, Roma.

Lelu A., Hallab M., Delprat B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes. In Mellet S. editor, Proc. of JADT'98 (4emes Journées Internationales d'Analyse Statistique des données Textuelles), pages 391-400, UPRESA " Bases, Corpus et Langage ", Université de Nice.

LSA : cf. <http://lsa.colorado.edu/>

NeuroNav : cf. www.diatopie.com

Plante P., Dumas L. et Plante A. (1997). Atelier FX. ATO, Département de Linguistique, Université du Québec à Montréal.

<http://www.ling.uqam.ca/Ato/FX>

- Reinert M. (1993). "Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars", *Langage et Société*, 66, 5-39.
- Renyi A. (1966). *Calcul des probabilités*. Dunod, Paris.
- Rhissassi H. and Lelu A. (1998). Indexation assistée et cartographie sémantique pour la génération automatique d'hypertexte. In Mojahid M. editors, *Proc. of CIDE'98*, pages 131-139, Europa Productions, INPT, Rabat, Maroc.
- Salton G. (1968). *Automatic Information Organization and Retrieval*. Mac Graw Hill, NY.
- Salton G. and Mac Gill M.J (1983). *Introduction to Modern Information Retrieval*. International Student Edition.
- Zamir O., Etzioni O. (1999). Grouper : a dynamic Clustering Interface to Web Search Results. *Proc. of WWW'99 (8th International World Wide Web Conference)*.