

**Ministère de l'enseignement supérieur et de la
recherche scientifique**

**Institut National de formation en Informatique (I.N.I)
Oued-Smar Alger**

Mémoire de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en informatique

Option : SI (Système d'Information)

**Thème: Conception et réalisation d'un système de
génération automatique des sites portails spécialisés**

**Réalisé par :
NEMAR ABDELLAH
NOUI FODIL**

**Proposé et Encadré par :
M^{me} S.Kouici
Organisme d'accueil :
CERIST**

Promotion: 2005/2006

Sommaire

Introduction Générale

1. Contexte général.....	I
2. Problématique.....	II
3. Organisation du document.....	III

Chapitre I

les outils de recherche sur Internet.

Introduction.....	1
1. Définition de l'Internet.....	1
2. Les outils de recherches sur Internet.....	2
2.1 Les navigateurs.....	2
2.2 Les moteurs de recherche.....	3
2.2.1 Les avantages des moteurs de recherche.....	4
2.2.2 Les inconvénients.....	4
2.3 Les métamoteurs.....	5
2.3.1 Intérêts d'utilisation des métamoteurs	5
2.3.2 Limite des métamoteurs.....	5
2.4 Les annuaires.....	6
2.4.1 Les avantages des annuaires	6
2.4.2 Les inconvénients.....	6
2.5 Les sites portails.....	7
2.5.1 Introduction.....	7
2.5.2 Exemple d'un site portail.....	8
2.5.3 Techniques de génération des sites portails.....	9
a. Approche manuelle.....	9
b. Approche automatique.....	9
Conclusion.....	10

D)	Introduction à la classification.....	11
1.	Contexte de la classification.....	11
2.	Définitions de base.....	13
1.1	Dissimilarité.....	13
1.2	Distance.....	13
1.3	Similarité.....	13
II)	Les méthodes de classification non supervisée.....	14
1.	Les méthodes de partitionnement.....	14
1.1	Définitions.....	14
1.2	Quelques méthodes de partitionnement	15
a.	Algorithme des centres mobiles.....	15
b.	Single-pass.....	15
c.	Les réseaux de neurones artificiels.....	16
1.3	Quelques applications à la recherche documentaire.....	16
2.	Les méthodes de classification hiérarchique.....	18
1.1	Définitions.....	18
1.2	Quelques méthodes de classification Hiérarchique.....	19
a.	Classification Ascendante Hiérarchique.....	19
b.	Les voisins réciproques.....	19
c.	AntTree	20
III)	Choix des méthodes.....	21
1.	Justification du choix de la méthode CAH.....	21
2.	Justification du choix de la méthode AntTree.....	21
IV)	Présentation détaillée des méthodes choisies.....	22
1.	Classification Ascendante Hiérarchique (CAH).....	22
a.	Principe de la méthode.....	22
b.	Choix d'un indice de dissimilarité.....	22
c.	Choix d'indice d'agrégation.....	23
2.	AntTree.....	24
a.	Propriété d'auto-assemblage chez les fourmis réelles.....	24
b.	Principe de la méthode AntTree.....	25
	Conclusion.....	28

1. Introduction.....	29
a. Description du problème.....	29
b. Solution	29
2. Description des procédures.....	31
I) Etape manuelle.....	31
Procédure1 : Collecte des sites Web.....	31
Procédure2 : Filtrage des pages.....	32
a. Définition d'une grille d'analyse.....	32
b. Buts poursuivis par les grilles d'analyse.....	33
Procédure3 : Conception de la base de données.....	35
a. Représentation des pages.....	35
b. Création de la base de données.....	37
b.1 Règles de gestion.....	37
b.2 Dictionnaire de données.....	37
b.3 Modèle conceptuel de données.....	38
b.4 Modèle logique de données.....	40
b.5 Modèle relationnel.....	40
II) Etape automatique.....	41
Procédure4 : Représentation mathématique des pages Web.....	41
a. Modèle vectoriel.....	41
a.1 Modèle vectoriel booléen.....	42
a.2 Modèle vectoriel avec fréquence des mots (TFIDF).....	42
b. Sac de mots.....	43
c. Analyse sémantique latente.....	43
Procédure5 : Classification des données.....	44
a. Classification Ascendante Hiérarchique	44
b. AntTree.....	46
Procédure6 : Interface utilisateur.....	49
Conclusion	51

I.	Environnement de mise en œuvre	52
1.	Choix du langage de programmation	52
2.	Choix des logiciels du système	53
II.	Utilisation du système	55
III.	Tests et résultats	60
1.	Jeux de données	60
2.	Méthodologie d'évaluation	60
3.	Résultats des tests	61
4.	Etude Comparative.....	63
	Conclusion.....	63

Conclusion et perspectives**Annexe : Métadonnées et Dublin Core****Bibliographie**

Résumé

Au début d'Internet, l'essentiel des outils de recherche étaient de type généraliste. Leur but fut de faciliter l'accès au grand nombre de ressources disponibles. Par la suite, la croissance phénoménale de l'information offerte sur le Web, a engendré l'apparition d'un nouveau moyen de recherche qui est les sites portails spécialisés. La construction de ces derniers requiert un effort considérable en temps et en ressources humaines impliquées. Ainsi, le principal objectif de ce travail est la conception et la réalisation d'un outil informatique permettant la génération automatique de sites portails. Pour le faire deux méthodes de classification automatique sont utilisées, à savoir, la méthode CAH (Classification Ascendante Hiérarchique) et la méthode AntTree.

Mots clés : Site portail, Classification Automatique, Classification Ascendante Hiérarchique, CAH, AntTree.

Abstract:

In the beginning of Internet, most of search tools were of general type. Their aim was to facilitate the access to the big number of available resources. Nevertheless, following the phenomenal growth of pieces of information offered on the Web, appeared a new search tool which is the specialized portal sites. But, unfortunately, the construction of the portal sites requires a considerable effort even in time and human resources used. That's why we have the idea to conceive automatic methods of portal sites generation. To do that, two clustering methods will be used: HAC (Hierarchical Ascendant Clustering) and AntTree.

Keywords: Portal site, Automatic Clustering, Hierarchical Ascendant Clustering, HAC, AntTree.