Comparaison de trois mesures de similarité utilisées en documentation automatique et analyse textuelle

Alain Lelu

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

Abstract

A small number of similarity indicators are underlying the rich diversity of methods and tools in the overlapping information retrieval 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.

Résumé

Sous-jacents à la diversité des méthodes et outils de la documentation automatique 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 sur un corpus réel par la méthode "rappel / précision".

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

1. 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 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, caractérisé par une série de choix situés de fait 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, chaque phrase, etc.

B - **Choix de codage des textes** : à chaque texte on fait correspondre des descripteurs codés. 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 en texte plein, ou par lemmes 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 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 l'environnement sémantique d'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 Otzioni, 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, 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 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 ne portent jamais 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 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 **M** carrée généralement symétrique, définie positive, intervenant dans la définition du produit scalaire : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \mathbf{x}' \mathbf{M} \mathbf{y}$

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

Les combinaisons décrites 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". 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 ses multiples domaines d'intervention (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…).

2. Trois mesures de similarité dans l'espace des unités textuelles et des mots

2.1. Le cosinus de Salton

Notations utilisées :

$x_{.t} = \Sigma_i x_{it}$; $x_{i.} = \Sigma_t x_{it}$; $x_{..} = \Sigma_i \Sigma_t x_{it}$; **Diag** [a] : matrice diagonale d'éléments a

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

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

(et par conséquent le carré de la norme de **x** est : $\|\mathbf{x}\|^2_{\mathbf{D}} = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{D}}$)

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). 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 :

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

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

Ce qui revient à effectuer le produit scalaire standard entre vecteurs-documents $\mathbf{x}_t$ transformés de la façon suivante :

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

Nous utiliserons cette formule dans notre évaluation pratique, à la section suivante. Mais pour les besoins de notre comparaison, on peut remarquer que N/n_i est égal à de $x_{..}/x_{i.}$ si les mots sont codés en présence/absence, ou peu différent, 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 :

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

En résumé, $\text{Cos Salton}\ (t_1,t_2) \cong \langle \mathbf{x}_{t1}, \mathbf{x}_{t2} \rangle_{\mathbf{Ds}} / \| \mathbf{x}_{t1} \|_{\mathbf{Ds}} \times \| \mathbf{x}_{t2} \|_{\mathbf{Ds}}$

2.2. 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 sur le simplexe, hyperplan lieu des points $\mathbf{z}$ tels que $\Sigma_i z_i = 1$ 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, du moins 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 type *inverse document frequency* (IDF) :

Diag($x_{..}/x_{i.}$) pour l'AFC, **Diag**($\log(N/n_i)$), souvent équivalent à **Diag**($\log(x_{..}/x_{i.})$), 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.

2.3. Le cosinus dans l'espace distributionnel

Une lignée ancienne de travaux (Kameo et 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*).
- Elle est liée à la mesure du gain d'information de Renyi d'ordre ½ (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 $\mathbf{y}_t$ 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, on démontre que les cosinus ci-après se déduisent des directions propres :

$$\text{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} \sqrt{(\lambda^{(k)} / x_{.t})} \qquad (1)$$

$$\text{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} \sqrt{(\lambda^{(k)} / x_{i.})} \qquad (2)$$

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 $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :

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

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

Ce qui permet de les calculer à partir de l'extraction des éléments propres du moins important des deux nuages de points.

Nos algorithmes K-means axiales (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 1er vecteur propre 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 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.

3. Comparaisons empiriques

3.1. Le corpus utilisé

L'organisation non-gouvernementale Fondation pour le Progrès de l'Humanité édite depuis une dizaine d'années une base documentaire consacrée aux problèmes du développement dans les pays du tiers monde et dans les milieux défavorisés des pays riches, dans un but de capitalisation d'expérience. Nous en avons extrait en 1998 les 4165 résumés en français, qui sont de tailles et d'inspirations très diverses : compte-rendus d'expériences militantes, opinions et réflexions, compte-rendus de lecture, … ; ce qui représente 13 Mo de textes bruts, sans mise en page, soit en moyenne autour de 3000 signes par texte.

3.2. Les traitements effectués

Ce corpus avait déjà fait l'objet d'une indexation assistée dans notre environnement spécifique (Rhissassi et Lelu, 1998), mais pour minimiser la part de subjectivité inhérente à tout nettoyage manuel de vocabulaire, 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), hors les mots grammaticaux, les verbes opérateur comme *avoir* et *être*, et les hapax, soient 47 319 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.

3.3. La requête de test

Notre but a été de choisir une requête 1) ni trop générale (comme l'aurait été "drogue et sous-développement") - elle drainerait plusieurs centaines de documents pertinents et serait propice aux incertitudes de frontières entre documents pertinents ou non, 2) ni trop précise (comme le serait "Les pays émergents face aux accords de Kyoto") - il pourrait ne lui correspondre que 2 ou 3 documents pertinents, interdisant de ce fait la méthodologie décrite plus bas.

D'où le choix de notre thème de requête : "Effets pervers de l'action humanitaire", après utilisation de notre procédure d'expansion lexicale Proxilex (Lelu et al., 1998) sur chacun de ses composants, puis d'une série d'itérations entre calculs de similarité ({mots}→ docs puis {docs}→ mots) utilisant successivement les trois mesures de similarité testées décrites ci-dessus, afin de n'en favoriser aucune ; ce choix nous a permis 1) d'isoler un sur-ensemble de quelques centaines de documents, parmi lesquels la lecture directe des textes a permis de sélectionner les 28 titres de documents pertinents ci-après :

```
. Confrontation.de.logiques.endogènes.et.exogènes...Effets.indésirables.
. Contre.l'humanitaire-alibi...L'humanitaire.instrumentalisé.-2871;.0389
. 'Aucun.enfant.n'a.choisi.de.naître.dans.un.camp.ou.dans.l'autre'.-2884
. Journal.de.voyage.en.Bolivie.mars-avril.1988.-0239;.00255.............
. Le.rôle.de.l'église.au.Guatemala.au.cours.des.dix.dernières.années.-30
. Le.nouvel.ordre.mondial.:.un.rôle.nouveau.pour.l'ONU.?..-.1809;.02279.
. La.composition.'hétérogène'.des.membres.d'Enfants.Réfugiés.du.Monde,.d
. Non.assistance.à.population.en.danger...L'inadéquation.de.la.philosoph
. L'accueil.des.réfugiés.bosniaques.et.des.personnes.déplacées.croates.-
. Du.Liban.au.Rwanda,.en.passant.par.le.Guatemala,.quinze.ans.d'action.h
. Perspectives.de.la.politique.agricole.-0892;.01042...................
. L'acteur.et.le.système.-3132;.04321.................................
. L'acteur.et.le.système...1...Les.contraintes.de.l'action.collective.-0
. L'aide.internationale.en.faveur.de.l'Afrique.en.question.(2)...Le.rôle
. Pratiques.de.formation.dans.les.projets.de.développement.-0409;.00446.
. Un.projet.a.priori.'innocent'.peut.révéler.des.enjeux.insoupçonnés.et.
. L'aide.alimentaire.:.bénéfice.ou.préjudice?..-.1701;.02129............
. Réflexions.sur.l'histoire.de.la.réhabilitation.du.quartier.Mermoz.à.Ly
. Aide.alimentaire.exagérée.au.Guatemala.suite.au.séisme.de.1976.-3012;.
. Pour.une.nouvelle.politique.d'aide.en.faveur.d'un.développement.durabl
. La.situation.actuelle.des.groupements.de.femmes.au.Bénin.-2373;.03137.
. 'Shonar.Bangla.Friends'.-2-...Acquis.et.problèmes.-2038;.02638........
. Méthodologie.du.travail.en.droit.alternatif.-0923;.01073..............
. La.médiation.:.une.autre.justice.-1761;.02211.........................
. Capitalisation:.quelle.diffusion.pour.le.témoignage?..-.2138;.02760...
. Promouvoir.la.paix.et.la.réconciliation.au.Cambodge.à.travers.une.phil
. Pour.une.nouvelle.philosophie.de.l'action.humanitaire...De.la.compassi
```

2) une liste de 25 mots reformulant notre requête initiale, auxquels nous n'avions le plus souvent pas pensé à l'origine :

```
effet_pervers..........
effet_indésirable.......
effet_inverse..........
effet_néfaste..........
effet_de_système........
effet_nocif............
effet_négatif..........
effet_contraire........
effet_sensible.........
destabiliser...........
déstabilisation........
déstabilisatrice........
démarche_humanitaire....
effort_humanitaire......
opération_humanitaire...
action_humanitaire......
```

```
assistance_humanitaire..
intervention_humanitaire
mission_humanitaire.....
mécénat_humanitaire.....
soutien_humanitaire.....
caractère_humanitaire...
aide_humanitaire........
organisation_humanitaire
association_humanitaire.
```

Cette liste constitue notre requête de test pour la suite de notre comparaison.

Le caractère itératif de cette démarche, au bout de laquelle aucune expansion ne fournissait plus d'idées de mots nouveaux à ajouter ni de nouveaux textes pertinents, nous a convaincu d'avoir fait raisonnablement le tour du problème, sans devoir lire la totalité des documents…

3.4. 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 deux indices. Ces valeurs sont reportées sur la figure 1 où nous avons dessiné également, après lissage, les trois courbes Rappel / précision respectives.

3.5. 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, du moins jusqu'au taux de rappel de 0,7. Le cosinus de Salton a un comportement intermédiaire entre celui-ci et la distance du khi-deux, 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.

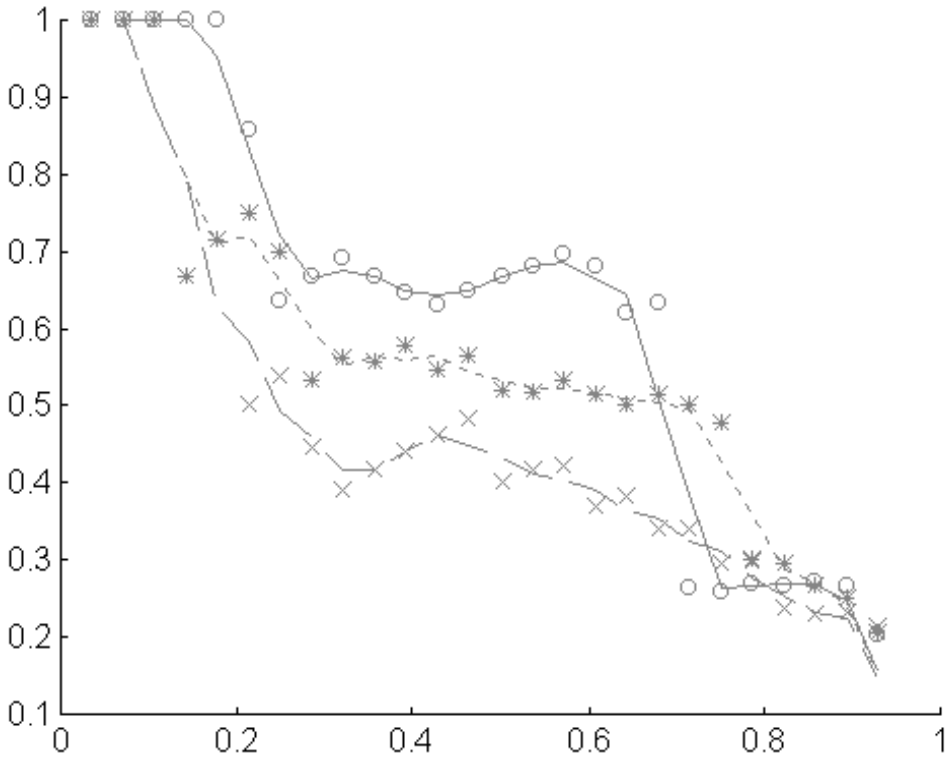

Figure 1 - Courbes Rappel / précision pour les 3 mesures de similarité
En abscisse : taux de précision ; en ordonnée : taux de rappel
o : cosinus dans l'espace distributionnel
* : cosinus de Salton
× : distance du khi-deux

4. Conclusions et perspectives

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 donne de meilleurs résultats que le cosinus « TF-IDF » de Salton, lui-même meilleur que la distance du khi-deux.

Nous comptons présenter à la conférence une évaluation plus systématique, à partir d'un jeu de données documentaire publié, accompagné d'une quinzaine de questions / réponses validées.

Il serait nécessaire de prolonger cette recherche par des comparaisons par rapport à d'autres indicateurs de liens, comme l'écart réduit ou ceux issus des théories de l'information.

Un autre axe intéressant à explorer serait la comparaison des résultats, sur un même corpus, et à mesure de similarité égale, de l'indexation sémantique latente (LSI) et de l'indexation sémantique *patente* que réalisent nos algorithmes KMA et ACL – ils extraient des axes

obliques et interprétables, au lieu d'axes orthogonaux et en majeure partie ininterprétables. Cette comparaison devrait être éclairante quant au nombre de dimensions pertinentes sous-jacentes à un corpus textuel typique de quelques Mo - moins d'une dizaine, d'après la pratique recommandée pour le logiciel SPAD/T (Lebart et Salem, 1994), de 200 à 300 d'après l'équipe LSI.

Remerciements

A la Fondation pour le Progrès de l'Humanité dont la collaboration avec l'UFR LIT de l'université Paris 8 en 1997-98 a permis d'obtenir les résultats présentés.

Références

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 editor, *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.

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, *4e Journées Analyse des données et Informatique*. INRIA, Rocquencourt.

Kameo et Matusita (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 Hypermédias*, 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, Universita La Sapienza, Roma.

Lelu A., Halleb 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.

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>*

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, Europia 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).*