



République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et d'Informatique

Département Informatique

Projet de fin d'étude pour l'obtention du diplôme Master

Option
Ingénierie des Logiciels

Thème
Data Mining pour la création des communautés
d'utilisateurs
- Cas d'étude le domaine de la recherche scientifique du
CERIST-

Sujet Proposé par :

M^{elle} Nour El Houda Boulkrinat

Présenté par :

M^{elle} BOUCHABOU Aida
M^{elle} KHELLOUF Hassina

Soutenu le : 20/06/2013

Devant le jury composé de:

M ^r DJOUADI	Président.
M ^r BABA ALI	Membre.
M ^r BESSAA	Membre.

Binôme N° : 127/2012

Remerciement

Louange à ALLAH, le miséricordieux, sans Lui rien de tout cela n'aurait pu être. Nous remercions Dieu le tout puissant d'avoir guidé nos pas vers les portes du savoir tout en illuminant notre chemin, et de nous avoir accordé la foi et la force, secret de l'achèvement de notre travail dans de bonnes conditions.

Nos sincères remerciements à notre promotrice Mlle Nour El Houda Boulkrinat de nous avoir fait confiance, pour sa patience et pour tous les efforts qu'elle a fournis durant toutes les étapes de notre travail, ainsi pour toute la peine qu'elle a prise pour nous épauler, nous soutenir, en répondant toujours présente.

Un grand remerciement aux membres du jury, Mr Djouadi, Mr Bessaa et Mme Baba Ali pour nous avoir honorées en consentant à juger notre modeste travail. Nous les remercions aussi pour leurs conseils lors des commissions de suivi.

Merci aux ingénieurs de l'USTHB, Benkherfallah Sabrina et Haddadj Souhila pour leur aide.

*Nous tenons fermement à mentionner le plaisir que nous avons eu à faire notre stage pratique au sein de Centre de recherche sur l'information Scientifique et Technique (**CERIST**).*

Nos remerciements vont aussi à tous les enseignants du département d'Informatique qui ont contribué à notre formation.

Nous remercions également nos familles respectives qui nous ont aidées, encouragées et soutenues dans les moments difficiles tout au long de la préparation de ce mémoire.

Nous tenons à remercier vivement tous ceux qui nous ont aidés de près ou de loin à l'élaboration de ce mémoire.

Table des Matières

Introduction générale.....	1
Chapitre 1 : Recherche d'Information Personnalisée à Base de Profil d'Utilisateur	
1 Introduction	3
2 Les fondements de la recherche d'information.....	3
2.2 Définition	3
2.2 Processus de Système d'Informations	3
2.2.1 Notions de base	4
2.2.2 Principales phases du processus de RI	5
2.3 Les outils de la recherche d'information sur le web	6
2.3.1 Les annuaires ou index ou répertoires	6
2.3.2 Les moteurs de recherche	6
2.4 Taxonomie des modèles de RI.....	7
2.4.1 Le modèle booléen	7
2.4.2 Le modèle vectoriel	7
2.4.3 Le modèle probabiliste	7
2.5 Évaluation des systèmes de recherche d'information	8
3 De la RI classique à la RI adaptative	8
4 La RI adaptative	8
4.1 Reformulation de requêtes	9
4.2 Désambiguïsation du sens des mots de la requête	9
4.3 Regroupement thématique des résultats de recherche	9
4.4 Limitations de la RI adaptative.....	10
5 Personnalisation de l'Information et le Profil Utilisateur	10
5.1 La notion de personnalisation de l'information	10
5.2 Le profil utilisateur.....	11
5.3 La notion de profil.....	11
5.4 Domaine d'utilisation informatique des profils.....	11
5.5 Modélisation du profil utilisateur	11
5.6 Exploitation du profil utilisateur.....	12
6 Les systèmes d'accès personnalisé à l'information	13
6.1 Gestion des profils	14

6.1.1	Représentation du profil	14
6.1.2	Construction du profil	17
6.1.3	Acquisition du profil utilisateur.....	17
6.1.4	Évolution du profil.....	17
7	Communauté d'utilisateur	18
8	Les travaux existants	19
9	Conclusion.....	21
Chapitre 2 : Data Mining		
1	Introduction.....	22
2	Motivation du Data Mining	22
3	Définition du Datamining.....	23
4	Data mining et le KDD.....	23
5	Processus Data Mining.....	24
6	Domaines d'application.....	26
7	Taxinomie des méthodes Data mining	27
7.1	Méthodes supervisées ou prédictive	27
7.1.1	La classification	27
7.1.2	Estimation	27
7.1.3	La prédiction	28
7.2	Méthodes non supervisées ou exploratoires	28
7.2.1	Le groupement par similitudes (Règles d'association)	28
7.2.2	Clusterisation (segmentation)	29
7.2.3	La description (supervisée et no supervisée en même temps)	29
8	Techniques du Data Mining	29
8.1	Analyse du panier de la ménagère	29
8.2	Le raisonnement basé sur la mémoire (RBM)	30
8.3	La détection automatique de clusters.....	30
8.4	Analyse des liens	31
8.5	Les arbres de décision.....	31
8.6	Les réseaux neurones.....	31
8.7	Les algorithmes génétiques.....	31
9	Comparaison entre les méthodes du data mining.....	32
10	Conclusion	33
Chapitre 3 : Conception		
1	Introduction.....	34
2	Etapes de construction de communautés	34
3	Base Profil Existante.....	35

3.1 Architecture fonctionnelle de la partie profil	36
3.1.1 Module d'acquisition	36
3.1.2 Module d'enrichissement.....	36
3.2 Modélisation du profil utilisateur.....	37
3.2.1 Choix d'une représentation du profil.....	37
3.2.2 Description des dimensions de profil utilisateur.....	37
3.3 Structures des données de profils	38
3.3.1 Classes du système	38
3.3.2 Description des classes du système	39
4 Data Mining	40
4.1 Préparation et exploitation des données	40
4.1.1 Compréhension du problème.....	40
4.1.2 Compréhension des données.....	41
4.1.3 Préparation des données	41
4.2 Segmentation des profils	46
4.2.1 Préparation des données	47
4.2.2 Choix de l'algorithme.....	47
4.2.3 Application et validation de l'algorithme de clustering.....	50
4.3 Elaboration de règles d'affectation (classification)	59
4.3.1 Description de l'algorithme	59
4.3.2 Algorithme de classification par k-PPV	59
4.3.3 Définition de la distance	60
4.3.4 Illustration.....	60
5 Les diagrammes UML représentatifs du système	61
5.1 Diagramme des cas d'utilisation	61
5.1.1 Identification des Acteurs	61
5.1.2 Identification des cas d'utilisation	61
5.2 Diagrammes de séquences.....	62
5.2.1 Diagramme de séquences du cas « s'inscrire »	62
5.2.2 Diagramme de séquences du cas « s'identifier »	63
5.2.3 Diagramme de séquence « Gestion de profil »	64
5.2.4 Diagramme de séquence « Accès à l'espace utilisateur »	64
5.3 Diagramme de classe.....	64
5.4 Modèle relationnel du système.....	65
6 Conclusion	65

1	Introduction.....	66
2	Environnement de développement.....	66
2.1	Choix du Système d'exploitation.....	66
2.2	Choix de l'éditeur d'ontologies.....	66
2.3	Choix de l'outil d'exploitation de l'ontologie.....	67
2.4	Choix du Langage de programmation.....	67
2.5	JSP	68
2.6	Apache Tomcat 6.0.26	68
2.7	SGBD Oracle	68
3	Les méthodes du système.....	69
3.1	Classes correspondant à la méthode " K_means "	69
3.2	Classes correspondant au " K plus proches voisins "	69
3.3	Classes correspondant au module "la recherche"	70
3.4	Classes correspondant au module "acquisition de profil utilisateur"	70
4	Description du système D_C_C_U	70
4.1	Interface d'accueil du système D_C_C_U	70
4.2	Interface de description générale sur les communautés.....	71
4.3	Interface de connexion au système D_C_C_U	72
4.4	Interface d'espace utilisateur	73
4.5	Interface de membre de la communauté	74
4.6	Interface d'espace communauté de l'utilisateur.....	74
4.7	L'interface de la recherche avancée.....	75
4.8	Interface résultats de la recherche avancée	75
4.9	Interface d'inscription de l'utilisateur.....	76
4.10	Portion de code de l'algorithme K_means	77
5	Etude comparative	78
5.1	Mode de fonctionnement R 2.15.1	78
5.2	Intégration d'une source de données.....	79
5.3	Exploitation des clusters	80
5.4	Discussion	81
6	Conclusion	82
	Conclusion générale	83
	Annexes	
	Références Bibliographique	