



La recherche d'informations

du texte intégral au thésaurus

Philippe Lefèvre

Hermès

3146

La recherche d'informations

du texte intégral au thésaurus



$\frac{1}{1}$

Philippe Lefèvre

Hermès
Science
— publications —

Table des matières

Préface	11
Introduction	15
Préambule	19
Chapitre 1. Caractéristiques du langage naturel. Traitements linguistiques.	21
1.1. Implicite et pragmatique	22
1.2. Redondance et synonymie	23
1.2.1. Equivalence entre mots ou entre mots et expressions : synonymie	23
1.2.2. Equivalence entre expressions : paraphrase	24
1.2.3. Synonymie et glissements de sens	25
1.3. Ambiguïté et polysémie	26
1.3.1. Ambiguïtés sur les mots : homonymie et polysémie	26
1.3.2. Ambiguïtés de niveau syntaxique : homotaxies et ambiguïtés structurales	28
1.3.3. Ambiguïtés de rapport au contexte	30
1.4. Sens de base et rôles complémentaires	32
1.5. Réduction des ambiguïtés et mots composés	33
Atomicité sémantique	35
Institutionnalisation de l'usage	35
Inséparabilité des composants	35
1.6. Niveaux de traitement linguistique	37
1.6.1. Découpage	37
1.6.2. Morphologie et lexique	38
1.6.3. Syntaxe	39

1.6.4. Sémantique	40	3.5.1. Principes généraux d'un thésaurus.....	91
1.7. Les ordinateurs pourront-ils un jour comprendre le langage naturel ?	41	3.5.2. Enrichissement de l'indexation par autopostage	95
Facultés d'abstraction et de modélisation de l'être humain	42	3.5.3. Exemple du thésaurus EDF.....	96
Que signifie comprendre pour un être humain ?	42	3.5.4. Utilisation du thésaurus à EDF pour l'indexation automatique..	97
Limitations actuelles de la compréhension automatique des textes....	43		
Et dans l'avenir ?	44		
Chapitre 2. Problématique de la recherche d'informations	47	Chapitre 4. Analyse et indexation documentaires	101
2.1. Cadre de la recherche d'informations.....	47	4.1. Catalogage et notices documentaires	101
2.1.1. Distinction des types d'information, méthodes et usages.....	47	4.2. Analyse documentaire.....	102
2.1.2. Plan de classement, analyse et recherche documentaire.....	52	4.3. Définition de l'indexation documentaire	105
2.1.3. Principe de fonctionnement et types de requêtes.....	54	4.4. Les méthodes d'indexation documentaire possibles.....	106
2.1.4. Pertinence des réponses et efficacité de la recherche	55	4.4.1. Type 0 : Indexation libre par fichier inverse brut	109
2.2. Difficultés dues aux aspects linguistiques	58	4.4.2. Type 1 : Indexation libre par fichier inverse de mots significatifs	110
2.2.1. Silence dû aux phénomènes de synonymie, de paraphrase et d'implicite	59	4.4.3. Type 2 : Indexation contrôlée par fichier inverse de mots appartenant à une liste	111
2.2.2. Silence et bruit dus aux phénomènes de synonymie et de rôles ..	61	4.4.4. Type 3 : Indexation libre par fichier inverse de racines ou par fichier inverse de mots significatifs associés par un radical commun.....	111
2.2.3. Bruit engendré par la polysémie et les rôles des mots employés ..	62	4.4.5. Type 4 : Indexation libre par fichier inverse de lemmes	113
2.2.4. Multilinguisme	64	4.4.6. Type 5 : Indexation libre par fichier inverse de syntagmes ou mots composés	114
2.3. Difficultés propres à la recherche d'informations.....	66	4.4.7. Type 6 : Indexation libre par syntagmes nominaux étendus	115
2.3.1. Support des documents qui constituent la base	66	4.4.8. Type 7 : Indexation contrôlée par liste terminologique	117
2.3.2. Formats de représentation des textes.....	67	4.4.9. Type 8 : Indexation contrôlée par champs sémantiques ou liste d'autorité.....	118
2.3.3. Ergonomie et méthode d'interrogation	68	4.4.10. Type 9 : indexation contrôlée par thésaurus.....	120
2.3.4. Généralité ou précision et but de la question	68	4.4.11. Type 10 : indexation à rôles	121
2.3.5. Profondeur de l'analyse ou de l'indexation	69	4.4.12. Type 11 : indexation contrôlée par domaines ou vedettes matières	122
2.3.6. Explosion combinatoire engendrée par la recherche.....	70	4.4.13. Type 12 : indexation structurée	123
Chapitre 3. Langages documentaires	71	4.5. Typologie des méthodes d'indexation	124
3.1. Les langages documentaires, réponse au problème de l'équivocité du langage naturel	71	4.6. Autres applications des techniques de TALN et d'indexation	127
3.2. Typologie des langages documentaires	74	4.6.1. Applications principales	128
3.2.1. Coordination	74	4.6.2. Besoins génériques.....	130
3.2.2. Coordination et composition.....	76	4.6.3. Applications secondaires.....	131
3.2.3. Contrôle.....	77	4.6.4. Applications complexes.....	132
3.2.4. Précision	78		
3.2.5. Synthèse de ces caractéristiques dans une typologie.....	79	Chapitre 5. Techniques et modes de requêtes	135
3.3. Classifications hiérarchiques. Exemple de la classification de Dewey	82	5.1. Requêtes et indexation	135
3.4. Systèmes de vedettes-matières à facettes. Exemple de RAMEAU.....	86	5.2. Référentiels terminologiques : réseaux sémantiques, terminologies et thésaurus.....	136
3.5. Langages combinatoires. Exemple du thésaurus.....	89		

5.2.1. Réseaux sémantiques. Exemple de WordNet.....	136	6.5.2. Choix des logiciels.....	181
5.2.2. Terminologies.....	138	6.5.3. Caractérisation indexation/requêtes des logiciels étudiés.....	185
5.2.3. Référentiel terminologique, réseau sémantique et thésaurus.....	140		
5.2.4. Correspondance des référentiels terminologiques.....	141	Chapitre 7. Internet et les évolutions de la recherche d'informations.....	193
5.2.5. Expansion des requêtes.....	141	7.1. Recherche d'informations sur l'internet/spécificités par rapport	
5.3. Processus de recherche, historique et sauvegarde des requêtes.....	142	aux systèmes documentaires traditionnels.....	194
5.3.1. Processus itératif de recherche d'informations.....	142	7.1.1. Diversité des sources sur l'internet.....	195
5.3.2. Recherche par similarité et filtrage des réponses.....	143	7.1.2. Quantité et croissance des informations.....	196
5.3.3. Historique des requêtes et combinaison des lots-résultats.....	145	7.1.3. Informations primaires et secondaires.....	197
5.3.4. Sauvegarde et réutilisation des requêtes-profil.....	146	7.1.4. Documentation interne et externe.....	197
5.4. Opérateurs utilisés dans les requêtes. Langage des requêtes.....	147	7.1.5. Hétérogénéité et mélange des informations et diversité	
5.4.1. Opérateurs booléens.....	147	des formats.....	198
5.4.2. Opérateurs booléens pondérés.....	148	7.1.6. Structuration ou non des bases.....	199
5.4.3. Opérateurs de comparaison numérique.....	149	7.1.7. Méthodes d'indexation : utilisation ou non d'un langage	
5.4.4. Opérateurs sur le texte intégral : variantes sur les mots.....	149	documentaire.....	199
5.4.5. Opérateurs sur le texte intégral : proximité et appartenance.....	150	7.1.8. Indexation totale ou partielle du corpus. Web invisible.....	200
5.4.6. Opérateurs sur les concepts et le domaine.....	151	7.1.9. Qualité et fiabilité des informations.....	201
5.5. Types de requêtes possibles.....	152	7.1.10. Multilinguisme.....	201
5.5.1. Type 1 : requête par grille ou formulaire sur les champs		7.1.11. Méthodes d'interrogation et mélange des modes de recherche.....	202
de catalogage, et utilisation de descripteurs d'un langage contrôlé.....	153	7.1.12. Consultation par des spécialistes et des non-spécialistes.....	202
5.5.2. Type 2 : requête par utilisation de mots du langage naturel		7.1.13. Evolution du profil des responsables de l'alimentation	
(langage non contrôlé).....	154	de certains serveurs.....	203
5.5.3. Type 3 : requête par utilisation de phrases courtes		7.1.14. Synthèse des conditions de recherche d'informations	
en langage naturel.....	156	sur l'internet par rapport à l'approche traditionnelle.....	203
5.5.4. Type 4 : requête par utilisation de textes ou de documents		7.2. Quels types d'indexation vaut-il mieux utiliser ?.....	204
en langage naturel. Requête par l'exemple ou par similarité.....	158	7.2.1. Inversion simple et limitation nécessaire du niveau d'indexation	
5.5.5. Récapitulation des modes de requête.....	158	des moteurs de recherche sur le Web.....	204
		7.2.2. Nécessité de la lemmatisation sur les bases internes.....	206
Chapitre 6. Caractérisation des moteurs de recherche.....	161	7.2.3. Problématique de l'indexation à des niveaux trop élevés.....	206
6.1. Présentation des résultats de la recherche.....	161	7.2.4. Y a-t-il un niveau optimal d'indexation automatique ?.....	208
6.2. Stratégies de sélection des documents : méthodes de calcul		7.2.5. L'indexation multiple, clef du problème ?.....	209
de la pertinence.....	163	7.3. Outils de recherche et d'analyse de l'information pour l'internet.....	210
6.2.1. Qu'est-ce que la pertinence ?.....	163	7.3.1. Les outils de base : répertoires thématiques et moteurs	
6.2.2. Modèle booléen pondéré.....	166	de recherche.....	210
6.2.3. Modèle vectoriel.....	168	7.3.2. Les outils annexes.....	216
6.2.4. Modèle probabiliste.....	170	7.3.3. Les méta-outils et outils d'analyse.....	216
6.2.5. Autres modèles.....	175	7.3.4. Les agglomérats d'outils et les portails.....	221
6.3. Gestion des index et performances des accès.....	176	7.4. Quelques perspectives d'avenir pour la recherche d'informations.....	224
6.4. Mesures de performances.....	177	7.4.1. Inventer de nouvelles méthodes d'indexation.....	224
6.5. Typologie indexation / requêtes appliquée à des logiciels		7.4.2. Internet : indexation et recherche en deux étapes	
du commerce.....	180	par constitution de bases intermédiaires.....	225
6.5.1 Typologie.....	180		

10 La recherche d'informations

7.4.3. Améliorer et utiliser systématiquement le classement automatique	226
7.4.4. Internet : mettre en place une classification universelle normalisée ?	227
7.4.5. Disposer de logiciels de recherche translingues	228
7.4.6. Développer l'interactivité permettant à l'utilisateur d'affiner sa demande	228
7.4.7. Technologies push, profils, et agents de recherche semi-autonomes	229
7.4.8. Associer plusieurs méthodes d'accès à l'information : les portails	230
7.4.9. Doit-on remettre en cause la conception de la recherche documentaire et de la constitution du savoir ?	231
Conclusion	233
Abréviations	237
Bibliographie	239
Index	243

Comment y voir clair dans l'effervescence qui règne aujourd'hui parmi les outils d'accès à l'information : logiciels de gestion documentaire, moteurs de recherche, agents de toutes sortes, portails... Sur quelle base comparer leurs fonctionnalités très diverses ? Quels sont par exemple les rapports entre la fonction d'indexation-recherche, le classement des informations, le filtrage et le *push* ? Autant de questions auxquelles cet ouvrage apporte des réponses claires.

La recherche d'informations – du texte intégral au thésaurus propose en effet une vision de synthèse réalisant le lien entre plusieurs mondes : la documentation traditionnelle, la GEIDE et l'ingénierie linguistique, enfin le nouveau monde de l'internet.

Les difficultés à trouver les informations, y compris sur l'internet, sont surtout dues aux caractéristiques du langage. Partant de ce constat, ce livre fait l'inventaire des problèmes liés à la recherche d'informations ; puis il décrit les techniques appliquées pour les résoudre, des plus anciennes aux plus récentes.

L'indexation est à la base de tous les traitements de l'information textuelle ; cette fonction est étudiée en détail, ce qui amène à distinguer treize modes d'indexation différents. Une telle typologie permet de caractériser sur un plan fonctionnel les principaux moteurs de recherche du marché.

Enfin, l'internet est abordé en montrant ses différences avec les systèmes documentaires traditionnels, puis une analyse systématique est utilisée pour présenter les outils de recherche sur ce nouveau média : moteurs, métamoteurs... ainsi que les approches les plus récentes comme les portails.

L'auteur

Philippe Lefèvre, ingénieur ESE, est chercheur à la division Recherche et Développement d'EDF. Après avoir travaillé sur la reconnaissance de documents ou lecture optique, il s'intéresse depuis quelques années à la recherche d'informations et au traitement automatique du langage naturel.

Hermès
Science
PUBLICATIONS

