

UNIVERSITÉ LUMIÈRE LYON2

Année 2003

THÈSE
pour obtenir le grade de
DOCTEUR
en
INFORMATIQUE

présentée et soutenue publiquement par

Radwan JALAM

le 4 juin 2003

Apprentissage automatique et catégorisation de textes multilingues

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Jean-Hugues CHAUCHAT

DEVANT LE JURY, COMPOSÉ DE :

Annie MORIN, Rapporteur	Maître de conférences habilitée, IRISA, Rennes
Yves KODRATOFF, Rapporteur	Directeur de recherche, CNRS, LRI Orsay
Martin RAJMAN, Rapporteur	Professeur, Ecole Polytechnique Fédérale, Lausanne
Geneviève BOIDIN-LALLICH, Examineur	Professeur, Université Claude Bernard-Lyon 1
Ludovic LEBART, Examineur	Directeur de recherche, CNRS, ENST Paris
Jean-Hugues CHAUCHAT, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

Introduction	1
I. Catégorisation de textes monolingues	5
1. Catégorisation de textes	7
1.1. Introduction	8
1.2. Définition de la catégorisation de texte	8
1.3. Comment catégoriser un texte ?	9
1.3.1. Représentation, le codage, des textes	10
1.3.2. Choix de classifieurs	11
1.3.3. Évaluation de la qualité des classifieurs	12
1.4. Applications de la catégorisation de texte	12
1.4.1. Catégorisation de textes : une fin en soi	13
1.4.2. Catégorisation de textes : un support pour différentes appli- cations	13
1.5. Difficultés particulières de la catégorisation de textes	13
1.5.1. Grandes dimensions	14
1.5.2. Imprécision des fréquences	15
1.5.3. Déséquilibre	15
1.5.4. Ambiguïté	15
1.5.5. Synonymie	15
1.5.6. Subjectivité de la décision	16
1.6. Lien avec la recherche documentaire	16
1.7. Jeu de données utilisé pour l'évaluation	18
1.8. Conclusion	19
2. Approches pour la représentation de textes	21
2.1. Introduction	22

2.2.	Choix de termes	22
2.2.1.	Représentation en « sac de mots »	22
2.2.2.	Représentation des textes par des phrases	23
2.2.3.	Représentation des textes avec des racines lexicales et des lemmes	24
2.2.4.	Méthodes basées sur les n-grammes	24
2.3.	Codage des termes	26
2.3.1.	Codage $TF \times IDF$	27
2.3.2.	Codage TFC	27
2.4.	Réduction de la dimension	28
2.4.1.	Réduction locale de dimension	29
2.4.2.	Réduction globale de dimension	29
2.4.3.	Sélection de termes	29
2.4.4.	Extraction de termes	30
2.5.	Conclusion	30
3.	Sélection multivariée de termes	33
3.1.	Introduction	34
3.2.	Méthode du χ^2 univariée	34
3.3.	Méthode du χ^2 multivarié	36
3.4.	Expérimentation	36
3.5.	Conclusion	38
4.	Pourquoi les n-grammes fonctionnent	39
4.1.	Introduction	40
4.2.	Intérêt du codage en n-grammes	40
4.3.	Étapes de la recherche des mots caractéristiques	41
4.3.1.	Recherche des n-grammes caractéristiques et des mots qui les contiennent	41
4.3.2.	Filtrage des mots « parasites »	42
4.3.3.	Algorithme complet	42
4.4.	Exemple d'application	42
4.4.1.	Données indexées de Reuters	42
4.4.2.	Quelques résultats	44
4.4.3.	Discussion des résultats sur la collection Reuters	44
4.5.	Conclusion	46
5.	Techniques pour la construction de classifieurs	51
5.1.	Introduction	52
5.1.1.	Manière de construction du classifieur	52
5.1.2.	Caractéristique du modèle	53
5.2.	Méthode de Rocchio	53
5.3.	Arbres de décision	55

5.3.1.	Phase d'apprentissage	56
5.3.2.	Phase de classification	61
5.3.3.	Critiques de la méthode	61
5.4.	Classifieurs à base d'exemples	62
5.4.1.	K-plus proches voisins	63
5.5.	Fonctions à bases radiales	67
5.6.	Machine à Vecteurs de Support	68
5.6.1.	Cas des classes linéairement séparables	69
5.6.2.	Cas des classes non séparables	70
5.7.	Évaluation de classifieurs de textes	71
5.7.1.	Évaluation des classifieurs, l'approche « binaire »	72
5.7.2.	Évaluation des classifieurs, l'approche « multi-classes »	76
5.8.	Contributions personnelles	77
5.8.1.	Nouvelle utilisation des SVM	77
5.8.2.	Nouvelle utilisation des réseaux RBF	77
5.8.3.	Nos expérimentations	77
5.9.	Conclusion : quel est le meilleur classifieur ?	79

II. Catégorisation de textes multilingues 83

6.	Catégorisation multilingue : les solutions proposées	85
6.1.	Introduction	86
6.2.	Intérêt accru aux traitements multilingues	86
6.2.1.	Davantage de collections numériques	86
6.2.2.	Plus de personnes connectées en ligne	87
6.2.3.	Plus de globalisation et de pays unifiés	87
6.2.4.	Réseau plus rapide et plus souple	88
6.3.	Recherche documentaire multilingues	88
6.3.1.	Approches basées sur la traduction automatique	89
6.3.2.	Thésaurus multilingues	90
6.3.3.	Utilisation de dictionnaires	90
6.4.	Nos solutions pour catégoriser des textes multilingues	91
6.4.1.	Premier schéma : le schéma trivial	92
6.4.2.	Deuxième schéma : choisir une seule langue d'apprentissage	93
6.4.3.	Troisième schéma : mélanger les ensembles d'apprentissage	93
6.5.	Conclusion	94
7.	Identification de la langue	97
7.1.	Introduction	98
7.2.	Approches linguistiques	99
7.2.1.	Présence de certains chaînes de caractères spécifiques	99
7.2.2.	Présence de certains mots	100

7.2.3.	Approche lexicale	101
7.2.4.	Approche plus linguistique	101
7.3.	Approches statistiques et probabilistes	102
7.3.1.	Mots les plus fréquents	102
7.3.2.	Méthodes basées sur les n -grammes	103
7.4.	Expériences pour la reconnaissance de langue	107
7.5.	Conclusion	109
8.	Traduction automatique	111
8.1.	Introduction	112
8.2.	Premières approches historiques de la traduction automatique	112
8.2.1.	Décryptage	112
8.2.2.	Analyse par micro-contexte	112
8.2.3.	Imiter la traduction humaine	113
8.3.	Nouvelles approches, plus modestes	113
8.3.1.	Mémoire de traduction	113
8.3.2.	Sous-langages et langages contrôlés	114
8.4.	Évaluer la traduction automatique	114
8.5.	Conclusion	115
9.	Cadre pour la catégorisation de textes multilingues	117
9.1.	Introduction	118
9.2.	Méthodes pour la catégorisation de textes multilingues	118
9.2.1.	Nouveau cadre pour la catégorisation multilingue	118
9.2.2.	Détection de la langue du texte à classer	119
9.2.3.	Traduction du texte à classer	119
9.3.	Application sur les corpus CLEF	120
9.3.1.	Constitution du corpus	120
9.3.2.	Représentation des textes	122
9.3.3.	Algorithmes d'apprentissage	123
9.3.4.	Reconnaissance de la langue	123
9.3.5.	Catégorisation des articles	123
9.4.	Discussion	128
9.5.	Conclusion	129
	Conclusion et perspectives	133
	Index des auteurs cités	137
	Bibliographie	141

Index des auteurs cités

- Aas, K. 11, 23, 28, 41
Adam, Chai K. 11
Afnor 121
Aha, David W. 63, 123
Albert, Marc K. 63, 123
Allan, James 28
Amati, Gianni 75
Amini, Massih-Réza 37, 53, 71
Androutsopoulos, Ion 11, 13, 75
Apté, Chidanand 11, 23, 29, 53, 62, 76
Auray, J.P. 58
- Ballesteros, Lisa 90
Beesley, K. 98, 99, 103, 104
Benzecri, J. P. 30, 106, 107
Bernick, Myrna 11
Biskri, I. 2, 26, 40
Borko, Harold 11
Breiman, L 66
Broomhead, D.S. 67
Buckley, Chris 28
Buckley, Christopher 14, 23, 55
Burgess, Christopher J. C. 70, 77
- Callan, James P. 54, 55
Carbonell, Jaime 12
Caropreso, Maria Fernanda 11, 23, 24, 29
Carreras, Xavier 12
- Catlett, Jason 62
Cavnar, William B. 11, 12, 24, 40, 46, 99, 104, 105
Chai, Kian M. 11
Chandioux, J. 112–114, 116
Chandrinou, Konstandinos V. 11, 13, 75
Chauchat, Jean Hugues 30
Chieu, Hai L. 11
Churcher, G. 99
Chute, C. 14
Chute, Christopher G. 11, 53
Clech, Jérémy 3, 122, 133
Cleverdon, C. 16
C.Nicholas 25, 26, 40, 41
Cohen, William W. 13, 54, 55, 62, 75, 76
Cormier, Monique C. 114
Cornuéjols, Antoine 68, 69, 71
Cover, T M 63
Cowie, J. 103
Creedy, Robert M. 63
Crestani, Fabio 75
Croft, W. Bruce 77, 90
- Damashek, Marc 40
Damerau, Fred J. 11, 23, 29, 53, 62, 76
Darken, C. 67
de Loupy, Claude 11, 24
Deerwester, S. 30
Delisle, S. 2, 26, 40

- Déjean, H. 101
Dorr, Bonnie 89
Dumais, Susan 11, 18, 23, 28–31, 53, 54, 71, 74, 80
Dunning, T. x, 40, 99, 100, 109
Duru, G. 58
- Eikvil, L. 11, 23, 28, 41
Elisseeff, André 69, 78
Escofier, Brigitte 30
Escudero, Gerard 12
- Fernandez, Rodrigo 68, 69
Fix, E 63
Fluhr, Christian 89
Forsyth, Richard S. 12
Friedman, J. 66
Fürnkranz, J. 41, 46
Fuhr, Norbert 13, 14, 23, 62
Furnas, G. 30
- Ganascia, Jean-Gabriel 76
Giguët, E. 99–101
Gilleron, Rémi 61, 63, 65
Gilli, Y. 22
Giroi, Federico 67
Goetz, Thilo 62
Goh, Wei B. 29, 31, 53
Good, I. 78, 79
Grefenstette, Gregory 40, 90, 99, 100, 102, 103
Grobelnik, Marko 29
Guan, Jihong 41
Guerneur, Y. 78
Gull, Aaron 133
- Hahn, Shang-Yoon 12
Hallab, M. 40, 46
Hampp, Thomas 62
Hancock-Beaulieu, Micheline 133
Harding, P. 13
Harshman, R. 30
Hart, P E 63
Hartmann, Stephan 13, 23, 62
Hastie, T. 66
Hatzivassiloglou, Vasileios 12, 29
Hayes, Ph. 13, 52, 99
He, Ji 11, 35, 71, 134
Heckerman, David 11, 18, 23, 29, 31, 53, 54, 71, 74, 80
Hirsh, Haym 62
Hodges, J L 63
Hovy, E. 114
Hughes, J. 99
Hull, David A. 11, 23, 24, 29, 35, 53, 54, 90, 134
- Iyer, Raj D. 12, 53
- Jain, Anil K. 29, 62, 63
Jalam, Radwan 3, 12, 40, 119, 122, 123, 133, 134
Joachims, Thorsten ix, 11, 15, 16, 18, 30, 53–55, 62, 63, 71, 76, 78
Johnson, David E. 62
Johnson, S. 99
- Kibler, Dennis 63, 123
Kim, Yu-Hwan 12
King, Margaret 114
Knorz, Gerhard 13, 62
Kodratoff, Y. 1
Kohavi, Ron 72
Koller, Daphne 31
Koutsias, John 11, 13, 75
K.Sparck-Jones 133
- Lam, Wai 63, 77
Landauer, T. 30
Larkey, Leah S. 63, 77
Lau, Marianna 133
Lebart, L. 1, 11, 30, 106
Lefèvre, Philippe 14, 15, 17
Lelu, A. 40, 46
Lewis, David D. 1, 11, 12, 17, 18, 23, 29–31, 53–55, 62, 72, 75, 76
Li, Yong H. 29, 62, 63

-
- Liddy, Elizabeth D. 13
 Liu, H. 34
 Liu, J. 25, 26, 40, 41
 Liu, Xin 11, 18, 29, 55, 63, 78–80
 Liu, Yan 12
 Loewe, D. 67
 Low, Kok L. 29, 31, 53
 Ludovik, E. 103
 Lustig, Gerhard 13, 62

 Manning, Christopher 26
 Maron, M.E. 12
 Márquez, Lluís 12
 Masand, Brij M. 63
 Matwin, Stan 11, 23, 24, 29
 McGill, M. 11, 14, 22, 53, 133
 Miller, E. 25, 26, 40, 41
 Mitchell, Tom M. 1, 37, 62, 123
 Mitra, Mandar 55
 Mladenić, Dunja 11, 29
 Moody, J. 67
 Morin, Annie 30
 Motoda, H. 34
 Moulinier, Isabelle 1, 13, 17, 18, 29, 53, 73, 74, 76, 80
 Muhlenbach, Fabrice 62
 Mustonen, S. 99

 Ng, Hwee T. 11, 29, 31, 53
 Nunberg, Geoffrey 87

 Oard, Douglas 89, 90
 Oles, Frank J. 62
 Olshen, R A 66

 Paik, Woojin 13
 Papineni, K. 115
 Papka, Ron 54, 55
 Paugam-Moisy, H. 78
 Pedersen, Jan O. 11, 23, 29, 31, 35, 53, 54, 63, 77, 134
 Peters, Carol 1, 86, 90

 Platt, John 11, 18, 23, 29, 31, 53, 54, 71, 74, 80
 Poggio, Tomaso 67
 Popescu-Belis, A. 114
 Porter, M. F. 24
 Powell, M. 67
 Péry-Woodley, M. P. 101

 Quinlan, J.R. 62, 123

 Rajman, M. 106
 Rakotomalala, Ricco 3, 56, 61, 88, 122, 133
 Reynar, J. 98, 106
 Rigau, German 12
 Ringuette, Marc 11, 29, 53, 62
 Robertson, Stephen 13, 133
 Rocchio, J.J. 53
 Roukos, S. 115
 Ruiz, Miguel E. 63, 77

 Sable, Carl L. 12, 29
 Sahami, Mehran 11, 18, 23, 26, 29, 31, 41, 53, 54, 71, 74, 80
 Salem, A. 11, 30
 Salton, G. 11, 14, 22, 28, 53, 55, 133
 Saporta, Gilbert 64
 Savoy, J. 120
 Schapire, Robert E. 12, 18, 53–55, 62, 75, 77
 Schmid, Helmut 24
 Schütze, Hinrich 11, 23, 26, 29, 35, 53, 54, 134
 Schwantner, Michael 13, 62
 Sebastiani, Fabrizio 1, 2, 10–14, 16, 17, 23, 24, 27–30, 37, 62, 72, 80, 134, 154
 Shannon, C. 25
 Shen, D. 25, 26, 40, 41
 Sheridan, Paraic 1, 86, 90
 Shortliffe, E. H. 52
 Sibun, P. 98, 106
 Singer, Yoram 12, 18, 53–55, 62, 75–77
 Singhal, Amit 12, 18, 28, 53, 55, 75

- Smith, Stephen J. 63
Soergel, Dagobert 90
Souter, C. 99
Sperduti, Alessandro 80
Spyropoulos, Constantine D. 11, 13, 75
Srinivasan, Padmini 63, 77
Stone, C J 66
Stone, M. 124
Stricker, Mathieu 11, 30, 53

Tan, Ah-Hwee 11, 35, 71, 134
Tan, Chew-Lim 11, 35, 71, 134
Teytaud, Olivier 3, 12, 40, 119, 123, 134
Thiel, E. 65
Tibshirani, R. 66
Tommasi, Marc 61, 63, 65
Trenkle, John M. 11, 12, 24, 40, 46, 99, 104, 105
Tzeras, Konstadinos 13, 23, 62

Uren, Victoria 16

Valdambrini, Nicola 80
Van Rijsbergen, C. J. 76
Vapnik, V 68

Viennet, Emmanuel 68, 69
Vinot, Romain 54, 55
Volk, Martin 113, 114

Walker, Steve 133
Waltz, David L. 63
Ward, T. 115
Weigend, Andreas S. 11, 29, 31, 35, 53, 134
Weinstein, Steven P. 13, 52
Weiss, Sholom M. 11, 23, 29, 53, 62, 76
Wiener, Erik D. 11, 29, 31, 35, 53, 134

Yang, Y. 14
Yang, Yiming 11, 12, 18, 29, 31, 35, 41, 44, 53, 55, 63, 74, 76–81, 134
Yu, Edmund S. 13
Yvon, François 54, 55

Zacharski, R. 103
Zhang, Byoung-Tak 12
Zhou, Shuigeng 41
Zhu, W. 115
Zighed, Djamel A. 56, 58, 60, 61, 88