Inclusion de sens dans la représentation de documents textuels : état de l'art

Emmanuel Eckard et Jean-Cédric Chappelier

22 juin 2007

Résumé

Ce document donne un aperçu de l'état de l'art dans le domaine de la représentation du sens dans les documents textuels.

1 Introduction

La chaîne de traitement en language naturel est constituée d'une séquence de processus dont l'importance varie selon la langue à traiter et le but recherché. Ce document se place du point de vue d'une recherche menée au niveau d'une application de recherche documentaire : le texte est supposé déjà déséquentialisé (les termes sont séparés les uns des autres), tokénisé (les formes de surface des termes sont remplacées par leur racine) et indexé (de ce point de vue, processus de bas niveau), et les étapes comme l'évaluation ou la repondération sont considérées comme des post-traitements de plus haut niveau.

Les documents en language naturel sont d'abord convertis en objets mathématiques abstraits utilisables par un ordinateur (phase d'*indexation*). Les données sous forme mathématique sont traitées de façon à produire le résultat désiré : recherche de documents pertinents (recherche d'information), classification de documents supervisée ou non supervisée, etc.

La partie sémantique du traitement est représentée par le couple Représentation-Proximité. C'est à ce niveau que résident les questions relatives à la recherche documentaire. La *représentation* consiste à associer à un document un objet mathématique dans un espace défini a priori.a *proximité* consiste en l'utilisation d'une mesure de similarité sur les représentations ; elle permet de quantifier la proximité sémantique de deux documents.

La recherche documentaire classique a étudié en particulier la phase de proximité, pour trouver des mesures de similarité optimales. Ces recherches ont produit des modèles comme TF-IDF et BM25.

Dans les dix dernières années, l'étape de représentation a été améliorée par l'introduction d'un processus de reconfiguration de la représentation. Ce processus est appliqué a posteriori sur la représentation, et tend à réduire la dimensionalité de l'espace de représentation.

2 Intégration de sens dans la représentation

Les textes en langage naturel ne sont pas compréhensibles tels quels par les systèmes d'information. Il est necessaire d'opérer une conversion en objets mathématiques abstraits, nommée *indexation*(Sebastiani, 2002) : aux documents sont associés des points dans un espace vectoriel de dimensionalité élevée[1], appelé *espace de représentation*. C'est cette étape qui permet d'effectuer une comparaison sémantique plutôt qu'une simple comparaison de la présence des mots dans un texte(Billhardt et al., 2002). L'indexation suppose deux choix : celui de l'ensemble des termes d'indexation (le vocabulaire pris en compte), et celui de la pondération des termes.

Les coordonées dans l'espace de représentation sont fonction de la présence ou non des termes dans un document ; le cas échant, de leur nombre d'occurences dans le document ; et de leur répartition à travers la collection de documents. La représentation repose également sur le choix des unités atomiques de sens dans un texte : c'est le problème de la *sémantique lexicale*. Un autre facteur est l'ensemble des règles de combinaisons de ces unités de sens : c'est le problème de la *sémantique compositionelle*.

En Recherche documentaire, un document est typiquement représenté par un vecteur de *poids* associés à des *termes* (Sebastiani, 2002).

2.1 Choix des termes d'indexation ("*indexing*")

L'indexation se typifie sur la base de deux critères (Lewis, 1992b) :
- Choix des termes : par un humain ou par l'ordinateur
- Nombre de termes : Indéfini (ils peuvent toujours s'additioner) ou déterminé, par avance ou en cours d'indexation

Statistiquement, on obtient un efficacité optimale avec un nombre de termes limité ; des fréquences faibles ; pas de redondances ; et peu de bruit — distortions ou inconsistances dans les poids.

La polysémie, la synonymie et le champs de validité des termes (termes techniques par exemple) compliquent la définition des algorithmes en aval de la représentation (Lewis, 1992b). L'approche naïve revient à les termes et les mots (avec des poids binaires ou non). Cette approche simple est très efficace : tant en IR (Salton and Buckley, 1987) qu'en catégorisation, il a été montré que les représentation plus sophistiquées ne donnent pas nécessairement de meilleurs résultats (Dumais et al., 1998; Cohen, 1995; Siolas and d'Alché Buc, 2000).

On a essayé d'utiliser des parties de texte comme termes d'indexation. La notion de « partie de texte » peut être définie soit *syntaxiquement* (Lewis, 1992a), soit *statistiquement* (ensemble des mots dont le groupement est statistiquement significatif, plutot que partie de texte formant une cohérence sytaxique (Caropreso et al., 2001)). Les résultats, en 2002, n'étaient pas concluants (Sebastiani, 2002), mais les recherches dans cette direction se poursuivent (Metzler and Croft, 2005; Metzler and Croft, 2006; Alvarez et al., 2004).

Dans le même ordre d'idées, la reconnaissance d'*entités textuelles* peut être employée pour raffiner l'indexation.

[1] De l'ordre de plusieurs milliers de dimensions.

2.2 Choix de la pondération

Ces deux pondérations ne tiennent pas compte de la place d'un terme dans un document : c'est le principe du « *bag of words* ».

Booléen : La pondération booléenne attribue un poids binaire de 1 si un terme est présent dans un document, et 0 dans le cas contraire.

Ce style de pondération est souvent associé à un système d'apprentissage non numérique.

`tf-idf :` `tf-idf` est le poids numérique traditionellement utilisé (Soucy and Mineau, 2005). tf désigne la fréquence du terme dans un document donné, et *idf* l'inverse du nombre de document contenant le terme. Il existe un certain nombre de variantes de `tf-idf` (Seydoux et al., 2006).

Les recherche se poursuivent pour trouver des pondérations plus performantes (Soucy and Mineau, 2005).

2.3 Réduction de dimensionalité

Le principe de *réduction de dimensionalité* consiste à réduire le nombre de dimension de l'espace vectoriel qui représente les documents. En classification, la réduction de dimensionalité est nécessaire à certains algorithmes d'apprentissage sophistiqués utilisés pour déterminer des classificateurs, comme l'algorithme LLSF (Yang and Chute, 1994) (cité par (Sebastiani, 2002) p.16). Sur les algorithmes d'apprentissage en général, la réduction de dimensionalité a aussi l'avantage de réduire les problèmes de sur-apprentissage.

Deux catégories de réduction de dimensionalité peuvent être distinguées :

- *Sélection de termes* : l'ensemble d'arrivée (réduit) des termes d'indexation est un sous-ensemble de l'ensemble de départ. Par exemple une sélection des dimensions les plus significatives après une analyse en composantes principales.
- *Extraction de termes* : l'ensemble d'arrivée (réduit) des termes d'indexation n'est pas un sous-ensemble de l'ensemble de départ. Par exemple une « fusion » par combinaison linéaire des dimensions correspondant à un même champs sémantique.

Sélection de termes : La réduction de dimensionalité par sélection de termes vise à extraire le sous-ensemble de l'espace de départ qui concentre en lui la plus grande part du sens du document.

La méthode évidente consiste à filtrer les termes en fonction d'un critère numérique qui en mesure l'importance : la fréquence dans le document, ou des critères plus sophistiqués basés sur la thréorie de l'information :

- facteur association DIA (Fuhr and Buckley, 1991)
- χ^2 (Caropreso et al., 2001; Galavotti et al., 2000; Schétze et al., 1995; Sebastiani et al., 2000; Yang and Pedersen, 1997; Yang and Liu, 1999)
- coefficient NGL (Ng et al., 1997; Ruiz and Srinivasan, 1999)
- gain d'information (Caropreso et al., 2001; Larkey, 1998; Lewis, 1992a; Lewis and Ringuette, 1994; Mladenic, 1998; Moulinier and Ganascia, 1996; Yang and Pedersen, 1997; Yang and Liu, 1999)

– information mutuelle (Dumais et al., 1998; Lam and Ho, 1998; Larkey and Croft, 1996; Lewis and Ringuette, 1994; Li and Jain, 1998; Moulinier and Ganascia, 1996; Ruiz and Srinivasan, 1999; Taira and Haruno, 1999; Yang and Pedersen, 1997)
– *odds ratio* (Caropreso et al., 2001; Mladenic, 1998; Ruiz and Srinivasan, 1999)
– relevancy score (Wiener et al., 1995)
– coefficient GSS (Galavotti et al., 2000)

Extraction de termes : La réduction de dimensionalité par extraction de termes a un effet positif sur l'othogonalité induite par le Modèle Standard, en « fondant » les unes dans les autres des dimensions associées à des mêmes concepts (synonymie). C'est l'idée de base du modèle *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990; "T. K. Landauer and Laham", 1998). L'effet sur la synonymie est discuté dans (Dupret, 2003).

Les méthodes principales sont :
– LSA et ses dérivés
– Le regoupement de termes (*term clustering*). Possible supervisé ou non supervisé.

2.3.1 LSI et PLSI

LSI *Latent Semantic Indexing* et PLSI *Probabilistic Latent Semantic Indexing* sont des techniques qui substituent à la représentation « directe » des documents et des requêtes une représentation de dimension inférieure.

LSI consiste à appliquer une anaylse en composantes principales sur la matrice d'occurrences d'une collection de documents (matrice termes $\times$ documents). Il a été montré (Schétze et al., 1995) que LSA offre des performances au moins égales et parfois nettement supérieures à la méthode de sélection de termes par χ^2. Les perspectives à partir de LSA vont dans le sens d'implémentations sans recours à des matrices — « filtrage polynomial », (Kokiopoulou and Saad, 2004) ; d'optimisations dans le pré-traitement (Tang et al., 2004) ; ou de structures non linéaires (He et al., 2004).

Un développement important est PLSI (Hofmann, 1999), fondé sur un modèle qui postule que les documents sont des réalisations d'un mélange de probabilités conditionelles sur des *sujets*, sur les documents et sur les termes. Les paramètres du molèle sont $P(z)$ (probabilité d'un sujet donné survienne), $P(d|z)$ (probabilité qu'un nouveau document survienne à l'intérieur d'une catégorie donnée), et $P(w|z)$ (probabilité qu'un terme donné survienne à l'intérieur d'une catégorie donnée). Une phase d'apprentissage machine permet d'inférer ces probabilités de la collection de documents.

Un défaut souvent cité de PLSI est que le modèle n'est pas génératif : il ne donne pas explicitement un modèle permettant d'associer à un document donnée l'ensemble de ses paramètres latents (Blei et al., 2003). Seul est possible le calcul sur toute une collection de documents. Des palliatifs ont été proposés pour « projeter » un document extérieur à la collection sur l'espace calculé (*Folding in queries*) (Hofmann, 1999; Bast and Weber, 2005). À ce jour, PLSI a été testé sur des bases de petites tailles, ce qui relativise ses performances (Lawrie and Binkley, 2006).

2.4 Allocation Latente de Dirichlet et Distribution de Dirichlet lissée

L'Allocation Latente de Dirichlet *Latent Dirichlet Allocation* est un modèle génératif probabiliste (Blei et al., 2003). Partant de l'hypothèse que l'ordre des documents dans la collection et celui des mots dans un texte sont indifférents[2], LDA définit des modèles de mélanges finis sur des ensembles de sujets sous-jacents pour générer la collection, chaque sujet étant modélisé comme un mélange infini sur des probabilités des sujets sous-jacentes. Il a été démontré que PLSI est un cas particuler de LDA (Girolami and Kabán, 2003).

L'application à la recherche de documents estime la vraisemblance que la requête q soit générée, étant donné un document d et les paramètres du modèle. Dans un cas de classification textuelle, l'espace de représentation a été réduit de 99.6% de ses dimensions (Blei et al., 2003).

Les termes rares tendent à apparaître par « rafales » : si un terme rare apparaît dans un texte, il est probable qu'il y apparaisse plus qu'une seule fois. Les multinomiales sous-estiment notablement la queue de la distribution. Le modèle *Dirichlet Compound Multinomial* a été proposé pour produire des distributions à queue épaisse décrivant mieux la distribution des termes (Madsen et al., 2005), mais il entraîne des solutions à forme non close, impossibles à calculer de façon exacte. Le modèle « Distribution de Dirichlet lissée » (*smoothed Dirichlet allocation*) est un modèle génératif sur la base de la KL-divergence. Des approximations théoriquement fondées de ce modèle produisent des solutions à forme close similaires à celles produites par les multinomiales, mais avec des distributions à queue épaisse (Nallapati, 2006).

3 Fonction de proximité entre documents

Les métriques entre documents ont été étudiées par Bollman dans les années 80 (Bollmann and Cherniavsky, 1981; Bollmann, 1984). Six critères pour une « fonction de recherche documentaire raisonnable » ont été proposés (Fang et al., 2004) :

1. une relation d'ordre pour des paires de documents (pour deux documents de même longueur et pour une requête d'un seul terme, le document qui contient le plus d'occurences aura la plus haute place).
2. une relation d'ordre pour des différences entre trois documents (pour documents a, b et c, différence entre $a - b$ et $b - c$)
3. une contrainte sur la discrimination des termes (les documents qui contiennent plus d'occurences de termes discriminants devraient être préférés).
4. une relation sur l'influence des termes hors requête sur le classement
5. une relation sur la longueur d'une document, pour deux documents ayant la meme proportion de termes appartenant à la requête.
6. une contrainte sur l'interaction entre TF et la longueur du document.

Soient deux contraintes sur la fréquence `tf` des termes (relations 1 et 2), une contrainte sur la discrimination des termes (3), deux contrainte sur la normalisation de la longueur du document (4 et 5), et une contrainte sur l'interaction

[2]L'ordre des mots dans un texte est indifférent par l'hypothèse du *bag of words*.

entre TF et la longueur du document. Ces relations formalisent en fait les notions intuitives suivantes :

- on préfère un document qui contient plus d'occurences d'un terme (1)
- on préfère un document qui correspond à plus de termes (2)
- passer de 1 à 2 occurences a plus de conséquences que de 100 à 101 (2)
- on régule l'effet de TF et de IDF (3)
- à TF égal, on pénalise un document plus long (4), « mais pas trop » (5,6)
- on régule l'interaction entre TF et la longueur du document (6).

3.1 Noyaux de Fisher

Un *noyau* est une mesure de similarité entre deux éléments x et y ; il constitue un produit scalaire dans l'espace crée par une représentation alternative des éléments : $K(x, y) = s(x) \cdot s(y)$, où $s(x)$ est la re-représentation de x (Elkan, 2005).

Les *noyaux de Fisher* (Jaakkola and Haussler, 1999b; Jaakkola and Haussler, 1999a; Hofmann, 2000) combinent la souplesse et les performances des méthodes discriminatives avec la rigueur des modèles probabilistes génératifs et leur capacité à opérer sur des chaînes de longueur variable.

Ces noyaux ont été utilisés comme base pour l'apprentissage d'espaces sémantiques latents (Hofmann, 2000). Ils induisent aussi une métrique sur ces espaces (Siolas and d'Alché Buc, 2000; Siolas, 2003).

3.2 Autres fonctions de proximité

Il a été proposé une approche de la recherche documentaire à base de distances contextuelles entre termes tenant compte de de leur morphologie (Jing and Tzoukermann, 1999).

Deux domaines externes dans lesquels les questions de similarités sont récurrentes :

- ontologies (et Salotti, 2004)
- logique floue, il y a matière à reflexion sur des empruns dans ce domaine (Kehagias and Konstantinidou, 2003; Botana, 1999; Fan and Ma, 2002; Xuecheng, 1992; Fan and Xie, 1999)

De nouvelles fonctions de similarité apparaissent du fait de travaux théoriques (Billot et al., 2004; Lin, 1998) ou dans des contextes plus appliqués (Makkonen à propos d'*event tracking* par exemple (Juha Makkonen and Salmenkivi, 2002)

À venir : *Simplified Similarity Scoring Using Term Ranks*, SIGIR 2005, Vo Ngoc Anh (The University of Melbourne) Alistair Moffat (The University of Melbourne)

4 Catégorisation de texte et recherche documentaire

La classification partage avec la recherche d'informations un certain nombre de techniques (Sebastiani, 2002) :

- l'indexation
- les techniques typiques de Recherche Documentaire pour la comparaison entre les documents et les requêtes, et de reformulation des requêtes, sont souvent utilisées dans la construction des classificateurs

– l'évaluation

(Lewis, 1992b) va jusqu'à dire que les systèmes d'IR *sont* des classificateurs, qui fonctionnent en triant les documents en deux groupes : ceux qui seront retournés à l'utilisteur, et ceux qui ne le seront pas. Ce point de vue est appuyé par (Nallapati, 2004). Il peut être argué qu'il existe des différences plus fondamentales entre la recherche documentaire et les classificateurs :

– la recherche documentaire se fonde sur des requêtes a priori inconnues ; ceci entraîne des problèmes spécifiques, par exemple lorsque les documents subissent un traitement qu'il est nécessaire d'appliquer aux requête sans que l'opération soit clairement définie[3].
– les requêtes ont des spécificités qui les distinguent des documents. Par exemple, une question récurrente est celle des *requêtes courtes*, qu'il est nécessaire d'étoffer artificiellement pour obtenir de bonnes performances.

On a vu les rapports entre entités et recherche documentaire. Par ailleurs (Zhang et al., 2004) suggère que la reconnaissance d'entités peut aussi s'assimiler à de la catégorisation à deux catégories.

[3]C'est par exemple le cas des systèmes à réduction de dimensionalité basés su des modèles non génératifs, comme PLSI. Une phase d'apprentissage sur les documents définit un espace latent dans lequel il faut projeter les requêtes, problème non trivial dans le cas non génératif.

References

Carmen Alvarez, Philippe Langlais, and Jian-Yun Nie. 2004. Word pairs in language modeling for information retrieval.

Shun-Ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2) :251–276.

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where : geotagging web content. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA. ACM Press.

Holger Bast and Ingmar Weber. 2005. Insights from viewing ranked retrieval as rank aggregation. In *WIRI '05 : Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pages 232–239, Washington, DC, USA. IEEE Computer Society.

Paul N. Bennett. 2003. Using asymmetric distributions to improve text classifier probability estimates. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 111–118, New York, NY, USA. ACM Press.

Holger Billhardt, Daniel Borrajo, and Victor Maojo. 2002. A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 53(3) :236–249.

Antoine Billot, Itzhak Gilboa, and David Schmeidler. 2004. Axiomatization of an exponential similarity function. Technical Report 1485, Cowles Foundation, Yale University, September. available at http ://ideas.repec.org/p/cwl/cwldpp/1485.html.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation.

P. Bollmann and V. S. Cherniavsky. 1981. Measurement-theoretical investigation of the mz-metric. In *SIGIR '80 : Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 256–267, Kent, UK, UK. Butterworth & Co.

P. Bollmann. 1984. Two axioms for evaluation measures in information retrieval. In *SIGIR '84 : Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 233–245, Swinton, UK, UK. British Computer Society.

Francisco Botana. 1999. A fuzzy measure of similarity for instance-based learning. In *ISMIS '99 : Proceedings of the 11th International Symposium on Foundations of Intelligent Systems*, pages 439–447, London, UK. Springer-Verlag.

Maria Fernanda Caropreso, Stan Matwin, and Fabrizio Sebastiani. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. pages 78–102.

Katri A. Clodfelder. 2003. An lsa implementation against parallel texts in french and english. *HLT-NAACL 2003 Workshop : Building and Unsing Parallel Texts Data Driven Machin Transaltion and Beyond, pp. 111-114.*

William W. Cohen and Yoram Singer. 1999. Context-sensitive learning methods for text categorization. *ACM Trans. Inf. Syst.*, 17(2) :141–173.

W.W. Cohen. 1995. Learning to classify English text with ILP methods. In L. De Raedt, editor, *ILP95*, pages 3–24. DEPTCW.

Koby Crammer and Yoram Singer. 2002. A new family of online algorithms for category ranking. In *SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158, New York, NY, USA. ACM Press.

Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. 2001. Latent semantic kernels. In *ICML '01 : Proceedings of the Eighteenth International Conference on Machine Learning*, pages 66–73, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. 2002. Latent semantic kernels. *J. Intell. Inf. Syst.*, 18(2-3) :127–152.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis.

S. Dumais, G. Furnas, and T. Landauer. 1988. Using latent semantic analysis to improve access to textual information.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *CIKM '98 : Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, New York, NY, USA. ACM Press.

Georges Dupret. 2003. Latent concepts and the number orthogonal factors in latent semantic analysis. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 221–226, New York, NY, USA. ACM Press.

Charles Elkan. 2005. Deriving tf-idf as a fisher kernel. In Mariano P. Consens and Gonzalo Navarro, editors, *SPIRE*, volume 3772 of *Lecture Notes in Computer Science*, pages 295–300. Springer.

Zargayouna et Salotti. 2004. Mesure de similarité sémantique pour l'indexation de documents semi-structurés.

J. EUZENAT and P. VALTCHEV. 2003. An integrative proximity measure for ontology alignment.

Jiu-Lun Fan and Yuan-Liang Ma. 2002. Some new fuzzy entropy formulas. *Fuzzy Sets Syst.*, 128(2) :277–284.

Jiulun Fan and Weixin Xie. 1999. Some notes on similarity measure and proximity measure. *Fuzzy Sets Syst.*, 101(3) :403–412.

Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA. ACM Press.

Norbert Fuhr and Chris Buckley. 1991. A probabilistic learning approach for document indexing. *ACM Trans. Inf. Syst.*, 9(3) :223–248.

Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. In *ECDL '00 : Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 59–68, London, UK. Springer-Verlag.

Mark Girolami and Ata Kabán. 2003. On an equivalence between plsi and lda. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434, New York, NY, USA. ACM.

Xiaofei He, Deng Cai, Haifeng Liu, and Wei-Ying Ma. 2004. Locality preserving indexing for document representation. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 96–103, New York, NY, USA. ACM Press.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August.

T. Hofmann. 2000. Learning the similarity of documents : An information-geometric approach to document retrieval and categorization. In *Proceedings of Advances in Neural Information Processing Systems (NIPS'99) (vol. 12)*. Cambridge, MA : MIT Press.

Davad A. Hull. 1994. *Thesis for doctorate of Standford University, 1994*. Ph.D. thesis, Standford University.

T. Jaakkola and D. Haussler. 1999a. Probabilistic kernel regression models.

Tommi S. Jaakkola and David Haussler. 1999b. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA. MIT Press.

Hongyan Jing and Evelyne Tzoukermann. 1999. Information retrieval based on context distance and morphology. In *SIGIR '99 : Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 90–96. ACM.

Helena Ahonen-Myka Juha Makkonen and Marko Salmenkivi. 2002. Applying semantic classes in event detection and tracking. In *Proceedings of International Conference on Natural Language Processing (ICON 2002),*, pages pp. 175–183.

A. Kehagias and M. Konstantinidou. 2003. L-fuzzy valued inclusion measure, l-fuzzy similarity and l-fuzzy distance. *Fuzzy Sets Syst.*, 136(3) :313–332.

Yu-Hwan Kim, Shang-Yoon Hahn, and Byoung-Tak Zhang. 2000. Text filtering by boosting naive bayes classifiers. In *SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 168–175, New York, NY, USA. ACM Press.

E. Kokiopoulou and Y. Saad. 2004. Polynomial filtering in latent semantic indexing for information retrieval. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA. ACM Press.

M. Kurimo. 2000. Fast latent semantic indexing of spoken documents by using self-organizing maps.

Wai Lam and Chao Yang Ho. 1998. Using a generalized instance set for automatic text categorization. In *SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–89, New York, NY, USA. ACM Press.

Wai Lam, Ruizhang Huang, and Pik-Shan Cheung. 2004. Learning phonetic similarity for matching named entity translations and mining new translations. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 289–296, New York, NY, USA. ACM Press.

Pawat Laohawee and Arnon Rungsawang. 2000. Co-operative dsir text indexing system.

Leah S. Larkey and W. Bruce Croft. 1996. Combining classifiers in text categorization. In *SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297, New York, NY, USA. ACM Press.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95, New York, NY, USA. ACM Press.

Dawn J. Lawrie and David Binkley. 2006. The qalp projects use of information retrieval techniques in software engineering.

Verayuth Lertnattee and Thanaruk Theeramunkong. 2004. Effect of term distributions on centroid-based text categorization. *Inf. Sci. Inf. Comput. Sci.*, 158(1) :89–115.

David D. Lewis and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, Las Vegas, US.

David D. Lewis. 1992a. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92 : Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50, New York, NY, USA. ACM Press.

David D. Lewis. 1992b. Text representation for intelligent text retrieval : a classification-oriented view. pages 179–197.

David Dolan Lewis. 1992c. *Representation and learning in information retrieval*. Ph.D. thesis, Amherst, MA, USA.

Y. H. Li and Anil K. Jain. 1998. Classification of text documents. *Comput. J.*, 41(8) :537–546.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 266–272, New York, NY, USA. ACM Press.

Rasmus E. Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the dirichlet distribution. In *ICML '05 : Proceedings of the 22nd international conference on Machine learning*, pages 545–552, New York, NY, USA. ACM.

Ludovic Lebart Martin Rajman. 1998. Similarités pour données textuelles. In *MELLET, Sylvie (ed.). JADT 1998. 4èmes Journées internationales d'Analyse statistique des Données Textuelles. Nice : Université de Nice - Sophia Antipolis, 1998*, pages 545–556.

Donald Metzler and W. Bruce Croft. 2005. Modeling query term dependencies in information retrieval with markov random fields.

Donald Metzler and W. Bruce Croft. 2006. Beyond bags of words : Modeling implicit user preferences in information retrieval.

Dunja Mladenic. 1998. Feature subset selection in text-learning. In *ECML '98 : Proceedings of the 10th European Conference on Machine Learning*, pages 95–100, London, UK. Springer-Verlag.

Isabelle Moulinier and Jean-Gabriel Ganascia. 1996. Applying an existing machine learning algorithm to text categorization. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 343–354, London, UK. Springer-Verlag.

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 64–71, New York, NY, USA. ACM Press.

Ramesh Nallapati. 2006. *The Smoothed Dirichlet Distribution : Understanding Cross-Entropy Ranking in Information Retrieval*. Ph.D. thesis, University of Massachusetts, Amherst, MA, USA.

Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. 1997. Feature selection, perception learning, and a usability case study for text categorization. In *SIGIR '97 : Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–73, New York, NY, USA. ACM Press.

Arnon Rungsawang Pawat Laohawee, Athichat Tangpong. 2000. Parallel dsir text indexing system : Using multiple master/slave concept.

Martin Rajman and Romaric Besancon. 2002. Evaluation of a vector space similarity measure in a multilingual framework, sep.

Michael M. Richter. 1992. Classification and learning of similarity measures. Technical Report SR-92-18.

Miguel E. Ruiz and Padmini Srinivasan. 1999. Hierarchical neural networks for text categorization (poster abstract). In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 281–282, New York, NY, USA. ACM Press.

Arnon Rungsawang, Athichat Tangpong, and Pawat Laohawee. 1999. Parallel dsir text retrieval system.

A. Rungsawang. 1997. Distributional semantic based information retrieval.

A. Rungsawang. 1999a. Dsir : The first trec-7 attempt.

A. Rungsawang. 1999b. High-performance information retrieval.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.

Gerard Salton. 1989. *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Hinrich Schétze, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In *SIGIR '95 : Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 229–237, New York, NY, USA. ACM Press.

Fabrizio Sebastiani, Alessandro Sperduti, and Nicola Valdambrini. 2000. An improved boosting algorithm and its application to automated text categorization. Technical report, Paris, France, France.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1) :1–47.

Florian Seydoux, Martin Rajman, and Jean-Cédric Chappelier. 2006. *Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire*. Ph.D. thesis.

George Siolas and Florence d'Alch&#233 ; Buc. 2002. Mixtures of probabilistic pcas and fisher kernels for word and document modeling. In *ICANN '02 : Proceedings of the International Conference on Artificial Neural Networks*, pages 769–776, London, UK. Springer-Verlag.

Georges Siolas and Florence d'Alché Buc. 2000. Support vector machines based on a semantic kernel for text categorization. In *IJCNN '00 : Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)-Volume 5*, page 5205, Washington, DC, USA. IEEE Computer Society.

Georges Siolas. 2003. *Modèles Probabilistes et noyaux pour l'extraction d'informations à partir de documents*. Ph.D. thesis, Paris 6.

Pascal Soucy and Guy W. Mineau. 2005. Beyond tfidf weighting for text categorization in the vector space model. In *IJCAI*, pages 1130–1135.

Christopher Stokoe, Michael P. Oakes, and John Tait. 2003. Word sense disambiguation in information retrieval revisited. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166, New York, NY, USA. ACM Press.

P. W. Foltz "T. K. Landauer and D. Laham". 1998. Introduction to latent semantic analysis. In *Discourse Processes 25*, pages 259–284.

Hirotoshi Taira and Masahiko Haruno. 1999. Feature selection in svm text categorization. In *AAAI '99/IAAI '99 : Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, pages 480–486, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Chunqiang Tang, Sandhya Dwarkadas, and Zhichen Xu. 2004. On scaling latent semantic indexing for large peer-to-peer systems. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 112–121, New York, NY, USA. ACM Press.

K. Tsuda, M. Kawanabe, G. atsch, S. Sonnenburg, and K. uller. 2001. A new discriminative kernel from probabilistic models.

Koji Tsuda, Shotaro Akaho, Motoaki Kawanabe, and Klaus-Robert Müller. 2004. Asymptotic properties of the fisher kernel. *Neural Comput.*, 16(1) :115–137.

C. J. van Rijsbergen. 1979. *Information Retrieval*.

C. J. van Rijsbergen. 1992. *The Geometry of Information Retrieval*.

Ji-Rong Wen, Ni Lao, and Wei-Ying Ma. 2004. Probabilistic model for contextual retrieval. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 57–63, New York, NY, USA. ACM Press.

Erik D. Wiener, Jan O. Pedersen, and Andreas S. Weigend. 1995. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 317–332, Las Vegas, US.

Liu Xuecheng. 1992. Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy Sets Syst.*, 52(3) :305–318.

Yiming Yang and Christopher G. Chute. 1994. An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.*, 12(3) :252–277.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA. ACM Press.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML '97 : Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

ChengXiang Zhai and John Lafferty. 2002. Two-stage language models for information retrieval. In *SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA. ACM Press.

Cheng Xiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, New York, NY, USA. ACM Press.

Li Zhang, Yue Pan, and Tong Zhang. 2004. Focused named entity recognition using machine learning. In *SIGIR '04 : Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 281–288, New York, NY, USA. ACM Press.