

R

Massih-Reza AMINI - Éric GAUSSIER

Préface de Stephen Robertson

Recherche d'information

Applications, modèles et algorithmes

Fouille de données,
décisionnel et big data

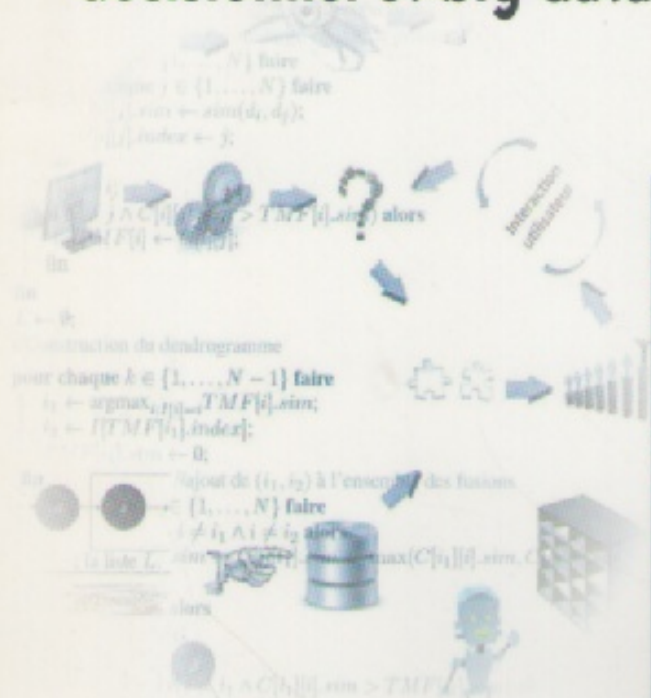


Table des matières

Préface	V
Liste des algorithmes	XI
Notations	XIII
Liste des tableaux	XV
Liste des figures	XVII
1 Introduction	1
1.1 Concepts étudiés dans ce livre	3
1.2 Organisation du livre	6
2 Représentation et indexation	9
2.1 Prétraitements linguistiques	10
2.1.1 Segmentation	11
2.1.2 Normalisation	13
2.1.3 Filtrage par un antidictionnaire	16
2.2 Les deux lois de base en recherche d'information	18
2.2.1 Loi de Heaps	19
2.2.2 Loi de Zipf	20
2.3 Représentation documentaire	21
2.3.1 Modèle vectoriel	21

2.3.2	Pondération des termes	23
2.4	Index inversé	26
2.4.1	Indexation dans des collections statiques	27
2.4.2	Indexation dans des collections dynamiques	30
2.5	Exercices	33
3	Recherche d'information	45
3.1	Modèles de recherche	46
3.1.1	Modèles booléens	47
3.1.2	Modèles vectoriels	49
3.1.3	Modèles probabilistes	53
3.1.4	Une approche axiomatique de la RI	66
3.2	Expansion de requêtes	68
3.2.1	La méthode « boucle de rétopertinence »	69
3.2.2	La méthode « boucle de rétopertinence en aveugle »	71
3.3	Mesures d'évaluation	71
3.3.1	Évaluation de résultats non ordonnés	72
3.3.2	Évaluation de résultats ordonnés	74
3.4	Exercices	80
4	Recherche sur le Web	99
4.1	Architecture de la Toile	100
4.2	Trois inventions à la base du Web	100
4.2.1	Langage HTML	101
4.2.2	Protocole de transfert hypertexte et adresses Web	103
4.3	Collecte et indexation des pages sur la Toile	104
4.3.1	Robot d'indexation	104
4.3.2	Index distribués	108
4.4	Nouvelles stratégies de recherche	109
4.4.1	Modèle d'apprentissage automatique pour la RI	110
4.4.2	PageRank	113
4.5	Exercices	116
5	Catégorisation de documents	121
5.1	Formalisme	122
5.2	Sélection de variables	124

5.2.1	Le seuillage sur la mesure <i>Document Frequency</i> (df)	125
5.2.2	L'information mutuelle ponctuelle (IMP)	125
5.2.3	L'information mutuelle (IM)	127
5.2.4	La mesure χ^2	128
5.3	Modèles génératifs	130
5.3.1	Modèle multivarié de Bernoulli	131
5.3.2	Modèle multinomial	134
5.4	Modèles discriminants	137
5.4.1	Modèle logistique	140
5.4.2	Séparateurs à vaste marge	142
5.5	Mesures d'évaluation	146
5.6	Exercices	149
6	Partitionnement de documents	163
6.1	Définitions	164
6.2	Les étapes du partitionnement	165
6.3	Principaux algorithmes de partitionnement	170
6.3.1	Partitionnement à plat : méthodes de réallocation	170
6.3.2	Partitionnement hiérarchique	178
6.4	Évaluation	188
6.5	Applications à l'accès à l'information	190
6.6	Exercices	191
7	Recherche de thèmes latents	201
7.1	Analyse sémantique latente	202
7.1.1	Décomposition en valeurs singulières	203
7.1.2	L'analyse sémantique latente pour la RI	205
7.1.3	Limitations	207
7.2	Analyse sémantique latente probabiliste	207
7.2.1	Remarques	210
7.3	Le modèle LDA	211
7.4	Exercices	213
8	Considérations pratiques	219
8.1	Logiciels libres pour la recherche d'information	220
8.1.1	dpSearch	221

8.1.2	Lucene/SolR	221
8.1.3	MG	221
8.1.4	Terrier	221
8.1.5	Zettair	222
8.2	Logiciels libres pour la catégorisation et le partitionnement	222
8.3	Le passage à l'échelle ou le Big Data	223
8.3.1	Traitement parallèle et distribué	223
8.3.2	Traitement de flux de données	224