

# Étude sur l'impact du sous-langage dans la classification automatique d'appels d'offres

François Paradis, Jian-Yun Nie  
Université de Montréal  
{paradifr,nie}@iro.umontreal.ca

**Résumé:** Dans cet article nous évaluons diverses approches pour filtrer le contenu « procédural » d'un document, et mesurons leur impact sur la classification d'une collection d'appels d'offres. Deux types d'approches sont testées : la sélection de termes à partir d'un vocabulaire de référence, et le filtrage de phrases. Nous ne trouvons pas de différence significative entre le vocabulaire de référence et celui de la collection d'entraînement. Par contre le filtrage par phrases donne d'excellents résultats sur notre collection, et peu même avantageusement être combiné à d'autres techniques de sélection.

## 1 Introduction

Les techniques de classification automatique assistée reposent sur la capacité de ces algorithmes à dégager des caractéristiques discriminantes pour les classes de documents qui leur sont données en entraînement. Or cette capacité est sensible à la « propreté » du corpus, c'est-à-dire que les documents contenant du texte hors-propos ou répétitif, tel qu'on en retrouve souvent sur le Web, en dégradent la performance. Nous nous intéressons dans notre travail au cas où le bruit est causé par la présence de sous-langage, qui contribue peu au thème du document, tel que le langage de base dans un article scientifique spécialisé, ou le langage procédural dans un appel d'offre.

La notion de sous-langage réfère à l'utilisation d'un langage spécifique comportant son propre vocabulaire et règles de composition, par rapport à une langue ou un langage de base. Les propriétés propres au sous-langage ont été étudiées dans [Leh82, Bib93]. On remarque entre autres que leur sujet est limité, que des restrictions lexicales, syntaxiques ou sémantiques s'appliquent (e.g. termes techniques), et qu'ils peuvent exhiber des règles « déviantes » de grammaire et l'utilisation de certains symboles. Plusieurs études

empiriques existent également, comme par exemple [Los95] qui étudie des sous-langages du domaine scientifique et leur construction automatique. Cet aspect de génération automatique est fortement lié à celui d'identification du sous-langage dans le document ou dans un passage du document [Sek95].

Le sous-langage est utilisé depuis longtemps en TALN (Traitement Automatique de la Langue Naturelle); ainsi les premiers succès de traduction automatique tels que TAUM-METEO [Isa87] ont-ils été obtenus sur des langages spécialisés. En recherche d'information et classification, l'utilisation du sous-langage passe généralement par l'emploi d'un thésaurus [Bru03]. Les thésaurus peuvent être génériques et construits manuellement, comme Wordnet [Voo93], ou spécifiques à un domaine, voire à une collection, et

```
<PRESOL>
<DATE> 0521
<YEAR> 99
<CLASSCOD> 75
<NAICS> 418210
<OFFADD> Office of Environmental Studies; 1323 Y Street, Washington, DC 22030
<SUBJECT> Office supplies and devices
<SOLNBR> N00140-04-Q-4555
<RESPDATE> 061399
<ARCHDATE> 07131999
<CONTACT> Mary Ann Deal, Contract Specialist, 301-
443-5329; Contracting Officer, Beatrice L. Woods, 301-
443-0043
<DESC> The office of Environmental Studies intends to procure
printer toner cartridges and supplies for the Naval Inventory
Control Point in Mechanicsburg, PA. Request for Quotation
(RFQ) N00140-04-Q-4555 contemplates an indefinite delivery
type firm fixed price order. This is a combined synopsis/solici-
tation for commercial items prepared in accordance with the
format in FAR Subpart 13.5, Test Program for Certain Com-
mercial Items, as supplemented with additional information
included in this notice. This announcement constitutes the
only solicitation; proposals are being requested, and a written
solicitation will not be issued. This is a 100% Total Small
Business Set-Aside. etc.
<URL> http://www.oes.gov
<EMAIL>
<ADDRESS> johndoe@usa.gov
<SETASIDE> Total Small Disadvantage Business
<POPZIP> 22030
<POPCOUNTRY> US
</PRESOL>
```

Figure 1: Exemple d'appel d'offres

construits automatiquement, comme dans [Bru97]. Bien que souvent utilisés pour l'expansion de requête [Mit98], on peut aussi concevoir de les utiliser directement dans la fonction de similarité.

Dans cet article nous évaluons diverses approches pour filtrer le contenu « procédural » d'un document, et mesurons leur impact sur la classification d'une collection d'appels d'offres. Deux types d'approches sont testées : la sélection de termes à partir d'un vocabulaire de référence, et le filtrage de phrases. Nous ne trouvons pas de différence significative entre le vocabulaire de référence et celui de la collection d'entraînement. Par contre le filtrage par phrases donne d'excellents résultats sur notre collection, et peu même avantageusement être combiné à d'autres techniques de sélection.

## 2 Contexte de l'étude

Cette étude se situe dans le cadre du projet MBOI (Matching Business Opportunities on the Internet) [Par04]. Ce projet a pour but de développer des outils pour aider à la veille d'appels d'offres, en développant les axes de la recherche d'information, de la classification et de l'extraction d'information, et du filtrage. Le volet qui nous intéresse ici, la classification, consiste à classer les appels d'offres par type d'industrie, selon les nombreuses normes en vigueur: SIC (*Standard Industrial Classification*), NAICS (North American Industry Classification System), FCS (Federal Supply Codes), CPV (Common Procurement Vocabulary), etc. Ces codes de classification ne sont pas toujours explicitement fournis dans les documents, et même lorsqu'ils le sont, il est intéressant de classer le même appel d'offre selon d'autres normes. Par exemple, un veilleur américain sera sûrement familier avec la norme NAICS, mais peut-être pas avec la norme CPV en vigueur dans l'Union Européenne. De même, ces normes sont régulièrement mises à jour, et la différence de codes entre deux versions peut être la source d'erreurs.

Afin de constituer une collection test, nous avons considéré les documents provenant d'un site d'appel d'offres du gouvernement américain, « FedBizOpps » (<http://www.fedbizopps.gov/>). Nous nous sommes limités aux documents qui contenaient deux codes de classification: FCS et NAICS (ceci afin de pouvoir mesurer plus tard la conversion de codes en utilisant la même base de documents). Nous avons ainsi obtenu 21945 appels d'offres (72Megs), répartis dans la période de septembre 2000 à octobre 2003. Nous avons scindé cette collection en deux: 60% pour l'entraînement des méthodes de classification, et 40% pour les tests.

La Figure 1 présente un exemple abrégé de document. Plusieurs méta- informations sont présentes, comme les dates (21 mai 1999), les codes de classification (« 75 » pour FCS, et « 418210 » pour NAICS), le courriel, etc. Le corps du document est formé du titre (Subject) et de la description : seuls ces deux champs sont conservés pour la classification. Cependant on remarque que la description comporte beaucoup de contenu qui n'est pas indicateur du type d'appel d'offre (en fait, toute la description, sauf la mention de « printer toner cartridges and supplies »), mais qui renseigne plutôt sur les dates et modalités de soumission à respecter. C'est ce que nous appelons le langage procédural. Dans notre corpus, une vingtaine de documents n'ont pour seul contenu qu'une référence à un document externe<sup>1</sup>. Les techniques que nous présentons dans les sections suivantes auront pour but de sélectionner le « véritable » contenu au travers du langage procédural.

Les codes NAICS sont hiérarchiques: chacun des six chiffres formant le code représente un niveau de la hiérarchie. Ainsi, pour l'exemple de la Figure 1, le code de classe d'industrie est « 424120 » (grossistes-marchands de papeterie et de fournitures de bureau) et le secteur, « 424 » (grossistes-marchands de biens non durables). Les trois pays participants à la norme, le Canada, le Mexique et les États-Unis, ont chacun leur propre version de la norme, qui en principe diffèrent surtout au niveau des classes d'industries (i.e. 5 ou 6 chiffres). Il y a cependant des exceptions, comme le montre notre exemple: la classe équivalente dans la version canadienne serait « 418210 » (grossistes-distributeurs de papeterie et de fournitures de bureau), et le code de secteur, « 418 » (grossistes-distributeurs de produits divers).

Nous avons réduit l'espace des catégories en ne retenant que les trois premiers chiffres, c'est-à-dire le secteur. Nous avons ainsi obtenu 92 catégories NAICS dans notre collection (vs. 101 pour les codes FCS). Nous n'avons pas normalisé pour la distribution inégale de ces catégories. Ainsi pour NAICS, 34% des documents sont dans les deux classes les plus courantes, alors que pour FCS, 33% se retrouvent dans les cinq premières classes.

Des premiers résultats pour la classification de cette collection (FBO) sont présentés au Tableau 1, pour les méthodes NB (Naive Bayes) [Ren03] et SVM (Support Vector Machines) [Joa98]. Ces résultats sont en accord avec la littérature, en particulier [Yan99] qui trouve aussi SVM plus performante sur la collection Reuters. La méthode NB est choisie pour sa simplicité, et parce qu'elle est la référence de comparaison dans la littérature. La méthode SVM, bien

---

<sup>1</sup> Les documents d'appel d'offres sur le Web ne sont souvent que des notes ou des sommaires, et le requérant doit payer pour obtenir le document complet.

qu'elle soit supérieure, est beaucoup plus lourde à utiliser sur de grandes collections, et moins flexible.

Le logiciel rainbow [MC96] a été utilisé. Nous avons aussi appliqué des techniques de seuillage standards [Yan01], soit : scut 50/50 avec 5 itérations (i.e. entraînement de 50% de la collection test pour optimiser la mesure macro-F1, répété 5 fois), wcut 0.001 (seuil de poids fixe), et rcut 1 (seuil de rang fixe).

Le Tableau 1 montre aussi l'effet de la sélection de termes (feature selection) sur cette collection, sous les entrées « IG8K ». Tel que prévu, NB étant très sensible au bruit, des gains significatifs sont obtenus en filtrant les termes. Nous présentons ici les résultats en ne gardant que les 8000 premiers termes selon la mesure d'IG (Information Gain), mais des résultats similaires sont obtenus pour d'autres mesures telles que le Chi-Square. Pour SVM cependant, l'effet est néfaste : il est bien connu que la sélection de termes est inutile pour SVM. Dans la suite de notre discussion, à moins d'indication contraire, le classifieur NB+IG8K constituera la base de comparaison.

Tableau 1: Résultats de référence FBO  
(résultats obtenus avec rcut1, wcut 0.001,scut 50/50 5 it.)

| méthode          | macro-F1     | micro-F1     |
|------------------|--------------|--------------|
| NB               | .1923        | .5256        |
| <b>NB + IG8K</b> | <b>.3297</b> | <b>.5498</b> |
| SVM              | .4351        | .6454        |
| SVM + IG8K       | .4228        | .6211        |

Il est à noter que la sélection de termes constitue déjà une première mesure de l'impact du vocabulaire sur la collection. En effet puisque IG mesure les termes les plus discriminants pour identifier une catégorie, on peut donc supposer que le langage procédural, qui est commun à toutes les catégories, sera éliminé.

### 3 Collection de référence

Notre but est de filtrer de façon automatique le bruit dans les appels d'offres, ou d'en faire ressortir le langage descriptif. Une première approche consiste à utiliser les descriptions du schéma de classification NAICS comme vocabulaire de contrôle. Nous avons donc créé une nouvelle collection à partir des 2320 descriptions disponibles (1.4Megs). Un exemple de description est présenté à la Figure 2. Ces descriptions couvrent les 6 niveaux de hiérarchie NAICS; puisque nous ne nous intéressons qu'au troisième niveau, chaque classe aura sa propre description ainsi que celle des quatrième, cinquième et sixième niveaux.

Nous avons également extrait les cinq mille termes les plus pertinents par la mesure IG sur l'ensemble de cette collection. IG est une mesure similaire à l'entropie proposée dans [Sek95], qui fait l'hypothèse que des documents utilisant un vocabulaire différent auront une entropie élevée. Une mesure alternative serait une distribution de Poisson [Los95].

**Category 111 : Oilseed and Grain Farming**

This industry group comprises establishments primarily engaged in (1) growing oilseed and/or grain crops and/or (2) producing oilseed and grain seeds. These crops have an annual life cycle and are typically grown in open fields.

**Category 111110 : Soybean Farming**

This industry comprises establishments primarily engaged in growing soybeans and/or producing soybean seeds. Establishments engaged in growing soybeans in combination with grain(s) with the soybeans or grain(s) not accounting for one-half of the establishment's agricultural production (value of crops for market) are classified in U.S. Industry 111191, Oilseed and Grain Combination Farming.

Figure 2: Exemple de descriptions NAICS

Tableau 2: Collection de référence NAICS

| méthode                      | macro-F1                        | micro-F1         |
|------------------------------|---------------------------------|------------------|
| NB sélection NAICS           | .3125<br>(-5.3%)                | .5516            |
| NB sélection NAICS+FBO       | .3074<br>(-6.8%)                | .5494            |
| NB comité NAICS+FBO          | <b>.3647</b><br><b>(+10.6%)</b> | .5470            |
| NB entraînement<br>NAICS+FBO | .3366<br>(+2.1%)                | .5416<br>(-1.5%) |
| NB paires IG                 | .3323                           | .5396<br>(-1.9%) |

Le Tableau 2 présente quelques expériences tentant d'exploiter cette collection de référence. La première expérience (« sélection NAICS ») consistait à sélectionner les termes sur la collection de référence plutôt que sur le corpus d'entraînement FBO. On remarque peu de changement au niveau de la mesure micro-F1 mais en

revanche une nette diminution de la mesure macro-F1. Même en combinant avec les 8000 termes filtrés par la collection d'entraînement FBO (ceci résulte en 10279 termes distincts), le même comportement est observé. Ceci est un peu surprenant puisque lors de tests précédents une sélection de 10,000 termes sur FBO améliorerait légèrement la mesure micro-F1, au détriment de la mesure macro-F1.

Une autre approche consiste à utiliser les descriptions NAICS pour l'entraînement (i.e. pour remplacer le corpus d'entraînement FBO). Nous avons ainsi entraîné un autre classifieur NB de cette manière (« entraînement NAICS ») et ajouté son résultat avec le classifieur de base NB+IG8K par la combinaison linéaire :

$$P(t | c) = \lambda_{c,naics} P(t | c, naics) + \lambda_{c,fbo} P(t | c, fbo),$$

où  $P(t | c, X)$  représente la probabilité d'un terme pour une catégorie donnée, telle que retournée par le classifieur X. Les coefficients  $\lambda_{c,naics}$  et  $\lambda_{c,fbo}$  sont estimés pour chaque catégorie en scindant la collection-test en deux, et en calculant sur la partie d'entraînement :

$$\lambda_{c,X} = \sum_{rang/doc(rang,X,c)estpertinent} \frac{1}{rang}$$

c'est-à-dire la somme inverse des rangs des documents retournés pertinents par le classifieur X pour la catégorie en question. Les coefficients vont donc tenter de privilégier le classifieur qui retourne le plus de documents pertinents de rang élevé. Il va sans dire que d'autres estimés auraient pu être utilisés, y compris la recherche des coefficients maximum étant donné la fonction F1; le problème est cependant compliqué par le seuillage effectué par la suite. Nous avons essayé quelques autres techniques et trouvé que cette fonction ad hoc fonctionnait relativement bien.

Les résultats de cette combinaison (« comité NAICS+FBO ») montrent une nette amélioration de la mesure macro-F1 (10.6%). Le classifieur NAICS semble avoir pris le relais lorsque les classes comportaient peu d'exemples dans la collection FBO, d'où l'effet marqué sur la moyenne des classes, mais puisqu'il s'agit de peu de documents, le peu d'effet sur la mesure micro-F1.<sup>2</sup> Ceci est démontré, dans une moindre mesure, lorsqu'on entraîne à la fois sur les descriptions NAICS et le corpus FBO (« entraînement NAICS+FBO »), puisqu'on observe également une augmentation de la mesure macro-F1.

Nous avons également tenté d'utiliser le vocabulaire afin d'ajouter de l'information aux documents, plutôt que d'en retrancher

<sup>2</sup> Les résultats de ce classifieur seul sur la collection FBO sont de : macro-F1 .1481 et micro-F1 .3451.

comme c'est le cas avec la sélection de termes. Pour ce faire nous avons utilisé la mesure IG sur des couples de termes, en gardant les 10,000 plus discriminants. Nous avons ensuite ajouté un terme complexe dans le document lorsque deux termes simples de la liste s'y retrouvaient. Par exemple, lorsqu'on retrouvait « product » et « manufacturing » dans un document, on a ajouté « product\_manufacturing ». Les résultats de cette expérience ne sont pas très concluants : ils conduisent à une légère diminution de la mesure micro-F1 (« NB paires IG »).

## 4 Sélection de phrases

Une autre approche pour filtrer le contenu consiste à identifier les phrases les plus porteuses de sens. Cette approche est très utilisée pour la production automatique de résumés. Par exemple, on peut définir une mesure de pertinence pour une phrase basée sur sa position, sa longueur la fréquence de termes et sa similarité avec le titre [Nob95]. Dans un premier temps, afin de mieux évaluer le problème, nous avons créé encore une fois une collection de référence.

Nous avons créé une collection de 1000 phrases prises parmi 41 documents. Nous avons étiqueté ces phrases manuellement selon leur pertinence, en gardant une phrase dès qu'elle avait du contenu descriptif; mais par contre en éliminant les phrases discutant uniquement de la procédure de soumission, des règlements à respecter, des dates de livraison, etc. Ainsi, dans l'exemple de la Figure 1, seule la première phrase serait conservée. En général, près du quart des phrases (243) ont été jugées pertinentes. Bien que les premières phrases soient souvent pertinentes, ce n'est pas toujours le cas. En fait il semble difficile de dégager des règles d'enchaînement de phrases.

Nous avons entraîné un classifieur NB sur cette collection de phrases, en définissant deux classes : pertinente ou non pertinente. Il semble d'emblée que ce problème de filtrage soit beaucoup plus simple, puisqu'on obtient une mesure micro-F1 de 85% (classifieur NB entraîné sur un sous-ensemble de 600 phrases). Nous avons ainsi filtré tout le corpus avec ce classifieur, pour le faire passer de près de 600,000 phrases à 96,811. Les nouveaux documents ont alors été classifiés comme auparavant.

Le Tableau 1 présente les résultats, sans sélection de termes et avec la sélection de termes IG8K. La différence entre ces deux résultats semble indiquer que notre sélection de phrases est additive par rapport aux sélections de termes, une propriété qui serait très désirable, et qui n'est habituellement pas le cas entre les techniques de sélection de termes.

Puisque le corpus d'entraînement de phrases était très petit, nous avons aussi essayé de normaliser certaines constructions telles que les dates dans l'espoir que le classifieur de phrase pourrait apprendre des « patrons » sans avoir vu toutes les instances. Ainsi, nous avons remplacé par des étiquettes uniques dans la collection de départ toutes les instances de dates, numéros de règlements FAR, courriels, numéros de téléphone et montants monétaires. Nous avons ensuite refait le processus d'entraînement et de filtrage de phrases. Les résultats (« NB patrons+IG8K ») constituent les meilleurs résultats obtenus jusqu'à présent; une augmentation de 8% de la micro mesure sur NB+IG8K.

Tableau 3: Sélection de phrases

| méthode            | macro-F1                     | micro-F1                      |
|--------------------|------------------------------|-------------------------------|
| NB                 | .2366<br>(+23%) <sup>3</sup> | .5635<br>(+7.2%) <sup>3</sup> |
| NB+IG8K            | .3223<br>(-2.2%)             | .5918<br>(+7.6%)              |
| NB<br>patrons+IG8K | .3497<br>(+6.1%)             | <b>.5939</b><br><b>(+8%)</b>  |

## 5 Conclusion

Nous n'avons pas trouvé de différence significative entre le vocabulaire de référence NAICS et celui de la collection d'entraînement FBO. Cependant un gain significatif peut être réalisé sur les classes comportant peu de documents en entraînement; le vocabulaire extrait de la collection de descriptions NAICS vient alors compléter l'information manquante. L'approche de comité que nous avons préconisé pourrait également nous permettre de combiner le classifieur NB avec le classifieur SVM, ce qui devrait aussi améliorer la macro-F1 de ce dernier.

Les résultats préliminaires pour la sélection de phrases sont très encourageants, et le fait qu'elle puisse se combiner avec la sélection de termes est très avantageux. La prochaine étape sera de trouver des critères automatiques pour cette sélection basés sur le vocabulaire, c'est-à-dire de faire le lien entre nos deux approches. Nous comptons aussi étudier la correspondance des entités nommées telles que les dates et unités monétaires (c'est-à-dire les patrons dans notre expérience) et la pertinence des phrases. Enfin, une autre piste pour la sélection de phrases serait la détection de style : en effet les appels d'offres ont souvent un style télégraphique, emploient la forme impérative, et incluent des « identificateurs » tels que des numéros de produit.

<sup>3</sup> Par rapport à NB sans sélection d'information.

## 6 Remerciements

Ce projet de recherche a été financé conjointement par Nstein Technologies et le CRSNG.

## Bibliographie

- [Bib93] Biber, D. Using register-diversified corpora for general language studies, Computational linguistics, vol19, no2, 1993.
- [Bru97] Marie-France Bruandet, Jean-Pierre Chevallet et François Paradis, Construction de thésaurus dans le système de recherche d'information IOTA : application à l'extraction de la terminologie, in 1eres Journees Scientifiques et Techniques du Reseau Francophone de l'Ingerierie de la Langue de l'AUPELF-URF , Avignon - France, pp537-544, 15-16 Avril, 1997.
- [Bru03] Marie-France Bruandet, Jean-Pierre Chevallet, Utilisation et construction de bases de connaissances pour la Recherche d'Informations, in Assistance Intelligente à la Recherche d'Information , M.-H. Stefanini, E. Gaussier, Hermes, chapter 3, pp85-118, 2003.
- [Isa87] Isabelle P., Machine Translation at the TAUM Group, in Margaret King (ed.), Machine Translation Today: The State of the Art, Edinburgh University Press, 1987.
- [Joa98] Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, in ECML-98, 10th European Conference on Machine Learning, 1998.
- [Leh82] Lehrberger, J., Automatic translation and the concept of sublanguage, in Sublanguage, 1982.
- [Los95] Losee, Robert M ; Haas, Stephanie W, Sublanguage Terms: Dictionaries, Usage, and Automatic Classification, Journal of the American Society for Information Science, 46(7), p. 519-529, 1995.
- [MC96] McCallum, Andrew Kachites, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, <http://www.cs.cmu.edu/~mccallum/bow>, 1996.
- [Mit98] Mandar Mitra, Amit Singhal, Chris Buckley, Improving Automatic Query Expansion, SIGIR, Melbourne, Australia, pp206-214, 1998.
- [Nob95] Chikashi Nobata, Satoshi Sekine, Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Hitoshi Isahara, Sentence Extraction System Assembling Multiple Evidence, Published in the Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, 2001; Tokyo, Japan.
- [Par04] François Paradis, Quing Ma, Jian-Yun Nie, Stéphane Vaucher, Jean-François Garneau, Robert Gérin-Lajoie and Arman Tajarobi, MBOI: Un outil pour la veille d'opportunités sur l'Internet, in Colloque sur la Veille Strategique Scientifique et Technologique, Toulouse, France, 25-29 October, 2004.
- [Ren03] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan and David R. Karger, Tackling the Poor Assumptions of Naive Bayes Text Classifiers, in Proceedings of the Twentieth International Conference on Machine Learning, 2003.
- [Sek95] Satoshi Sekine, A New Direction for Sublanguage NLP, Journal of Gengo

Syori Gakkai (Natural Language Processing),1995.

- [Teu97] Simone Teufel et Marc Moens. Sentence Extraction as a Classification Task. In Inderjeet Mani and Mark T. Maybury, editors, Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 1997.
- [Teu98] Teufel, S., et Moens, M.. Sentence extraction and rhetorical classification for flexible abstracts. In Spring AAAI Symposium on Intelligent Text summarization, 1998.
- [Voo93] Ellen M. Voorhees, Using WordNet to disambiguate word senses for text retrieval, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, pp171-180, 1993.
- [Yan99] Yiming Yang et Xin Liu, 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, Berkeley, CA, pp42-49, August 15-19, 1999.
- [Yan01] Yiming Yang, A Study on Thresholding Strategies for Text Categorization, in Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval, 2001.