Sommaire :

Introduction générale 1

I.Chapitre 1 : La recherche d’informations 2

I.1Introduction 2

I.2La recherche d’information 2

I.2.1 Définitions 2

I.2.2 Concepts de base de la RI 3

I.3Processus d’indexation 6

I.3.1 L’analyse lexicale 7

I.3.2 L’élimination des mots vides 7

I.3.3 La normalisation 8

I.3.4 Le choix de descripteurs 8

I.3.5 La création de l’index 9

I.4Appariement document requête 9

I.5Les modèles de recherche d’information 9

I.5.1 Le modèle booléen 10

I.5.2 Le modèle vectoriel 11

I.5.3 Les modèles probabilistes 11

I.5.3.1 Le modèle probabiliste de base 13

I.5.3.2 Le modèle de langue 14

I.6La reformulation de requêtes 15

I.6.1 Expansion automatique des requêtes 16

I.6.2 Combinaison des présentations des requêtes 16

I.6.3 Réinjection de pertinence 17

I.7Evaluation des performances d’un système d’information 18

I.7.1 Les Collections de test 18

I.7.2 Mesures d'évaluation du SRI 20

I.8Conclusion 24

[II.Chapitre 2 : La pertinence à priori de document 25](#_Toc57331448)

[II.1Introduction 25](#_Toc57331449)

[II.2Définition de la pertinence à priori de documents 25](#_Toc57331450)

[II.3Les caractéristiques utilisées pour le classement à priori des documents 26](#_Toc57331451)

[II.3.1 La structure de liens 26](#_Toc57331452)

[II.3.2 La taille de document 27](#_Toc57331453)

[II.3.3 La date de création de document 28](#_Toc57331454)

[II.3.4 Le rapport information/bruit 28](#_Toc57331455)

[II.3.5 Type d’URL de document 29](#_Toc57331456)

[II.4Le modèle de langue et la pertinence à priori de documents 29](#_Toc57331457)

[II.5Combinaison du score à priori et du score Document/Requête 30](#_Toc57331458)

[II.6Conclusion 31](#_Toc57331459)

IIIChapitre3 : Implémentation en mise en œuvre de l’approche 32

III.1Introduction 32

III.2Description de l’approche 33

III.2.1 Les caractéristiques utilisées pour le calcul de la probabilité de pertinence à priori de document : 33

III.2.1.1 La longueur de document : 33

III.2.1.2 Nombre de termes uniques : 33

III.2.1.3 Moyenne IDF : 33

III.2.1.4 Écart type IDF : 34

III.2.1.5 Écart type TF : 34

III.2.1.6 Rapport TFmin/TFmax 34

III.2.1.7 Entropie 35

III.2.2 L’apprentissage de la probabilité de pertinence à priori de documents : 35

III.2.3 La combinaison des scores : 36

III.2.4 Les outils et langages utilisés : 36

III.2.4.1 Le langage de programmation Java : 37

III.2.4.2 NetBeans : 37

III.2.4.3 Terrier : 38

III.2.4.4 RStudio : 39

III.3Architecture de notre approche : 40

III.3.1 Indexation de documents 41

III.3.2 Calcul des caractéristiques utilisées pour le calcul de la probabilité de pertinence à priori de documents 41

III.3.3 Apprentissage et calcul de scores à priori de documents : 42

III.3.4 Extension du modèle de recherche avec la score à priori de document : 44

III.4Evaluation des résultats 44

III.4.1 La collection de test et les requêtes utilisés : 44

III.4.2 Résultats obtenus avant et après l’extension des modèles de recherche : 44

III.5Conclusion : 46

Conclusion générale :………………………………………………………………………...46

Bibliographie………………………………………………………………………………….48

Liste des tableaux :

Tableau ‎I‑1 :Les mesures de similarités utilisées dans le modèle vectoriel 12

Tableau ‎I‑2 :Exemple de calcul de rappel et de précision pour une requete. 22

Tableau ‎III‑1 : Précisions avant et après l'extension du modèle de recherche 43

Tableau ‎III‑2 : Les requetes améliorées par notre approche avec le modèle de recherche Dirichlet 45

Liste des Figures :

Figure ‎I‑1 : Architecture générale d’un SRI. 4

Figure ‎I‑2 : Indexation d'un document. 7

Figure ‎I‑3 : Taxonomie des modèles de RI [Baeza-Yates et al.1999]. 10

Figure ‎I‑4 : Techniques d'améliorations des SRI par reformulation de requete. 15

Figure ‎I‑5 : Exemple d'un document TREC. 19

Figure ‎I‑6 : Exemple d'une requête TREC. 20

Figure ‎I‑7 : Courbe de Rappel et Précision. 23

Figure ‎III‑1: Extrait du fichier Qrels 35

Figure ‎III‑2:Interface NetBeans de notre approche. 37

Figure ‎III‑3: Architecture générale de notre approche 39

Figure ‎III‑4: Fichier Excel contenant les valeurs des caractéristiques utilisées pour le calcul du score à priori de pertinence de documents 40

Figure ‎III‑5 : Commandes utilisées pour la mise en œuvre de la fonction logistique 40

Figure ‎III‑6 : Résultats de la corrélation entre les variables explicatives et la pertinence 41

Figure ‎III‑7: Coefficients de corrélation entre les caractéristiques et la Pertinence 41

Figure ‎III‑8 : Le fichier contenant les scores à priori de documents 42

Figure ‎III‑9 Analyse requête par requête sur la collection AP88. 44