

TRAITÉ DES SCIENCES ET TECHNIQUES DE L'INFORMATION

Les systèmes de recherche d'informations

modèles conceptuels



*sous la direction de
Madjid Ihadjadene*

Hermès

Lavoisier

G173

(2)

28929

COMITÉ SCIENTIFIQUE DU

TRAITÉ DES SCIENCES ET TECHNIQUES DE L'INFORMATION

Les systèmes de recherche d'informations

modèles conceptuels



sous la direction de

Madjid Ihadjadene

hermes
Science
— publications —

Table des matières

| | | |
|--|---|----|
| <p>+ Introduction</p> <p>Madjid IHADJADENE</p> | 15 | |
| <p>Chapitre 1. Le modèle booléen</p> <p>Madjid IHADJADENE et Hubert FONDIN</p> | | 19 |
| <p>+ 1.1. Introduction : architecture d'un SRI</p> <p>1.2. Le modèle booléen ou ensembliste</p> <p>1.3. Formulation de la question dans le système booléen.</p> <p> 1.3.1. Les opérateurs booléens</p> <p> 1.3.2. La troncature.</p> <p> 1.3.3. Les opérateurs de proximité.</p> <p> 1.3.4. Les opérateurs de comparaison (ou d'échelle)</p> <p> 1.3.5. Priorité entre les opérateurs</p> <p>1.4. Stratégie de recherche dans le modèle booléen</p> <p>1.5. Vocabulaire d'accès et SRI booléens.</p> <p>1.6. Les fonctions d'analyse statistiques et de tri dans le modèle booléen.</p> <p>1.7. Les limites du modèle booléen</p> <p>• 1.8. Extension du modèle booléen</p> <p>1.9. Conclusion</p> <p>1.10. Bibliographie</p> | <p>19</p> <p>20</p> <p>22</p> <p>23</p> <p>24</p> <p>25</p> <p>25</p> <p>25</p> <p>26</p> <p>27</p> <p>28</p> <p>30</p> <p>30</p> <p>32</p> <p>33</p> | |
| <p>Chapitre 2. Technologies statistiques pour la recherche d'informations : les modèles vectoriels</p> <p>Romarc BESANÇON</p> | | 35 |
| <p>2.1. Introduction.</p> <p>2.2. Le modèle vectoriel standard</p> | <p>35</p> <p>36</p> | |

| | |
|--|------------|
| 2.2.1. La sélection des termes d'indexation | 37 |
| 2.2.2. Les schémas de pondération | 39 |
| 2.2.2.1. Pondération locale | 39 |
| 2.2.2.2. Pondération globale | 40 |
| 2.2.2.3. Normalisation | 40 |
| 2.2.2.4. Combinaison des pondérations | 42 |
| 2.2.3. La matrice d'occurrence | 43 |
| 2.2.4. Les mesures de similarité | 43 |
| 2.3. Variantes et extensions du modèle vectoriel | 45 |
| 2.3.1. Expansion de requêtes et intégration de co-occurrences | 45 |
| 2.3.2. Modèle vectoriel généralisé | 46 |
| 2.3.3. Espace des documents et espace des requêtes | 46 |
| 2.3.4. Le retour de pertinence | 46 |
| 2.3.5. Le modèle <i>Latent Semantic Indexing</i> | 47 |
| 2.4. Conclusion | 50 |
| 2.5. Bibliographie | 51 |
| Chapitre 3. Modèles probabilistes en recherche d'informations | 55 |
| Jian-Yun NIE et Jacques SAVOY | |
| 3.1. Introduction | 55 |
| 3.2. Le modèle de recherche probabiliste | 57 |
| 3.2.1. Un modèle simple | 57 |
| 3.2.2. Le modèle probabiliste de base | 58 |
| 3.2.3. Rétroaction | 61 |
| 3.2.4. Estimation <i>a priori</i> | 62 |
| 3.3. Indexation probabiliste | 62 |
| 3.4. Le modèle unifié | 65 |
| 3.5. Extensions récentes du modèle probabiliste | 67 |
| 3.6. Conclusion | 70 |
| 3.7. Bibliographie | 71 |
| Chapitre 4. Connexionisme et génétique pour la recherche d'informations | 77 |
| Mohand BOUGHANEM et Lynda TAMINE | |
| 4.1. Introduction | 77 |
| 4.2. Evaluation de la pertinence en recherche d'informations | 78 |
| 4.2.1. L'utilisateur au centre du processus | 79 |
| 4.2.2. Apprentissage : émergence du besoin | 80 |
| 4.3. Le modèle connexionniste | 80 |
| 4.3.1. Le modèle PIRCS | 81 |
| 4.3.2. Le modèle MERCURE | 83 |
| 4.3.3. Synthèse | 86 |
| 4.4. Génétique et recherche d'informations | 86 |
| 4.4.1. Qu'est-ce qu'un algorithme génétique ? | 87 |
| 4.4.1.1. Modélisation génétique | 88 |
| 4.4.1.2. Analyse formelle | 89 |
| 4.4.2. Recherche d'informations basée sur la génétique | 89 |
| 4.4.2.1. Traitement génétique des documents | 90 |
| 4.4.2.2. Traitement génétique des requêtes | 92 |
| 4.4.2.3. Processus génétique de recherche | 94 |
| 4.4.3. Synthèse | 98 |
| 4.5. Conclusion | 98 |
| 4.6. Bibliographie | 99 |
| Chapitre 5. Modélisation logique pour la recherche d'informations | 105 |
| Jean-Pierre CHEVALLET | |
| 5.1. Introduction | 105 |
| 5.2. Modéliser la correspondance | 107 |
| 5.2.1. Relation de correspondance | 108 |
| 5.2.2. Orientation de la correspondance | 109 |
| 5.2.3. Classification de la correspondance | 110 |
| 5.2.3.1. Correspondance par équivalence | 111 |
| 5.2.3.2. Correspondance par appartenance | 111 |
| 5.2.3.3. Correspondance d'inclusion | 111 |
| 5.2.3.4. Correspondance d'intersection | 111 |
| 5.3. Modélisation logique de la correspondance | 112 |
| 5.3.1. Une logique pour la RI : que choisir ? | 115 |
| 5.3.2. Logique des propositions | 115 |
| 5.3.3. Application à la modélisation de la RI | 118 |
| 5.3.4. Limites de cette modélisation | 120 |
| 5.3.5. Extension à logique trivaluée | 121 |
| 5.4. Modélisation par une logique modale | 121 |
| 5.4.1. La logique modale | 122 |
| 5.4.2. Un modèle modal de RI | 123 |
| 5.4.3. Un modèle théorique modal flou | 124 |
| 5.5. Autres pistes de modélisations logiques | 126 |
| 5.5.1. Logique abductive | 127 |
| 5.5.2. Logique non monotone | 127 |
| 5.5.3. Logique <i>imaging</i> | 128 |
| 5.5.4. La théorie des situations | 128 |
| 5.6. Les modèles logiques opérationnels | 129 |
| 5.6.1. Indexation par des arbres sémantiques | 129 |

| | |
|--|-----|
| 5.6.2. La logique terminologique | 130 |
| 5.6.3. Les graphes conceptuels | 131 |
| 5.6.3.1. Les graphes conceptuels et la logique | 133 |
| 5.6.3.2. Le processus de correspondance | 134 |
| 5.7. Conclusion | 134 |
| 5.8. Bibliographie | 135 |

Chapitre 6. SRI et traitement du langage naturel 139

Claude DE LOUPY et Eric CRESTAN

| | |
|---|-----|
| 6.1. Introduction | 139 |
| 6.2. Le niveau morphologique | 140 |
| 6.2.1. La graphie | 140 |
| 6.2.2. Les variantes grammaticales | 141 |
| 6.2.3. Détection d'éléments importants | 142 |
| 6.2.4. Les variations morphologiques | 144 |
| 6.3. Le niveau syntaxique | 146 |
| 6.4. Le niveau sémantique | 146 |
| 6.4.1. Quelques expériences | 147 |
| 6.4.2. Quelle ressource pour la gestion de la synonymie et de la polysémie ? | 150 |
| 6.4.3. Le découpage thématique | 151 |
| 6.5. Le multilinguisme | 152 |
| 6.6. Aide à la navigation et évaluation des difficultés | 153 |
| 6.6.1. Aide à la navigation | 154 |
| 6.6.2. Evaluation des difficultés | 154 |
| 6.7. Conclusion | 156 |
| 6.8. Bibliographie | 157 |

Chapitre 7. Modèles de langue pour la recherche d'informations 163

Mohand BOUGHANEM, Wessel KRAAIJ, Jian-Yun NIE

| | |
|--|-----|
| 7.1. Introduction | 163 |
| 7.2. Modèles de langue en linguistique informatique | 164 |
| 7.2.1. Idée de base | 164 |
| 7.2.2. Lissage | 166 |
| 7.3. Modèle de langue en RI – principes | 170 |
| 7.4. RI comme génération de la requête par le document | 172 |
| 7.4.1. Modèle de Ponte et Croft | 172 |
| 7.4.2. Modèle de Hiemstra <i>et al.</i> et de Miller <i>et al.</i> | 174 |
| 7.5. La RI comme ratio de vraisemblance | 176 |
| 7.6. Modèle basé sur l'entropie croisée | 177 |

| | |
|---|-----|
| 7.7. La RI comme traduction statistique | 180 |
| 7.8. Discussions | 181 |
| 7.9. Bibliographie | 182 |

Chapitre 8. L'évaluation des systèmes de recherche d'informations 185

Stéphane CHAUDIRON

| | |
|--|-----|
| 8.1. Introduction | 185 |
| 8.2. Les origines de l'évaluation des SRI | 186 |
| 8.2.1. Les tests de Cranfield | 187 |
| 8.2.1.1. Cranfield I | 187 |
| 8.2.1.2. Cranfield II | 189 |
| 8.2.1.3. La mesure de la performance dans les tests de Cranfield | 190 |
| 8.2.1.4. Les caractéristiques des tests de Cranfield | 193 |
| 8.2.2. MEDLARS | 194 |
| 8.2.3. Le projet SMART | 195 |
| 8.2.4. Le projet STAIRS | 197 |
| 8.3. Les campagnes TREC | 198 |
| 8.3.1. La méthodologie TREC | 200 |
| 8.3.1.1. Les collections TREC | 200 |
| 8.3.1.2. Les thèmes (topics) | 200 |
| 8.3.1.3. La construction du référentiel : la <i>pooling method</i> | 201 |
| 8.3.1.4. Une phase d'entraînement et une phase de test | 201 |
| 8.3.2. Les mesures d'évaluation dans TREC | 202 |
| 8.3.3. Limites et interrogations | 202 |
| 8.4. Conclusion | 204 |
| 8.5. Bibliographie | 205 |

| | |
|-----------------|-----|
| Index | 209 |
|-----------------|-----|

Le Traité des Sciences et Techniques de l'Information répond au besoin de disposer d'un fonds commun de connaissances dans les domaines où les technologies ont transformé la production et la gestion de l'information. Absorbant et transformant des disciplines traditionnelles, comme la documentation et la bibliothéconomie, les sciences de l'information occupent en effet un champ à part, où se rencontrent l'informatique et les sciences humaines et sociales.

Les ouvrages du Traité des Sciences et Techniques de l'Information analysent les différents aspects d'une discipline étendue et diverse : les documents et leur gestion, l'information dans la société, la communication homme-machine, la représentation et l'instrumentalisation des connaissances. Ils abordent les technologies de base, les institutions, les grandes options méthodologiques et les savoir-faire.

Chaque volume étudie aussi bien les aspects fondamentaux qu'expérimentaux. Une classification des différents chapitres contenus dans chacun, une bibliographie et un index détaillé orientent le lecteur vers ses points d'intérêt immédiats : celui-ci dispose ainsi d'un guide pour ses réflexions ou pour ses choix.