SIMILARITÉS POUR DONNÉES TEXTUELLES

Martin Rajman
LIA
Laboratoire d'Intelligence Artificielle
Ecole Polytechnique Fédérale
CH-1015 Lausanne, Suisse
e-mail: rajman@lia.di.epfl.ch

Ludovic Lebart
CNRS
Ecole Nationale Supérieure des
Télécommunications
Paris, France
e-mail: lebart@enst.fr

Abstract: Similarities for textual data

The evaluation of similarities between textual entities (documents, sentences, words...) is one of the central issues for the implementation of efficient methods for tasks such as description and exploration of textual data, information retrieval or knowledge extraction (text mining).

The main purpose of this contribution is to propose a comparative presentation of different approaches used to define the notion of similarity in fields such as Textual Data Analysis, Information Retrieval or Text Mining.

We first discuss some of the linguistic treatments (tagging, lemmatization, …) necessary for the pre-processing of the textual data and then analyze some of the measures (cosinus, chi-square, Kullback-Leibler) used to quantify the similarities (or dissimilarities) between textual entities.

Finally, we present some techniques allowing to improve the quality of the similarities in the case where additional knowledge, such as external corpora or semantic graphs, is available.
Keywords: similarity measure, textual data analysis, information retrieval, text mining

0 Introduction

Évaluer des similarités entre entités textuelles est un des problèmes centraux dans plusieurs disciplines comme l'analyse de données textuelles, la recherche documentaire ou l'extraction de connaissances à partir de données textuelles (Text Mining). Dans chacun de ces domaines, les similarités sont en effet utilisées pour une large variété de traitements :

- en analyse de données textuelles (ADT), les similarités sont utilisées pour la description et l'exploration de données, pour l'identification de structures cachées et pour la prédiction ;
- en recherche documentaire (RD), l'évaluation des similarités entre documents et requêtes est utilisée pour identifier les documents pertinents par rapport à des besoins d'information exprimés par les utilisateurs;
- en Text Mining (TM), les similarités sont utilisées pour produire des représentations synthétiques de vastes collections de documents, dans le cadre de procédures d'extraction d'information à partir de données textuelles.

Les techniques mises en œuvre pour calculer les similarités varient bien évidemment selon les disciplines, mais elles s'intègrent cependant le plus souvent dans une même approche générale en deux temps :

1. Les entités textuelles sont tout d'abord associées à des représentations spécifiques

qui vont servir de base au calcul des similarités ;

La nature précise des représentations utilisées dépend fortement du domaine d'application considéré.

En ADT, on utilise souvent les profils lexicaux alors qu'en RD et TM des distributions (éventuellement pondérées) de mots-clés ou des vecteurs contextuels de co-occurrences sont mis en œuvre.

Il est important de remarquer que, dans tous les cas, les structures associées sont représentées sous la forme d'éléments d'un espace vectoriel de grande dimension que nous appellerons ci-après « l'espace de représentation ».

2. Un modèle mathématique est choisi pour mesurer, dans l'espace de représentation, les proximités qui seront utilisées pour estimer les similarités entre entités textuelles.

En ADT, la distance du chi-deux (χ^2) est un choix fréquent. En RD, des similarités dérivées de mesures à base de cosinus sont utilisées alors qu'en TM on préfère souvent des mesures issues de la théorie de l'information (comme la « distance » de Kullback-Leibler à base d'entropie relative).

L'objectif de cette contribution est de décrire différentes approches, méthodes et techniques utilisées, dans les domaines mentionnés ci-dessus, pour traiter les similarités entre entités textuelles.

La section 1 est consacrée à une discussion sur les unités linguistiques qui peuvent être utilisées pour représenter les textes pour le calcul des similarités.

La section 2 décrit quelques unes des mesures fréquemment utilisées pour évaluer les similarités (ou dissimilarités) et présente rapidement certaines de leurs propriétés.

Enfin, la section 3 examine les différentes possibilités d'amélioration des similarités utilisées dans le cas où des informations additionnelles comme des corpus externes ou des graphes sémantiques sont disponibles.

1. Représenter les textes pour l'évaluation des similarités

Afin de produire les structures qui vont être utilisées pour représenter les textes lors du calcul des similarités, les données textuelles doivent tout d'abord être décomposées en unités lexicales plus simples. Plusieurs choix sont possibles et les différentes unités retenues auront des degrés de pertinence variables selon le domaine d'application particulier choisi.

Une approche classique pour définir les unités textuelles dans une corpus est d'utiliser les formes de surface (« mots ») pouvant être produites par des techniques simples de segmentation automatique. Cependant, ces unités élémentaires peuvent également faire l'objet de traitements additionnels permettant l'intégration de connaissances linguistiques plus sophistiquées dans les représentations. L'étiquetage morpho-

syntaxique (i.e. l'affectation automatique aux mots d'étiquettes grammaticales) et/ou la lemmatisation (i.e. la réduction automatique des formes fléchies à une représentation canonique – infinitif pour les verbes, singulier pour les noms, …) sont de bons exemple de telles techniques de pré-traitement des données textuelles.

De plus, du fait que le sens des mots est fortement lié à la manière dont ils apparaissent en combinaison (par exemple, des expressions composées comme « sécurité sociale » ou « niveau de vie » ont des significations qui ne peuvent être simplement dérivées du sens de leurs constituants), il peut également être utile de prendre en compte des unités plus larges constituées de plusieurs mots. L'utilisation des « segments répétés » (Salem, 1987) ou des « quasi-segments » (Becue, 1993), reposant sur la détection automatique des séquences répétitives, constituant ou non des formes ou expressions composées, est une solution possible mais des approches combinant des connaissances linguistiques et statistiques pour identifier de façon automatique les formes composées (ou termes) sont aujourd'hui également disponibles (Daille, 1994).

Il est cependant à noter que l'utilisation de techniques d'extraction plus sophistiquées pour les unités servant à la représentation des textes présuppose la disponibilité des ressources linguistiques nécessaires (ce qui n'est pas forcément le cas pour toutes les langues) et, de plus, augmente de façon sensible le nombre total d'unités à prendre en compte dans les étapes ultérieures de traitement (calcul des similarités, sélection et visualisation des entités textuelles, …). Le choix de la nature des constituants des représentations est de ce fait un compromis nécessaire entre, d'une part, la qualité et la disponibilité des outils de pré-traitement linguistique et, d'autre part, les contraintes (tailles des données, temps de traitement, ressources informatiques disponibles, …) imposées pour le calcul des similarités.

2. Mesures de similarité

Pour toute représentation de document prenant la forme d'un tuple de valeurs D, soit :

$|D|$, la norme de D : $|D| = \sqrt{D^2}$ (où $D^2 = D \bullet D$ et $\bullet$ représente le produit scalaire sur les tuples); D_1 / D_2, la restriction de D_1 aux coordonnées non nulles de D_2, i.e. la représentation obtenue en remplaçant par 0 toutes les coordonnées de D_1 pour lesquelles la coordonnée correspondante dans D_2 est nulle; en termes informels, D_1 / D_2 est « l'intersection » de D_1 avec D_2 ; $D_1 \setminus D_2$, la restriction de D_1 aux coordonnées nulles de D_2, i.e. la représentation obtenue en remplaçant par 0 toutes les coordonnées de D_1 pour lesquelles la coordonnée correspondante dans D_2 est non nulle; et $\max(D)$, la coordonnée maximale de D.

Notons que, par définition, D_1 / D_2 and $D_1 \setminus D_2$ vérifient les propriétés suivantes :

$D_1 = D_1 / D_2 + D_1 \setminus D_2$ and $D_1 / D_2 \bullet D_1 \setminus D_2 = 0$

2.1 Distance du chi-deux pour l'analyse de données textuelles

L'analyse de données textuelles s'intéresse essentiellement à l'évaluation de similarités

entre documents. Usuellement, chaque document est représenté par son profil lexical, i.e. un tuple D_i qui contient les fréquences des unités textuelles dans le document ($D_i = \left(f_{i,j}\right)_j$, où $f_{i,j}$ est la fréquence de la j^{ème} unité dans le document D_i). Le corpus est alors représenté par une matrice **T** dont la i^{ème} ligne est la représentation du du i^{ème} document.

La similarité entre les documents est mesurée par une distance, appelée la *distance du chi-deux* (χ^2), très proche de la distance euclidienne (somme des carrés des différences entre les composantes des profils) mais avec une pondération ($1/f_{.j}$) associée à chacun des termes de la somme.

Plus formellement, on a :

$$\mathrm{d}_{\chi 2}(D_i, D_{i'})^2 = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2 \text{, où } f_{i.} = \sum_j f_{i,j} \text{ et } f_{.j} = \sum_i f_{i,j}$$

ou, si l'on intègre les pondérations dans la représentation des documents :

$D_i = \left(w_{i,j}\right)_j$, avec $w_{i,j} = p_{i,j} / \sqrt{f_{.j}}$, où $p_{i,j}$ est la fréquence relative de la j^{ème} unité dans le document D_i ($p_{i,j} = f_{i,j} / f_{i.}$)

et $\mathrm{d}_{\chi 2}(D_i, D_{i'}) = \left| D_i - D_{i'} \right|$

<u>*Quelques propriétés de la distance du chi-deux:*</u>

(1) Équivalence distributionnelle (Escofier, 1978)

Les distances entre les lignes (resp. colonnes) restent inchangées lors de la fusion de deux colonnes (resp. lignes) de même profil. Cette propriété d'invariance induit une certaine stabilité des résultats pour les analyses textuelles: en effet, deux textes ayant le même profil lexical pourront être indifféremment considérés comme une seule entité ou deux entités distinctes sans que cela n'affecte en rien les autres distances.

(2) $\mathrm{d}_{\chi 2}(D_1, D_2)^2 = \mathrm{d}_{\chi 2}(D_1 / D_2, D_2 / D_1)^2 + (D_1 \setminus D_2)^2 + (D_2 \setminus D_1)^2$

La propriété (2) montre que la distance du chi-deux est une mesure de proximité particulièrement sensible aux « différences hors intersection » entre les entités textuelles. La sensibilité aux différences n'a bien sûr rien de surprenant puisqu'une distance est, par définition, une dissimilarité et est de ce fait une fonction qui croît lorsque les différences entre les entités comparées augmentent. Ce qui, par contre, est notable (et ne sera pas le cas pour la distance KL ci-après) est que les « différences hors intersection » jouent un rôle important dans le calcul de la valeur de la dissimilarité. La conséquence de cette propriété est que la distance du chi-deux est a priori peu adaptée aux situations où les tailles des entités textuelles comparées sont fortement différentes (ce qui est par exemple souvent le cas en recherche documentaire lors de l'évaluation de similarités entre courtes requêtes et longs documents).

3.2 Similarités à base de cosinus pour la recherche documentaire

En recherche documentaire, le problème principal est d'évaluer les similarités entre les éléments stockés dans une base documentaire et des requêtes représentant les besoins d'information exprimés par les utilisateurs. Dans le cadre du modèle vectoriel classique, les approches utilisant des métriques à base de cosinus sont les plus fréquentes (Salton and Buckley, 1990). Différentes variations de cette approche ont été implémentées dans le système SMART, bien connu dans le domaine de la recherche documentaire (Salton and Buckley, 1988).

Mesures de dissimilarité dans SMART:

$D_i = \left(w_{i,k}\right)_k$

avec $w_{i,k} = 0.5 \cdot \left(1 + p_{i,k} / \max_l(p_{i,l})\right) \cdot \log(N / n_k)$ si $p_{i,k} \neq 0$

$w_{i,k} = 0$ sinon

où $w_{i,k}$ est le poids du terme T_k dans le document D_i, $p_{i,k}$ est la fréquence relative de T_k dans D_i, N représente le nombre total de documents dans la base documentaire et n_k le nombre de documents contenant le terme T_k.

et

$\text{atn}(D_i, D_j) = D_i \bullet D_j$ (dissimilarité SMART atn), où $\bullet$ représente le produit scalaire

et

$\text{atc}(D_i, D_j) = \cos(D_i, D_j)$ (dissimilarité SMART atc).

Propriétés des dissimilarités à base de cosinus :

Les dissimilarités atc et atn vérifient les propriétés suivantes :

(3) si $\max(D_1) = \max(D_1 / D_2)$ et $\max(D_2) = \max(D_2 / D_1)$,

alors $\text{atn}(D_1, D_2) = \text{atn}(D_1 / D_2, D_2 / D_1)$

La propriété (3) signifie que, sous les conditions portant sur les maxima, la dissimilarité atn n'est sensible qu'aux parties communes (i.e. les parties partagées par les profils lexicaux) des entités textuelles comparées. La dissimilarité atn est de ce fait bien adaptée pour le calcul de similarités dans les cas où les similarités entre parties de documents sont suffisantes pour entraîner les similarités entre les documents pris dans leur ensemble.

En termes informels, la dissimilarité atn est sensible au « nombre » de mots communs entre les documents comparés.

Notons cependant que cette propriété n'est vérifiée que si $\max(D_1 \setminus D_2)$ (resp. $\max(D_2 \setminus D_1)$), i.e. la coordonnée maximale "hors intersection" est inférieure à $\max(D_1/D_2)$ (resp. $\max(D_2/D_1)$), la coordonnée maximale "dans l'intersection". Ceci signifie que la dissimilarité possède une sensibilité aux coordonnées "hors intersection",

mais limitée à l'intégration des valeurs maximales.

Dans le domaine de la recherche documentaire, la dissimilarité atn peut être utilisée pour rechercher de l'information « à l'intérieur » des documents, i.e. de l'information ne correspondant qu'à une partie, une phrase par exemple, de ces derniers. Dans ce type de situation en effet, des similarités sensibles à la « proportion » de termes communs (comme la dissimilarité atc par exemple) ne sont pas adéquates.

(4) $\quad \text{atc}(D_1, D_2) = \text{atn}(D_1, D_2) / \left(|D_1| \cdot |D_2| \right)$

En termes informels, la propriété (4) indique que la dissimilarité atc est sensible à la « proportion » de mots communs dans les documents comparés.

2.3 La « distance » de Kullback-Leibler pour les applications de Text Mining

Dans le domaine du Text Mining, les distributions de co-occurrences de mots-clés peuvent être utilisées comme support pour l'exploration de grandes collections de documents (Feldman, 1995). Pour ce type de tâche, une mesure est nécessaire pour quantifier le degré d'intérêt d'une distribution observée par rapport à un modèle donné. Un choix possible est la mesure d'entropie relative (ou « distance » de Kullback-Leibler) qui quantifie le degré de « surprise » associé à l'observation d'une distribution p alors qu'une distribution q était attendue :

$$\text{KL}_0(D_i, D_j) = \sum_{k | p_{i,k} \cdot p_{j,k} \neq 0} p_{i,k} \cdot \log(p_{i,k} / p_{j,k})$$

Pour faciliter la comparaison avec les autres mesures de similarité analysées, nous utiliserons ci-après une version symétrisée de la distance de Kullback-Leibler :

$$\text{KL}(D_i, D_j) = \sum_{k | p_{i,k} \cdot p_{j,k} \neq 0} \left(\left(p_{i,k} - p_{j,k} \right) \cdot \left(\log(p_{i,k}) - \log(p_{j,k}) \right) \right)$$

Notons que le terme de « distance » n'est pas strictement correct car la « distance » KL ne vérifie par l'inégalité triangulaire. Pour cette raison, en toute rigueur, la version symétrisée de la distance KL n'est qu'une dissimilarité.

Propriétés de la version symétrisée de la « distance » KL :

(5) $\quad \text{KL}(D_1, D_2) = \text{KL}(D_1 / D_2, D_2 / D_1)$

Comme, par définition, le distance KL est une somme de termes positifs calculés uniquement pour les paires de coordonnées $(p_{i,k}, p_{j,k})$ strictement positives, elle correspond de ce fait à une dissimilarité exclusivement sensible aux différences « dans l'intersection ». Comme la dissimilarité atn, elle peut donc être utilisée pour comparer des représentations de tailles sensiblement différentes.

2.2 Quelques propriétés générales des différentes mesures de similarité :

Une première propriété importante de toutes les mesures de similarité mentionnées ci-dessus est qu'elles sont toutes définies en termes de fréquences relatives en non pas en

termes de fréquences absolues. Ceci entraîne la propriété désirable de garantir qu'un document composite obtenu par concaténation d'un nombre quelconque de copies d'un même document élémentaire sera, pour ce qui est des similarités, strictement équivalent au document élémentaire lui-même.

Une seconde propriété caractéristique des différentes similarités (ou dissimilarités) décrites ici est que toutes, sauf la distance KL, sont associées à une pondération des dimensions de l'espace de représentation (par le biais du facteur $\log(N/n_k)$ pour les dissimilarités utilisées en recherche documentaire ou du facteur $1/\sqrt{f_{.j}}$ pour la distance du chi-deux en analyse des données textuelles). De telles pondérations peuvent être vues comme des procédures de normalisation dont l'objectif est d'intégrer, dans l'évaluation des similarités, la notion de « pouvoir de discrimination » sélectivement associé avec les différentes dimensions de l'espace de représentation. L'idée sous-jacente est qu'une dimension qui « sélectionne » un nombre important d'entités textuelles (par exemple une dimension pour laquelle un nombre important d'entités ont des coordonnées strictement positives) est également une dimension qui discrimine peu au sein de l'ensemble des entités textuelles et donc une dimension dont le poids devra être réduit lors de l'évaluation des similarités. Ceci explique par exemple l'utilisation de facteurs de pondération inversement proportionnels à la fréquence en document $1/n_k$.

3. Améliorer les similarités : graphes sémantiques et analyse de contiguïté

Une analyse de similarités appliquée à une matrice de profils peut mener à des résultats décevants voire trompeurs.

1. La matrice des profils peut en effet être extrêmement creuse : beaucoup de lignes n'auront alors aucun élément commun.
2. Une quantité non négligeable d'information additionnelle (meta-information) est souvent disponible (relations syntaxiques, réseaux sémantiques, corpus externes, thesaurus, …) et l'on doit alors envisager de la mettre à profit.

Ainsi, pour rendre les dissimilarités entre profils lexicaux plus signifiantes, il peut être utile de prendre en compte des informations sémantiques sur les unités textuelles et, en particuliers, sur les similarités (sémantiques) entre ces unités.

Bien que l'on ne dispose d'aucune règle universelle permettant d'établir si deux mots sont sémantiquement équivalents, il est habituellement reconnu (Lewis et Croft, 1990) que les co-occurrences (i.e. les mots avec lesquels un mot apparait au sein d'une même phrase) sont des éléments pertinents pour la détermination du sens (les mots se désambiguïsent les uns les autres). Pour prendre en compte des relations de co-occurrence, un « graphe sémantique » peut être utilisé. Les nœuds d'un tel graphe sont les différentes unités textuelles (les « mots ») et ses arcs (non orientés) sont pondérés en fonction d'un indice d'intensité de co-occurrence entre les unités correspondantes. Un graphe sémantique est ainsi complètement décrit par une matrice de poids **M**, d'ordre (p, p) où p est le nombre total d'unités distinctes.

Un graphe sémantique peut être construit à partir d'une source externe d'information (un dictionnaire de synonymes ou un thésaurus par exemple) ou dérivé des associations effectivement observées dans un corpus. Le corpus utilisé peut être un corpus extérieur ou même la collection de documents sur laquelle porte l'analyse. Dans ce dernier cas, les similarités entre deux unités pourront être dérivées à partir des proximités entre leurs profils lexicaux dans la collection.

Si un graphe sémantique est disponible, les techniques développées dans le cadre de l'analyse de données textuelles peuvent être adaptées par le biais du calcul d'un nouvel indice de similarité entre textes (Becue, 1996).

Ainsi, dans le cas d'un graphe sémantique externe (i.e. dérivé à partir d'informations autres que le corpus analysé lui-même), une façon simple pour prendre en compte les voisinages sémantiques est de remplacer la matrice des profils **T** par la nouvelle matrice **T(I + αM)**, où **I** est la matrice identité, **M** la matrice des poids définissant le graphe sémantique et α un paramètre numérique permettant de calibrer l'importance accordée aux voisinages sémantiques. Cette approche revient à munir l'espace de représentation de dimension p d'une nouvelle métrique définie par $\mathbf{(I + \alpha M)^2}$ ce qui mène de façon immédiate à un nouvel indice de similarité qui peut être utilisé pour le calcul des similarités textuelles. Du fait de la taille des tables de données manipulées, les calculs effectifs sont souvent réalisés à l'aide des coordonnées sur les premiers axes principaux.

Si un graphe sémantique interne (i.e. dérivé à partir des co-occurrences dans le corpus analysé lui-même) est utilisé, une matrice des poids possible est la matrice **M = C - I**, où **C** est la matrice des corrélations entre mots (ce qui permet des poids négatifs correspondant à des intensités de co-occurrence négatives). Cependant, il est à noter que, si les colonnes de **C** sont de variance unité, l'effet de la nouvelle métrique se réduit à une simple re-pondération des axes principaux (l'importance relative des premières valeurs propres est fortement augmentée) lors du calcul des distances. De telles propriétés contribuent à souligner le rôle primordial des premiers axes principaux dans le calcul des similarités définies pour analyse de données textuelles.

Dans le domaine de la recherche documentaire, différentes techniques à base de co-occurrences ont également été utilisées, comme par exemple les vecteurs de co-occurrence moyens de l'approche « sémantique distributionnelle » proposée dans (Rajman, Rungsawang, 1995) ou les vecteurs de contexte dans (Schütze, 1992). Plusieurs de ces approches s'appuient sur des outils d'analyse factorielle très similaires à ceux proposés par Benzecri (1977) ou par Lebart (1982) pour le traitement de matrices creuses de grandes dimensions. Par exemple, Furnas et al. (1988) et Deerwester et al (1990) suggèrent, sous le nom de « Latent Semantic Indexing », une approche dans laquelle l'hypothèse fondamentale est que les relations terme/document, implicitement représentées dans la matrice des profils, sont en fait obscurcies par les phénomènes de variabilité lexicale (Berry, 1996) et que la matrice des profils doit donc être traitée par le biais d'une décomposition en valeurs singulières (SVD) qui permet de

remplacer les profils lexicaux par les coordonnées des documents dans le sous-espace engendré par les k premiers vecteurs principaux produits par la SVD. Cette nouvelle représentation a l'avantage d'encoder les relations d'association entre mots et documents d'une façon qui repose exclusivement sur les mots : deux documents pourront alors être proches dans le sous-espace engendré même s'ils ne possèdent aucun mot commun.

Ces méthodes sont en fait très similaires à celles utilisées dans le cas de l'analyse discriminante réalisée sur le premiers axes factoriels d'une analyse des correspondances (Bartell, 1992), ce qui n'est en fait guère surprenant puisque la décomposition en valeurs singulières constitue une base commune pour l'analyse des correspondances et l'analyse en composantes principales.

5. Conclusion

L'analyse des similarités entre entités textuelles dans des espaces de représentation de haute dimensionnalité constitue un domaine actif de recherche. Dans cette contribution, nous avons présenté quelques unes des mesures de similarité usuellement utilisées dans des domaines comme l'analyse des données textuelles, la recherche documentaire ou le Text Mining. Nous avons ensuite analysé certaines des propriétés générales de ces mesures et présenté quelques méthodes permettant l'amélioration des similarités obtenues. Les résultats mentionnés constituent une première étape dans le processus complexe de la définition de mesures de similarité bien adaptées pour les différentes applications opérant sur des données textuelles et devront être articulés avec les travaux de recherche en cours portant sur les différentes techniques de validation des mesures de similarité (visualisation, catégorisation, …).

References

Bartell B.T., Cottrell G.W., Belew R.K. (1992) - Latent semantic indexing is an optimal special case of multidimensional scaling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al. Ed., 161-167, ACM Press, New York.

Becue Bertaut M., Lebart L. (1996) Clustering of Texts using Semantic Graphs. Application to Open-ended Questions in Surveys, Proceeding of the IFCS 96 Symposium, Kobe.

Becue, M., Peiro, R. (1993): Les quasi-segments pour une classification automatique des réponses ouvertes, in *Actes des 2ndes Journées Internationales d'analyse des données textuelles*, (Montpellier), 310-325, ENST, Paris.

Benzecri J.-P.(1977) - Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Données*, II, n °4, 369-406.

Berry M. W. (1996) - Low-Rank Orthogonal Decompositions for Information Retrieval Applications, Numerical Linear Algebra with Applications, vol 1(1), 1-27.

Cover T. M., Thomas J. A. (1991) - *Elements of Information Theory,* John Wiley and Sons, 1991.

Daille B. (1994) – Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, Las Cruces, 1994.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990) - Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6), 391-407.

Escofier B. (1978) - Analyse factorielle et distances répondant au principe d'équivalence distributionnelle, *Revue de Statist. Appl.*, vol. 26, n°4, 29-37.

Feldman R. and Dagan I. (1995). KDT – Knowledge Discovery in texts*, Proceedings of the 1st Int. Conf. On Knowledge Discovery (KDD-95)*, Aug., 1995.

Furnas G. W., Deerwester S., Dumais S.T., Landauer T.K., Harshman R. A., Streeter L.A., Lochbaum K.E. (1988) - Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, 465-480.

Gifi A. (1990) - *Non Linear Multivariate Analysis*, Wiley, Chichester.

Greenacre M.(1984) - *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Lebart L. (1982) - Exploratory analysis of large sparse matrices, with application to textual data, *COMPSTAT,* Physica Verlag, 67-76.

Rajman M., Rungsawang A. (1995) - "Textual Information Retrieval based on the Concept of Distributional Semantics", *3rd International Conference on Statistical Analysis of Textual Data (JADT'95)*, Rome, Bolasco S. et al., (eds), p 151-162

Salem A. (1984) - La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, *Les Cahiers de l'Analyse des Données,* 9, n° 4, 489-500.

Salton G. (1988) - *Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer,* Addison-Wesley, New York.

Salton G., Mc Gill M.J. (1983) - *Introduction to Modern Information Retrieval,* International Student Edition.

Salton G. and Buckley C. (1988) - Term Weighting Approaches, in *Automatic Text Retrieval, Information Processing and Management,* 24:5, 513-523.

Schütze H.(1992a) - Context Space, Working Notes of the AAAI *Fall Symposium on Probabilistic Approaches to Natural Language*, 1992