Vérification sémantique de liens hypertextes avec LSA

Philippe Dessus

Dessus, P. (1999). Vérification sémantique de liens hypertextes avec LSA. In J.-P. Balpe, A. Lelu, S. Natkin, I. Saleh (Eds). *Hypertextes, hypermédias et internet (H2PTM'99)*. Paris : Hermès, 119-129.

Vérification sémantique de liens hypertextes avec LSA

Philippe Dessus

Laboratoire des sciences de l'éducation,
Bât SHM, 1251 av. Centrale, BP 47,
Université Pierre-Mendès-France
38040 GRENOBLE CEDEX
Philippe.Dessus@upmf-grenoble.fr

RESUME : L'objet de cet article est de vérifier si une méthode d'analyse factorielle multidimensionnelle issue de la recherche automatique de documents prédit de manière adéquate les liens hypertextes de deux types de bases textuelles sur l'astronomie : un dictionnaire encyclopédique et un ensemble de cours sur Internet. Notre première étude a montré que la relation de proximité sémantique entre deux termes (une entrée et une cible) liés hypertextuellement dans le dictionnaire croît avec la typicalité de l'entrée. À un niveau plus général, LSA prédit correctement l'ordonnance de documents WWW formant un cours. Il pourrait donc être utilisé à cette fin dans un logiciel générant des documents didactiques qui reste à implémenter. L'avantage de LSA est qu'il ne requiert pas, comme la plupart des systèmes issus de la recherche documentaire, d'index « manuel » dont la qualité est trop dépendante de la personne qui le construit.

ABSTRACT : The purpose of this paper is to test whether LSA, a factorial analysis IR method, predicts hypertext links of two text corpora. Two documents were used: a hypertext dictionary and an Internet course, both in astronomy. A first study shows that given two linked terms (an entry and its target), the more the entry is typical, the more both terms are semantically close. In a second study we show that LSA predicts the WWW course structure. LSA does not require complex human pre-processing and could be used to automatically create hypertext courses.

*MOTS-CLES : Génération automatique de liens, LSA (*latent semantic analysis*), Recherche automatique de documents, Documents WWW, Encyclopédie sur CD-ROM.*

KEY WORDS : Automatic Hypertext Links Processing, LSA (Latent Semantic Analysis), Information Retrieval, Web documents, CD-ROM encyclopedia.

1. Introduction

L'objet de cet article est de vérifier si une méthode d'analyse factorielle multidimensionnelle issue de la recherche automatique de documents prédit de manière adéquate les liens hypertextes de deux types de bases textuelles sur l'astronomie : un dictionnaire encyclopédique et un ensemble de cours sur Internet. LSA (*latent semantic analysis*), le modèle utilisé ici, rend compte assez fidèlement de l'apprentissage et de l'utilisation du langage humain, pour peu qu'on lui ait fait traiter, au préalable, de grands *corpora* [LAN 97]. Il a également été testé avec succès dans l'évaluation de la cohérence textuelle [FOL 98]. Pour ces raisons, nous allons ici tester ses capacités pour révéler la nature sémantique de liens hypertextes dans de grands *corpora* textuels, en vérifiant les relations de synonymie entre deux termes liés hypertextuellement (une entrée et une cible), ainsi que la grandeur de l'effet de l'introduction d'un document de cours dans une base textuelle du même domaine. Si notre test s'avère concluant, LSA pourrait être utilisé comme aide à la génération de liens hypertextes ou à l'ordonnance de documents de cours, sans requérir préalablement de formalisation sémantique du domaine de la part du concepteur.

On peut distinguer trois approches visant à hypertextualiser automatiquement un document [BAL 96] :

— *l'approche structurelle* qui permet une conversion automatique de textes en hypertexte, une fois que la structure du document a été définie par l'auteur [WEN 96]. Si la navigation dans un ensemble de liens hypertextes induit une charge mentale importante pour le lecteur [EDW 89], il en est de même pour l'auteur à propos de la conception de ces liens [1]. La conversion de plus larges *corpora* de textes devient donc encore plus longue et problématique [2] ;

— *l'approche linguistique* qui permet, à l'aide d'une indexation préalable (au niveau syntaxique), de proposer des mots-clés que le lecteur pourra suivre ;

— *l'approche statistique* où l'on construit, par analyse factorielle, un espace vectoriel multidimensionnel où les documents et/ou termes sont représentés d'après leurs co-occurrences. Des requêtes [CUN 97] ou des cartes [CHE 97] permettent à l'utilisateur de se repérer au sein de cet espace vectoriel et de récupérer les documents correspondants.

L'approche que nous allons développer ici est de type statistique, car LSA manipule un espace vectoriel multidimensionnel, sur lequel il réalise une forme d'analyse factorielle. La différence avec les autres modèles présentés plus haut est qu'il réalise sur les données textuelles un mécanisme inductif qui rapproche sémantiquement des termes co-occurrents, ou des termes proches de termes co-occurrents. Ainsi, aucune connaissance du domaine, formalisée par un humain, n'est nécessaire au préalable (par exemple des *thesauri*). Après avoir décrit plus en détail le modèle de LSA ainsi que quelques-unes de ses validations, nous tenterons de répondre à deux questions : 1) LSA peut-il prédire les liens hypertextes d'un

[1] Par exemple, HEFTI [CHI 91] demande une journée pour convertir un livre de 400 pages sous forme hypertextuelle.

[2] Balpe *et al.* [BAL 96] donnent quelques raisons à cela : les *corpora* et index sont volumineux, les données sont « diluées », c'est-à-dire que chaque unité-document n'utilise qu'une très faible proportion du vocabulaire et sont de type *pick-any*, c'est-à-dire que le choix des termes descripteurs pour l'indexation dépend fortement de la personne qui la réalise.

dictionnaire encyclopédique ? ; 2) LSA peut-il prédire l'ordonnance d'une série de documents WWW séquentiels ?

2. LSA, une méthode d'analyse factorielle multidimensionnelle

2.1. *Description du modèle*

LSA [3], pour *latent semantic analysis* (analyse sémantique latente) est un modèle statistique, fondé sur un type d'analyse factorielle [4], permettant d'analyser la proximité sémantique à l'intérieur d'un grand ensemble d'unités d'informations textuelles. Initialement, LSA a été conçu pour améliorer l'efficacité de l'interrogation de systèmes documentaires informatisés, la plupart du temps fondés sur un appariement lexical plutôt que sémantique. Le modèle de LSA suppose que, étant donné plusieurs « contextes » (unités d'information textuelle, soit phrases, paragraphes, discours…), il existe une *structure latente* dans l'utilisation des mots communs à ces contextes et qu'une analyse statistique permet de mettre en évidence cette structure. Le modèle de LSA pose que la similarité sémantique de deux mots est liée à la probabilité que deux mots se retrouvent dans le même contexte, ou dans deux contextes différents dans lesquels apparaissent des mêmes mots. En d'autres termes, LSA tient compte des différents contextes dans lesquels apparaissent les mots [5] et il considère aussi comme proches deux mots n'apparaissant jamais dans le même contexte, mais dont les contextes respectifs contiennent des mots similaires. LSA permet deux types de calculs : celui de la *centralité sémantique* globale de termes, un terme étant d'autant plus central, donc typique, au sein du corpus qu'il est souvent associé à d'autres termes de ce corpus ; le calcul de la *proximité sémantique* entre deux termes (ou un terme et un contexte, deux contextes) donne un indice d'autant plus élevé que ces deux entités sont de sens voisin ou bien ont été fréquemment associées [6]. Décrivons plus précisément quelques études validant les performances de LSA.

---

[3] LSA est écrit en langage C et fonctionne sur une station de travail Unix, il est déposé en 1990 par *Bell Communications Research Inc*. Le lecteur trouvera, sur le site Internet http://lsa.colorado.edu un grand nombre d'informations sur LSA par leurs auteurs, ainsi qu'une version de LSA interrogeable à distance.

[4] Nous évitons délibérément la description mathématique du modèle utilisé par LSA. Brièvement, voici comment le logiciel traite les données. À partir de la table de contigence rassemblant les occurrences, par document, des mots apparaissant au moins deux fois, LSA réalise une décomposition aux valeurs singulières de cette matrice, puis « filtre » les cent dimensions les plus significatives. Chaque mot et résumé se trouve représenté par un vecteur dans cet espace. Pour plus de précisions, on se reportera à Landauer et Dumais [LAN 97] ou à Deerwester *et al.* [DEE 90].

[5] LSA ne tient en revanche pas compte de la syntaxe, c'est-à-dire de l'ordre dans lequel sont écrits les mots.

[6] Ainsi, deux termes ayant une forte proximité sémantique d'après LSA ne sont pas forcément des synonymes, mais peuvent être des termes souvent associés. Ce problème apparaît notamment lorsque la base textuelle traitée par LSA est trop peu importante.

2.2. *Revue de quelques validations de LSA*

De nombreux travaux ont été réalisés pour tester la validité de LSA, on se reportera à Landauer et Dumais [LAN 97] pour une vue complète. Exposons quelques résultats en rapport avec notre travail. Dumais [DUM 91] teste LSA avec de larges bases de données (médicale, d'aéronautique, de magazine). On peut recueillir, à partir de requêtes dans chaque base, les documents ayant une similarité maximale avec chaque requête. Les résultats montrent que les documents recueillis par LSA sont 20 % plus pertinents que ceux recueillis par une traditionnelle requête par mots-clés. Landauer *et al.* [LAN 93] ont incorporé LSI — *latent semantic indexing*, une version antérieure de LSA dédiée à la recherche automatique de documents — à *SuperBook*, un logiciel hypertexte de navigation dans de larges bases de données textuelles ou imagées (manuels, romans). Une comparaison de recherches sur document papier *vs* LSI montre qu'avec ce dernier, les recherches sont significativement plus précises lorsque la requête ne mentionne que des mots présents dans le corps du texte ou qu'elle est formulée avec des synonymes ; les performances sont en revanche similaires lorsque la requête contient des mots présents dans les titres du texte. Ces deux études, même si elles font intervenir LSA comme moteur de recherche par mots-clés, montrent que LSA pourrait être également utilisé pour générer des liens hypertextes, les termes recueillis par les requêtes étant sémantiquement pertinents.

Deux autres études, l'une sur la mesure de la cohérence textuelle et l'autre sur les hypertextes nous incitent à penser cela. Foltz *et al.* [FOL 93] cités par Foltz [FOL 96], afin de mesurer la cohérence textuelle de productions écrites, ont tout d'abord « entraîné » LSA avec vingt et un articles d'un même thème. Ils ont ensuite pris quatre autres textes du même domaine, dans lesquels ils ont fait varier intentionnellement leur cohérence locale ainsi que leur macrocohérence. Ils ont calculé les proximités sémantiques des phrases de chacun des textes, prises deux à deux, afin d'obtenir une moyenne de chevauchements sémantiques pour chaque texte. Les résultats montrent que cette moyenne croît avec la cohérence attribuée aux textes. Les prédictions de cohérence textuelle calculées par LSA augmentent donc bien dans le sens attendu. Blustein et Webber [BLU 95] vérifient la qualité des liens hypertextes d'un corpus de messages d'un groupe de discussion en comparant la similarité sémantique, calculée par LSI, entre deux documents, au nombre minimal de liens hypertextes qui les séparent. Ils trouvent que la corrélation entre ces deux mesures est élevée (r = .73) lorsque les deux documents sont directement liés l'un à l'autre, sans nœud intermédiaire.

Ces résultats montrent que l'on peut, avec LSA, effectuer des requêtes, obtenir une mesure de la cohérence textuelle et vérifier la nature sémantique de liens hypertextes. Ces résultats sont obtenus alors que LSA travaille à partir d'un large ensemble de textes, sans aucune connaissance formalisée au préalable, avec des performances satisfaisantes et parfois même voisines de celles obtenues par des humains. Ces résultats nous incitent à mettre en place deux études, la première vérifie la nature sémantique des liens hypertextes au sein d'un dictionnaire électronique ; la deuxième vérifie l'éventuel lien sémantique entre deux documents d'un cours, en se basant sur les capacités de LSA en termes de calculs de macrocohérence.

3. LSA peut-il prédire les liens hypertextes d'un dictionnaire encyclopédique ?

Il est ici question de vérifier dans quelle mesure LSA peut prédire les liens hypertextes d'un dictionnaire encyclopédique sur l'astronomie. Soit une entrée de ce dictionnaire comportant, dans sa définition, un lien hypertexte vers une autre entrée, nous appelons « cible » un tel lien (*voir figure 1*). Notre hypothèse est qu'il existe une liaison positive [7] entre la centralité des entrées, calculée par LSA, et la moyenne de proximité entre ces dernières et les cibles figurant dans leur texte de définition. Autrement dit, plus une entrée est typique (centrale) au sein du dictionnaire, moins les termes utilisés pour sa définition et liés hypertextuellement depuis cette dernière entretiendront avec elle des rapports de proximité sémantique ; à l'inverse, les termes atypiques nécessiteront une définition avec des cibles sémantiquement proches. Cette explication renvoie à la théorie de l'information [COO 75] : un message (une définition) délivre d'autant moins d'informations qu'il est plus probable. Ainsi, en ce qui concerne les termes les plus typiques, leurs définitions devront, pour être informantes, renvoyer à des cibles moins probables, donc entretenant une faible proximité sémantique. À l'inverse, les termes moins typiques devront avoir des définitions plus informantes, car ils sont censés être moins connus. Les cibles contenues dans leurs définitions devront donc être sémantiquement proches [8]. Cette hypothèse nous permet de vérifier si les liens hypertextes obéissent à une logique sémantique, selon la centralité des entrées.

absolute luminosity A measure of the actual rate of energy output of a star or other celestial object as opposed to the apparent luminosity, which depends on the distance to the object.
apparent luminosity The luminosity of a star or other astronomical object as it appears to an observer on Earth. The apparent luminosity depends on both the actual energy output of the object and its distance.
luminosity (symbol L) The energy radiated per unit time by a luminous body. See also: magnitude.

Figure 1. *Exemples d'entrées (en gras) et de cibles (en souligné) tirés du* Dictionnaire Penguin de l'astronomie *(ce n'est pas une copie d'écran).*

Nous avons utilisé la version anglaise électronique du *Dictionnaire Penguin de l'astronomie* [11] fournie avec le logiciel *Redshift* v. 1.2 de *Maris Multimedia* (*voir figure 1 ci-dessus pour un extrait*). Le corpus a une taille de 1 Mo, 3 000 liens hypertextes, 2 000 entrées. Ce dictionnaire, probablement tiré d'une version sur support papier où les liens hypertextes sont remplacés par des corrélats, peut être

[7] La relation est positive car le score de centralité est inversement proportionnel à la valeur qu'il mesure.
[8] À titre d'exemple, on peut vérifier que « *sun* » (terme très central) ne renvoie pas à « *star* », terme avec lequel il entretient pourtant une grande proximité sémantique.
[11] Ainsi, chaque entrée du dictionnaire a été traitée par LSA comme un terme, même si elle comporte plusieurs mots. Par exemple, « *main sequence star* » est traité comme « *main-sequence-star* ».

raisonnablement considéré comme ayant des liens hypertextes vérifiés et valides sémantiquement. Ainsi, il nous importe de vérifier si LSA peut les prédire.

Voici la procédure suivie pour cela :

— nous avons traité le dictionnaire avec LSA et récupéré arbitrairement les 2 000 termes les plus centraux (soit environ le tiers des 5 187 termes traités par LSA) ; de ces termes ont été retirés ceux ne faisant pas l'objet d'une entrée du dictionnaire, ce qui laisse 256 entrées parmi les 2 000 termes les plus centraux ;

— nous avons ensuite calculé, pour chacune de ces entrées restantes, la moyenne des proximités sémantiques de ces dernières avec leurs cibles (soit 630 cibles au total, *voir Tableau I*) [12].

Tableau 1. *Les trois entrées les plus centrales (première colonne), leurs cibles correspondantes et les calculs de proximité effectués. Nous examinons la corrélation entre les indices de centralité (première colonne) et la moyenne des proximités entre chaque entrée et ses cibles (quatrième colonne).*

Entrées les plus centrales (indice de centralité)	Cibles associées (par ordre alphabétique)	Indice de proximité entre entrée et cible calculé par LSA	Moyenne des proximités pour chaque entrée
constellation (0,270)	declination	0,175	0,144
—	right-ascension	0,114	
star (0,294)	main-sequence-star	0,693	0,693
sun (0,298)	aurora	0,067	0,298
—	chromosphere	0,427	
—	corona	0,361	
—	flare	0,348	
—	geomagnetic-storm	0,212	
—	granulation	0,284	
—	photosphere	0,489	
—	prominence	0,346	
—	solar-activity	0,360	
—	spectral-type	0,154	
—	spicules	0,304	
—	sunspot	0,227	

Les résultats montrent une liaison significative, bien que très faible, entre la centralité des termes et la moyenne des proximités (N = 256 ; *r* unilatéral = 0,12 ; $p < 0,05$). Les indices de centralité des entrées calculés par LSA sont donc faiblement mais significativement liés aux moyennes de proximité des cibles correspondant aux entrées. Ainsi, moins une entrée est centrale au sein du dictionnaire, plus elle a une proximité sémantique importante avec les cibles qui composent sa définition. Ce résultat, conforme à notre hypothèse, permet d'avancer que LSA pourrait être une aide à la vérification de liens hypertextes puisque, selon la centralité des termes, on pourrait calculer la proximité sémantique des cibles adéquates.

Une deuxième étude permet de nous situer à un niveau plus élevé : la macrocohérence, en vérifiant si LSA possède des capacités d'ordonnancement de

[12] Étant entendu que si toute cible est nécessairement une entrée, la réciproque n'est pas toujours avérée. Nous n'avons pas calculé la proximité des entrées avec les termes associés sans lien hypertexte, de type « *voir aussi…* ».

textes d'un même domaine, pour peu qu'on lui donne une base textuelle sur ce domaine.

4. LSA peut-il prédire l'ordonnance d'une série de documents hypertextes ?

Il est maintenant question de vérifier dans quelle mesure LSA peut prédire l'ordonnance de documents tirés d'un cours d'astronomie [RYD 97]. Ce cours est composé de 43 documents, chacun d'environ 10 000 caractères [13]. Les liens hypertextes entre ces documents sont beaucoup moins nombreux et moins sémantiquement riches que ceux de notre étude précédente — il s'agit d'une table des matières où chaque item forme un lien vers un document, forme habituelle des documents WWW à deux niveaux de hiérarchie. Un outil d'aide à une telle ordonnance peut toutefois être utile lorsque la base de textes sur laquelle on travaille est importante.

L'idée générale est de mesurer l'impact de l'introduction d'un texte à ordonner, dans un espace vectoriel comprenant les termes du dictionnaire encyclopédique suscité (*corpus* initial), sur les centralités de tous les termes traités par LSA. Notre hypothèse est que plus cet impact est important, plus le document introduit est riche sémantiquement, par conséquent, il doit suivre le ou les document(s) dont l'impact a été plus faible. À l'inverse, les documents plus pauvres sémantiquement, donc de faible impact sur le *corpus* initial, seront plutôt placés dans les premiers. Ainsi, nous devrions observer une liaison positive entre la variable du rang des documents (de 1 à 43) et la variable « impact » de l'introduction du document au sein du *corpus* initial.

Voici la procédure suivie :

— nous avons construit, avec LSA, l'espace vectoriel initial, à partir du dictionnaire de l'astronomie décrit dans la première étude. Les centralités des termes traités par LSA (soit environ 5 200 termes) sont récupérées ;

— ensuite, nous avons concaténé successivement et indépendamment chaque document du cours au corpus initial et construit avec LSA l'espace vectoriel correspondant. Les centralités des termes traités par LSA lors de ces traitements successifs ont été récupérées [14] ;

— nous avons enfin calculé, pour chaque terme de chaque cours, la différence de ses centralités au sein des deux espaces vectoriels (celui du corpus initial et celui de chaque cours concaténé au corpus initial). On a ainsi, en sommant tous les écarts, une information sur l'impact qu'a chaque document sur la centralité de tous les termes traités par LSA.

[13] Pour les besoins de l'étude, nous avons n'avons traité que les 3 000 premiers caractères de chaque document, car la taille des documents interagissait fortement avec les variables recueillies. Cette méthode a également été suivie par Foltz dans une étude sur la cohérence textuelle calculée par LSA [FOL 98]. Comme le premier document était d'une taille inférieure à ce seuil, nous ne l'avons pas pris en compte dans nos calculs.

[14] Arbitrairement, la centralité d'un terme non pris en compte dans le corpus initial est de 1, soit une centralité minimale.

Tableau 2. *Les plus grandes évolutions de centralités au sein du document 2. Le calcul effectué ensuite par document est une somme de tous les écarts de centralité (dernière ligne) des 5 200 termes traités par LSA.*

	lambda	wave	electro-magnetic	wave-lenght
Centralité au sein du corpus initial	0,796	0,668	0,502	0,469
Centralité au sein du corpus initial concaténé au document 2	0,829	0,684	0,511	0,474
Écart de centralité	0,033	0,016	0,009	0,005

Le résultat [15] montre une liaison moyenne entre les deux variables — rang du document et son « impact » dans le *corpus* initial (ρ unilatéral = .397 ; N = 43 ; $p < 0{,}005$). Cette relation, très significative, nous montre que des documents séquentiels de cours entretiennent avec leurs précédesseurs et leur successeurs des relations que LSA peut déterminer, ce que Foltz, cité plus haut, avait montré pour d'autres types de documents [FOL 96]. Plus précisément, l'impact de l'introduction d'un document dans une base textuelle du domaine, analysée par LSA est proportionnel au rang de ce document. Tout se passe comme si la force de cet impact rendait compte de la richesse sémantique du document. Concernant l'ordonnance de documents, l'auteur du cours utilisé ici a placé, conformément à notre hypothèse, les documents selon un impact croissant dans le corpus textuel initial. Bien évidemment, d'autres travaux sont nécessaires afin de vérifier ce résultat dans d'autres domaines. Pour le cours considéré, LSA peut être un outil d'aide à son ordonnance, activité difficile dès que le corpus a une taille importante.

5. Discussion

Le but de cet article est de vérifier dans quelle mesure LSA est une méthode permettant d'aider à construire des liens hypertextes au sein d'un dictionnaire ou d'un ensemble de documents formant un cours. Nous avons tout d'abord montré une liaison significative, bien que faible, entre la centralité d'entrées et la moyenne des proximités de leurs termes-cibles. Ensuite, à un niveau plus général, LSA prédit correctement l'ordonnance de documents WWW formant un cours. En vérifiant de telles relations, nous avons réalisé un travail descriptif et le passage à un logiciel prescrivant des liens hypertextes ne pourra se faire que prudemment, en disposant d'un modèle adéquat de la qualité des liens. Toutefois, un tel logiciel, qui reste à implémenter, pourrait avoir les fonctionnalités suivantes : vérifier *a priori* ou *a posteriori* la validité sémantique de liens hypertextes ; générer un ordre de lecture d'un ensemble de documents didactiques. Nous avons montré que LSA, même s'il ne peut encore prescrire efficacement des liens hypertextes, est une méthode qui rend assez fidèlement compte de la manière dont des auteurs hypertextualisent des documents. L'avantage de LSA est qu'il ne requiert pas pour ces tâches, d'index « manuel » dont la qualité est trop dépendante de la personne qui le construit.

[15] Nous avons converti les sommes en données de rang afin d'utiliser le ρ de Spearman.

Remerciements

Nous remercions Erica de Vries, Benoît Lemaire et Pascal Bressoux pour leurs commentaires d'une version antérieure de ce document. Selon la formule, toute erreur et/ou imprécision qui a subsisté nous est totalement imputable.

Bibliographie

[BAL 96] BALPE, J.-P., LELU, A., PAPY, F. & SALEH, I. (1996). *Techniques avancées pour l'hypertexte*. Paris : Hermès.

[BLU 95] BLUSTEIN, J. & WEBBER, R. E. (1995). Using LSI to evaluate the quality of hypertext links. *In* M. Agosti, J. Allan (Eds). *ACM SIGIR IR and Automatic Construction of Hypermedia : a research workshop*.

[CHE 97] CHEN, C. (1997). Structuring and visualising the WWW by generalised similarity analysis. *Proc. Hypertext '97*. Southampton : ACM.

[CHI 91] CHIGNELL, M. H., HORDHAUSEN, B., VALDEZ, F. & WATERWORTH, J. A. (1991). The HEFTI model of text to hypertext conversion. *Hypermedia,* 3-3, 187-205.

[COO 75] COOMBS, C. H., DAWES, R. M. & TVERSKY, A. (1975). *Psychologie mathématique, T. 2*. Paris : P.U.F.

[CUN 97] CUNLIFFE, D., TAYLOR, C. & TUDHOPE, D. (1997). Query-based navigation in semantically indexed hypermedia. *Proc. Hypertext '97*. Southampton : ACM.

[DEE 90] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41-6, 391-407.

[DUM 91] DUMAIS, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers,* 23-2, 229-236.

[EDW 89] EDWARDS, D. M. & HARDMAN, L. (1989). Lost in Hyperspace : cognitive mapping and navigation in a hypertext environment. *In* R. McAleese (Ed.), *Hypertext : theory into practice.* Norwood : Ablex.

[FOL 93] FOLTZ, P., KINTSCH, W. & LANDAUER, T. K. (1993). An analysis of textual coherence using latent semantic indexing. *Third Annual Conference of the Society for Text and Discourse*. Boulder.

[FOL 96] FOLTZ, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers,* 28-2, 197-202.

[FOL 98] FOLTZ, P. W., KINTSCH, W. & LANDAUER, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25-2-3, 285-307.

[LAN 93] LANDAUER, T., EGAN, D., REMDE, J., LESK, M., LOCHBAUM, C. & KETCHUM, D. (1993). Enhancing the usability of text through computer delivery and formative evaluation : the SuperBook project. *In* C. McKnight, A. Dillon & J. Richardson (Eds), *Hypertext, a psychological perspective*. Chichester : Ellis Horwood, 71-136.

[LAN 97] LANDAUER, T. K. & DUMAIS, S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

[RYD 97] RYDEN, B. (1997). *Introduction to Stellar, Galactic, and Extragalactic Astronomy.* Colombus : Ohio State University (http://www-astronomy.mps.ohio-state.edu/~ryden/ast162.html).

[WEN 96] WENTLAND FORTE, M. (1996). Outils d'aide à la génération automatique d'hypertextes pédagogiques. *In* E. Bruillard, J.-M. Baldner & G.-L. Baron (Eds), *Hypermédias et apprentissages, T. 3.* Paris : INRP, 47-56.