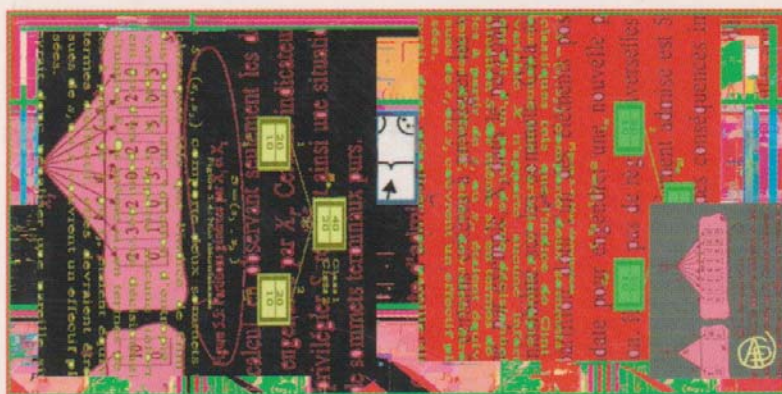


Annie Morin
Georges Hebrail

Patrick Bosc
Ludovic Lebart

2^e CYCLE • ÉCOLES D'INGÉNIEURS

Bases de données et statistique



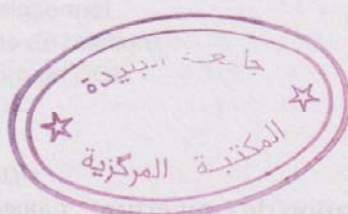
DUNOD

005-379-1

2-005-379-1

TABLE DES MATIÈRES

Bases de données et statistiques



Annie Morin

IRISA-IFSIC

Patrick Bosc

IRISA-ENSSAT

Georges Hebrail

EDF-Division R&D

Ludovic Lebart

CNRS-ENST

DUNOD

TABLE DES MATIERES

Introduction	vii
Les auteurs	xiii
Remerciements	xiii

PARTIE 1

Bases et entrepôts de données

Chapitre 1. Bases de données relationnelles et SQL	2
1.1 Généralités sur les bases de données	2
1.2 Brève introduction au modèle relationnel	4
1.3 Redondance et formes normales de relations	10
1.4 Principales constructions du langage SQL	14
1.5 Conclusion	19
Chapitre 2. Méthodes d'accès aux BDR	21
2.1 Introduction	21
2.2 Les méthodes d'accès	23
2.3 Présentation de quelques algorithmes	27
2.4 Généralités sur les optimiseurs de requêtes	31
2.5 Conseils, quelques manières d'opérer	35
2.6 Conclusion	38
Chapitre 3. Les systèmes de recherche d'informations	40
3.1 Présentation du domaine	40
3.2 Les modèles classiques de recherche d'informations	42
3.3 Les mesures qualitatives des systèmes de recherche d'informations	56
3.4 La problématique de l'indexation	60
3.5 Conclusion générale	68
Chapitre 4. Entrepôts de données et OLAP introduction	69
4.1 Introduction	69
4.2 Une architecture pour les entrepôts de données	72
4.3 Construction des entrepôts de données	78
4.4 Vue conceptuelle multidimensionnelle des données	79
4.5 Interrogation des entrepôts de données	84
4.6 OLAP et bases de données statistiques	89
4.7 Conclusion	96

PARTIE 2

Apprentissage et recherche de règles

Chapitre 5. Graphes d'induction : apprentissage automatique et Data Mining	98
5.1 Introduction	98
5.2 Un exemple	100
5.3 Définitions générales et notations	100

5.4	Principes des graphes d'induction	102
5.5	Du graphe d'induction aux règles de prédiction	106
5.6	Mesure de qualité d'une partition dans les GI	107
5.7	La méthode SIPINA	113
5.8	Application	123
5.9	Conclusion	124
Chapitre 6. Objets symboliques et analyse de données		125
6.1	Position du problème	125
6.2	Des données aux connaissances	126
6.3	Principes d'acquisition d'objets symboliques à partir d'une base de données	127
6.4	Construction d'assertions à partir d'une base de données	133
6.5	Méthode de classification divisive	138
6.6	Conclusion	142
Chapitre 7. Quelques contraintes symboliques sur le numérique en ECD et en ECT		144
7.1	Introduction	144
7.2	Le Paradoxe de Hempel et la théorie de confirmation	145
7.3	Des connaissances précises et significatives	146
7.4	Des connaissances utiles : les mesures d'intérêt	157
7.5	Conclusion : Définition de l'ECD et de l'ECT	160
Chapitre 8. Introduction à la recherche de règles d'associations et séquences fréquentes		163
8.1	Présentation du problème	164
8.2	La recherche de règles d'associations	165
8.3	La recherche de séquences fréquentes	178
8.4	Présentation des procédures de SAS Enterprise Miner	186
8.5	Conclusions et perspectives	190

PARTIE 3

Quelques aspects de l'analyse statistique des données

Chapitre 9. Problèmes statistiques et numériques liés à l'échelle des données.		
Données manquantes		192
9.1	Les nouveaux problèmes	192
9.2	Bases de données, structures de données et statistique	193
9.3	Problèmes statistiques et numériques liés à l'échelle des données	195
9.4	Problèmes de données manquantes ou sous-tableaux manquants	199
9.5	Conclusion	205
Chapitre 10. Sélection de variable		207
10.1	Introduction	207
10.2	Processus de sélection et pertinence d'un attribut	208
10.3	Approche statistique	211
10.4	Quantité d'information	214
10.5	Les algorithmes génétiques	215
10.6	Les réseaux de neurones	218
10.7	Sélection d'attributs et ensembles approximatifs	220
10.8	Conclusion	226

PARTIE 4

Traitement des données textuelles

Chapitre 11. Apprentissage et données textuelles	230
11.1 Introduction	230
11.2 Les Méthodes de l'apprentissage	231
11.3 La recherche d'information	236
11.4 L'extraction d'information	240
11.5 Accès à l'information	244
11.6 Un exemple	246
11.7 Conclusion	251
Chapitre 12. Analyse et visualisation de données mixtes : textuelles/numériques	253
12.1 Les étapes de la statistique textuelle	253
12.2 Le perfectionnement des outils	254
12.3 Les unités et les similarités	256
12.4 Les analyses statistiques : les trois outils de base.	262
12.5 L'apport des analyses linguistiques	264
Chapitre 13. Quelle représentation pour les textes dans les bases de données biomédicales ?	269
13.1 Enjeux et donnée de base	269
13.2 De l'article scientifique à la base de données bibliographiques	270
13.3 Constitution du corpus	271
13.4 Représentation des articles dans les bases bibliographiques	275
13.5 Traitement des données	278
13.6 Conclusion	281
Chapitre 14. Analyse des données textuelles et banque de données bibliographiques	282
14.1 Le choix des mots à analyser	282
14.2 Sélection de la bibliographie des maladies rares	288
14.3 Exemple de découverte d'une banque de données bibliographique	290
14.4 Discussion	297

PARTIE 5

Processus et mise en œuvre du Data Mining

Chapitre 15. Bases de données et analyse de données (Databases and Data Mining) vers des responsables informationnels	302
15.1 Introduction	302
15.2 Présentation des systèmes d'information, orientés décision	303
15.3 Bases de données	306
15.4 Poste de travail d'un responsable informationnel	314
15.5 Conclusions	318
Chapitre 16. Applications industrielles du Data Mining	319
16.1 Les entreprises et le Web	319
16.2 Pourquoi utiliser le Data Mining ?	322
16.3 Le data warehouse	323
16.4 Comment développer un projet Data Mining	328
16.5 Quelques exemples	328
16.6 Perspectives	333

Chapitre 17. Ressources statistiques disponibles sur la toile :	
de la convergence fonctionnelle à la veille informationnelle	335
17.1 Convergence des technologies de l'information	336
17.2 La veille informationnelle	344
17.3 Pratiques de la recherche d'informations	346
17.4 Innovation technique, propriété intellectuelle et éthique applicable aux traitements nominatifs	350
 BIBLIOGRAPHIE	 357