République Algérienne Démocratique et Populaire.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

Université Saad Dahlab, Blida
USDB.

Faculté des sciences.
Département informatique .

Mémoire pour l'obtention
d'un diplôme d'ingénieur d'état en informatique.
Option : **Système d'Information**

Sujet :

Conception et réalisation d'un moteur de recherche pour le réseau Intranet

Présenté par : Mr Mansour Samir

Promoteur : Melle BOUSTIA Narhimène

Organisme d'accueil : USDB.

Soutenue le: **Juin , devant le jury composé de :**

M. Bala	**Président**
M. Benouar	**Examinateur**
M^me^ Abed	**Examinatrice**

- /2005 -

Mémoire de fin d'étude
Pour l'obtention du diplôme d'ingénieur d'état en
GENIE : Informatique

Présenté par :

-Mansour Samir

Thème

Conception et réalisation d'un moteur de recherche pour le réseau Intranet

Dirigé par :

- Melle BOUSTIA Narhimène

Année : 2005

Remerciements

Nous tenons à adresser nos sincères remerciements à notre promoteur Melle BOUSTIA Narhimène qui a su nous conseiller par ses critiques constructives et nous guider dans notre modeste travail.

En deuxième lieu, nous tenons à remercier l'ensemble des professeurs du Département Informatique pour avoir assurer notre formation et l'environnement adéquat afin de réaliser notre travail.

Nos remerciements vont aussi à tous ceux et celles qui ont participé de près ou de loin à l'élaboration du présent travail.

Nos sentiments de profonde gratitude vont à nos professeurs qui tout au long des trois années d'études nous ont transmit leur savoir sans réserve.

Enfin, nous tenons à remercier tous nos amis et collègues pour leur soutien moral tout au long de la préparation de ce mémoire.

Samir

Dédicaces

Avec les sentiments de gratitude les plus profonds,
je dédie ce modeste travail aux deux êtres qui me sont les plus chers;
Ma mère et Mon père
qui n'ont pas cessé de prier pour moi et de m'encourager dans les moments difficiles.

A ma fiancée, futur épouse et mère de mes enfants incha'allah
Ainsi qu'à toute sa famille

A tous mes frères, à toutes mes sœurs ainsi qu'à leurs familles.
A tous mes amis ; Kheiredinne, Adel, Hafied, Rachid, …..
A tous ceux qui ont éclairé ma vie avec leur amour (El 3achra).

Samir

SOMMAIRE

INTRODUCTION

Dans les entreprises modernes, l'information est devenue une ressource vitale. Afin de faciliter le stockage, le traitement et l'échange de l'information au sein de ces entreprises, l'utilisation des systèmes d'informations a de nos jours une importance incontournable.

Au file des années, la quantité d'information échangée s'accroît et le nombre de documents contenus dans les systèmes d'informations se multiplie. Ce qui est devenu un phénomène de plus en plus inquiétant, du fait qu'il peut dévier l'objectif pour lequel les systèmes d'information ont été mis en œuvre, car cet ensemble de documents possède l'inconvénient de noyer l'information. De ce fait, la classification, la localisation et l'accès aux documents relèvent pratiquement de l'impossible.

Les informations sont souvent disponibles, mais inaccessibles et les demandes deviennent de plus en plus exigeantes: facilité d'accès aux informations, rapidité de la recherche et pertinence des résultats obtenus. Cela conduit à l'élaboration de méthodes de recherches performantes qui assurent des résultats de bonnes qualité et des délais d'attentes très raisonnables. En plus, dans un environnement professionnel d'entreprise, l'échange d'informations doit être contrôlé car certaines d'entre elles ne doivent être délivrées qu'à un nombre restreint d'utilisateurs pour des raisons de confidentialité, cela mène à imposer des restrictions d'accès aux informations disponibles .

Pour répondre aux besoins cités précédemment, les systèmes d'information doivent être dotés d'un système de recherche ***documentaire*** puissant et performant, permettant un accès efficace à l'information, non au document en tant que fichier, mais à son contenu lexical et sémantique (les mots clés, les concepts sémantiques...). Un tel système résout les problèmes liés à la recherche séquentielle qui oblige l'utilisateur à parcourir toute la base d'information (ou le corpus) pour trouver les informations désirées.

Notre travail consiste à concevoir et à réaliser un moteur de recherche de documents partageables dans un réseau Intranet. Le but principal de ce moteur est d'amélioré les systèmes d'information en facilitant l'accès à l'information. Pour cela, nous devons étudier les différentes techniques d'***indexation*** utilisées par les moteurs de recherche d'Internet (techniques sémantiques, statistiques, texte intégral ...), afin d'avoir un bon système de stockage qui assure la rapidité d'accès aux informations, ainsi que les techniques de ***recherche*** (recherche booléenne, textuelle...) et de ***présentation*** des documents réponses pour garantir une bonne qualité des résultats en terme de quantité et de pertinence.

Le présent mémoire est organisé en cinq chapitres :

Le premier décrira les différents outils de recherche utilisés sur le web, ainsi que leurs avantages et inconvénients. Nous donnerons une brève comparaison entre ces outils en mettant l'accent particulièrement sur les moteurs de recherche, leur mode de fonctionnement et les méthodes utilisées dans leur développement.

Le second chapitre, présentera dans le détail les principales techniques d'indexations des documents. Il se terminera par une critique de ces techniques en citant leurs avantages et inconvénients afin de faire le choix le plus adapté pour nos besoins.

Le troisième chapitre sera consacré à la recherche et la présentation des résultats. Les différents modèles et techniques de recherche y seront exposés ainsi que les informations concernant la présentation des documents résultats.

Enfin les deux derniers chapitres seront consacrés à la conception et la mise en œuvre du moteur de recherche.

Une conclusion générale viendra pour résumer ce mémoire afin de présenter les résultats obtenus et les perspectives qui peuvent servir à l'amélioration du système.

Chapitre I

LES OUTILS DE RECHERCHE D'INFORMATIONS

Dans ce chapitre :

- Les outils de recherche d'information
- Comparaison entre les outils de recherche
- Présentation détaillée d'un moteur de recherche
- Conclusion

Les réseaux d'entreprises évoluent sans cesse et sont en train de s'imposer comme un outil important dans le monde du travail du fait de l'accroissement du nombre de ses utilisateurs, du nombre d'ordinateurs connectés et de la quantité d'informations disponible. Cette quantité d'informations en tout genre rend parfois difficile la navigation et la recherche rapide. C'est pourquoi il est impératif d'introduire la notion d'outils de recherche permettant de lancer une requête et de trouver ainsi les documents susceptibles de renfermer l'information désirée par l'utilisateur. La recherche d'information par l'intermédiaire de ces outils spécifiques permet un gain de temps considérable.

Sur le web, Il existe à l'heure actuelle trois grandes familles d'outils de recherche :

- Les annuaires ou répertoires thématiques.
- Les moteurs de recherche.
- Les méta-moteurs.

I.1 LES OUTILS DE RECHERCHE D'INFORMATION

I.1.1 Les annuaires ou répertoires thématiques

I.1.1.1 Présentation d'un annuaire

Un annuaire ou répertoire est un outil de recherche dont les sites proposés sont vérifiés par des humains et classés selon des catégories appropriées. L'inscription d'un site dans un annuaire pour figurer dans la base de données se fait au moyen d'un formulaire où doivent être inscrites plusieurs informations dont une description du site, l'adresse (URL), le titre ainsi que la catégorie dans laquelle le propriétaire du site désire qu'il figure.

Les éléments requis lors d'une demande d'inscription sont très importants car la recherche au moyen de mots clés est basée sur ces éléments (adresse Web, titre, description) et non sur le contenu des pages du site en question. Il ne s'agit donc pas d'une indexation automatique effectuée par un Robot (programme informatique qui scrute le réseau afin de localiser les documents et les indexer dans une base de données), mais d'un référencement humain et volontaire sollicité par le titulaire du site lui-même **[OSMA 99]**.

Les principaux répertoires qui existent sur Internet sont: **La Toile du Québec, Yahoo, Open Directory Project, Looksmart, Galaxy, NetGuide.**

I.1.1.2 Avantages attribués aux annuaires

- Lors d'une recherche, une sélection de sites correspondant à une catégorie précise est rapidement obtenue.
- Comme le contenu des sites a été vérifié par des humains, il y'a moins de chance d'obtenir des résultats erronés.
- Une qualité significative des sites répertoriés.

I.1.1.3 Inconvénients attribués aux annuaires

- Les sites répertoriés doivent être inscrits.
- La mise à jour de la base de données est plus longue.
- Classement des résultats de la recherche par ordre alphabétique et non par pertinence des documents trouvés.
- Inscrire son site à un annuaire ne garantit aucunement son acceptation car les critères d'acceptations sont souvent sévères.

I.1.2 Les moteurs de recherche

I.1.2.1 Présentation d'un moteur de recherche

Un moteur de recherche est une immense base de données continuellement mise à jour par des "Robots" ou "Spiders" qui scrutent le Web en allant de page en page et qui les sauvegardent dans son index. Son fonctionnement est donc totalement automatisé.

Le rafraîchissement de la base se fait selon certains délais. Une fois les pages repérées, le moteur de recherche classe les pages par ordre de pertinence, selon un ordre et un algorithme basé sur des critères de tri qui lui sont spécifiques. C'est à la pertinence des résultats obtenus que l'on juge la qualité d'un moteur de recherche **[FORE 03]**.

Parmi les moteurs de recherche les plus connus, on peut citer : **Google, Altavista, HotBot, Excite,...** etc.

I.1.2.2 Avantages attribués aux moteurs de recherche

- Les recherches donnent plus de résultats car la base de données d'un moteur de recherche est très importante (par exemple La base de données de Google est stockée dans dix serveurs ou datacenters).
- La base de données est mise à jour fréquemment à intervalle régulier.

- Le classement des résultats des recherches est effectué par pertinence et non par ordre alphabétique.

I.1.2.3 Inconvénients attribués aux moteurs de recherche

- Les recherches peuvent générer une grande masse de résultats (bruit)
- Comme le contenu des sites n'est pas examiné par des humains, la qualité des résultats peut être moindre.
- Les moteurs de recherche n'interrogent pas directement le Web mais leurs bases de données contenant les termes décrivant chaque page Web ce qui inclut que les moteurs peuvent difficilement suivre les mises à jour des sites où l'information change quotidiennement et donc ne peuvent assurer une indexation totalement fiable d'un environnement dynamique comme celui du Web.

I.1.3 Les méta-moteurs

I.1.3.1 Présentation d'un méta-moteur

Le principe d'un méta-moteur est de permettre l'interrogation simultanée de plusieurs indexes de moteurs de recherche différents. La saisie de la requête s'effectue à travers une interface unique qui peut être accessible via un site Web ou via un logiciel qu'il est nécessaire d'installer sur un poste client. La requête est alors soumise aux différents moteurs de recherche interrogés dont le nombre varie de quelques-uns à plusieurs dizaines selon les méta-moteurs considérés. Les réponses provenant de ces différents moteurs subissent le plus souvent une élimination des doublons et sont présentées par ordre de pertinente à l'usager, ce dernier peut se rendre sur chaque page proposée, via un lien hypertexte. Les principaux méta-moteurs actuels sont: **infind**, **debriefing**, **dogpile**, **eo** ... etc.

I.1.3.2 Avantages attribués aux méta-moteurs

- Augmentation de la taille de l'index interrogé.
- Résultats de recherches beaucoup plus importants

I.1.3.3 Inconvénients attribués aux méta-moteurs

- Même si la plupart des méta-moteurs assurent une traduction de la requête pour l'adapter à la syntaxe de chacun des moteurs interrogés, l'utilisation de requêtes complexes entraîne le plus souvent des réponses très bruitées.
- Le temps de recherche est plus long.
- Difficulté d'intégration des résultats retournés par différents moteurs utilisant des méthodes et des techniques différentes.

I.2 COMPARAISON ENTRE LES OUTILS DE RECHERCHE

A travers les avantages et les inconvénients énumérés pour chacun des outils de recherche cités précédemment, il devient aisé d'établir une comparaison et de définir les principales différences entre ces outils. Ces différences ont des incidences sur les résultats de la recherche, ainsi, la connaissance de ces outils est nécessaire pour l'obtention de meilleurs résultats lors d'une recherche sur le Web (Internet ou Intranet).

Les annuaires sont généralement utilisés pour effectuer une recherche sur un sujet général, c'est-à-dire au début du processus de recherche, pour cerner le sujet et retrouver les sites de référence sur ce dernier. Les annuaires sont alors associés à une recherche thématique.

A la différence des annuaires, les moteurs de recherche indexent des pages Web, selon une approche totalement automatisée et sont généralement utilisés pour trouver des informations sur un sujet précis dont les termes clés sont déjà connus, on parle alors de recherche par interrogation via l'utilisation de mots clés. Chaque moteur de recherche possède un système de classement différent. Les résultats affichés par les moteurs sont des pages Web tandis que ceux provenant des annuaires sont des sites.

En fait, certains outils de recherche présentent à la fois une fonction de répertoire et de recherche automatique. De plus en plus, certains répertoires se sont associés avec un moteur de recherche qui fournit des résultats non trouvés dans la base de données (Le moteur de recherche Google intègre le répertoire Dmoz depuis mars 2000 **[FORE 03]**).

Les moteurs de recherche sont des outils très importants qui permettent de faciliter la recherche d'informations sur le Web, cependant il est nécessaire de souligner que les moteurs ne permettent pas un recouvrement total du Web et n'en recouvrent tous que 60% (Google:

18.6%, AltaVista: 37.1%, HotBot: 27.1%... etc. **[LAWR 98]**) et donc ne s'adaptent pas à l'évolution du Web.

Les méta-moteurs permettent d'adresser simultanément une même requête à différents moteurs de recherche pour augmenter le nombre de résultats. Les méta-moteurs sélectionnent les réponses en fonction de leur pertinence. Ils peuvent être considérés comme une évolution des moteurs de recherche, cependant ils souffrent d'un temps d'interrogation plus long et d'une difficulté d'interrogation des résultats provenant de sources d'information hétérogènes.

Nous nous intéressons dans notre étude tout particulièrement aux moteurs de recherche qui peuvent être facilement intégrés dans un contexte Intranet, en les dotant d'une interface d'indexation automatique ainsi qu'une autre interface pour l'indexation manuelle des documents fournis par leurs auteurs afin de les référencer auprès de notre moteur de recherche. Dans un intranet, vu que la densité de l'information circulante est moins intense qu'Internet et le nombre des utilisateurs et des documents est beaucoup plus réduit, il est préférable de fournir un outil hybride qui profite des avantages de toutes les technologies de recherches existantes.

I.3 PRESENTATION DETAILLEE D'UN MOTEUR DE RECHERCHE

Qu'est ce qu'un moteur de recherche ?

C'est un outil qui permet d'extraire d'une information, principalement textuelle, les mots ou les termes qui la représentent au mieux et de les stocker dans un index. Le même outil parcourt ensuite cet index afin d'identifier les termes les plus pertinents par rapport à ceux de la question de l'utilisateur, puis de trier les résultats à lui fournir **[BENA 00].**

En général, le fonctionnement d'un moteur de recherche comporte les phases suivantes :

- **Indexation** : C'est l'analyse des documents collectés par extraction des mots et des concepts qui caractérisent leur contenu sémantique. Ces mots seront ensuite mis dans un index pour faciliter la recherche. Cette phase n'est pas visible aux utilisateurs.
- **Recherche** : C'est le premier plan par rapport à l'utilisateur. Après la formulation et l'analyse de la requête, la recherche se fera au niveau de l'index, le résultat sera la liste des documents correspondants à la requête.
- **Présentation et Consultation** : Avant de présenter à l'utilisateur la liste des documents répondants aux critères de sa requête, une phase de classification de cette liste par ordre de pertinence est nécessaire pour aider l'utilisateur à choisir les documents à consulter.

Pourquoi l'indexation ?

C'est la phase la plus importante dans la vie d'un document car elle permet de le représenter sous un formalisme informatique. On peut dire que c'est la matière grise d'un moteur de recherche car l'efficacité de ce dernier dépend principalement de cette phase.

Pourquoi la recherche ?

C'est la phase de récupération de l'information enfouie dans la base de données et difficile à atteindre, en mobilisant toute la puissance et la rapidité d'un ordinateur. Sans cet outil de recherche, l'être humain sera contraint de sacrifier beaucoup de temps pour avoir l'information dont il a besoin. La figure suivante montre le fonctionnement général d'un moteur de recherche avec toutes ses phases **[BENA 00]**.

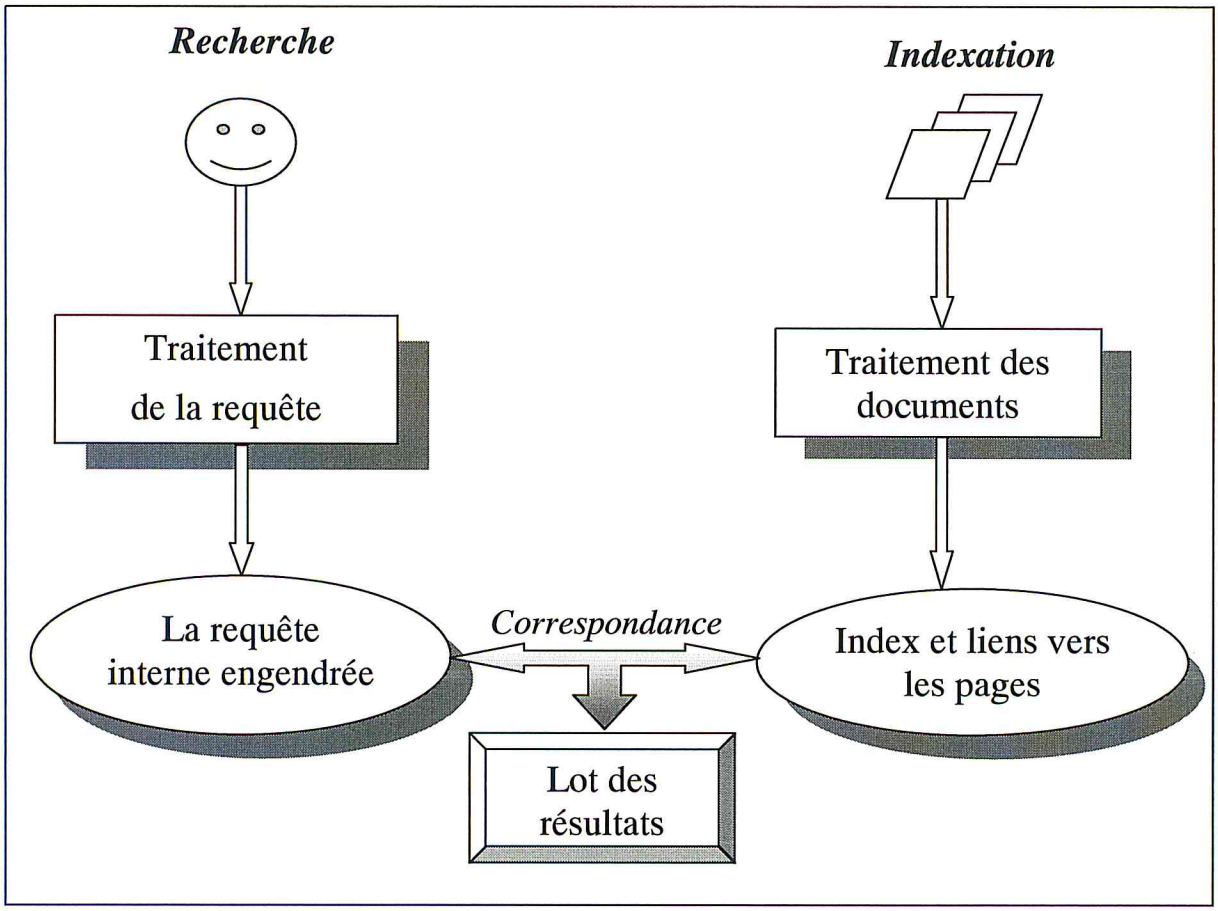


Figure I.1 Différentes phases du fonctionnement d'un moteur de recherche

L'objectif essentiel d'un système de recherche documentaire est de sélectionner des documents en réponse à une requête. Pour ce faire, il met en œuvre un mécanisme de comparaison entre une requête et l'ensemble des documents contenus dans le fond. L'organisation du système de recherche documentaire consiste à trouver une représentation intermédiaire des documents et des requêtes capables de favoriser ce mécanisme de comparaison.

I.3.1 Architecture d'un moteur de recherche

Afin d'assurer le rôle d'un moteur de recherche et couvrir tous les aspects liés à la recherche documentaire, il est principalement constitué de trois parties :

- Le robot (appelé aussi spider)
- L'index ou la base de données.
- Le logiciel/interface d'interrogation.

➢ Le robot

Le robot ou encore "spider" est la partie la plus importante du moteur de recherche, car celui-ci effectue la recherche directement sur le Web pour en extraire le plus grand nombre d'informations relatives aux documents présents sur le Web et les indexer au sein de sa base de données.

Un moteur de recherche utilise un robot qui balaie la structure hypertexte du Web par suivi récursif des liens pour en archiver intégralement son contenu. Il assure ainsi la lecture des données des pages Web et le repérage des liens pointant vers d'autres pages afin de constituer l'index. **[KOST 95]**

➢ L'index

L'index est le lieu de stockage et d'indexation des pages Web visitées par le robot, c'est la phase de structuration et de classification de l'information rapatriée. En général, les informations répertoriées sont : l'adresse (URL), le titre des pages, les mots clés voir même l'intégralité des pages. L'entrée de pages Web dans l'index est également rendue possible par la visite du robot pages soumises volontairement par leurs créateurs. Afin d'actualiser le contenu des pages Web dans l'index, le robot se rend à intervalle défini sur celles-ci, pour mettre à jour leur contenu dans l'index.

Les méthodes d'indexations diffèrent d'un moteur à l'autre influençant ainsi les résultats obtenus suite à une requête. Une fois les documents collectés par les robots, les moteurs d'indexations utilisent différentes méthodes pour indexer le contenu des pages, parmi les plus courantes:

- Indexation complète du texte (Full-Text)
- Indexation Manuelle par l'intermédiaire d'un fichier de soumission à compléter par tout opérateur de site désirant faire figurer ses pages Web dans la base de données du moteur.
- Indexation de balises spécifiques (les balises Méta, Title des pages HTML)
- Indexation statistique consistant à supprimer les mots de sens vide et à ne retenir que ceux qui dépassent un seuil de fréquence d'apparition (Scoring).

➢ **L'interface**

L'interface permet à l'utilisateur de saisir sa requête en utilisant un ou plusieurs termes significatifs qui seront ensuite recherchés dans l'index. L'interface sélectionne parmi les milliers de documents enregistrés dans l'index ceux qui satisfont cette requête et les propose sous forme d'une liste de pages Web énumérées selon un ordre de pertinence décroissant, pages sur lesquelles l'utilisateur peut ensuite se rendre via un lien hypertexte.

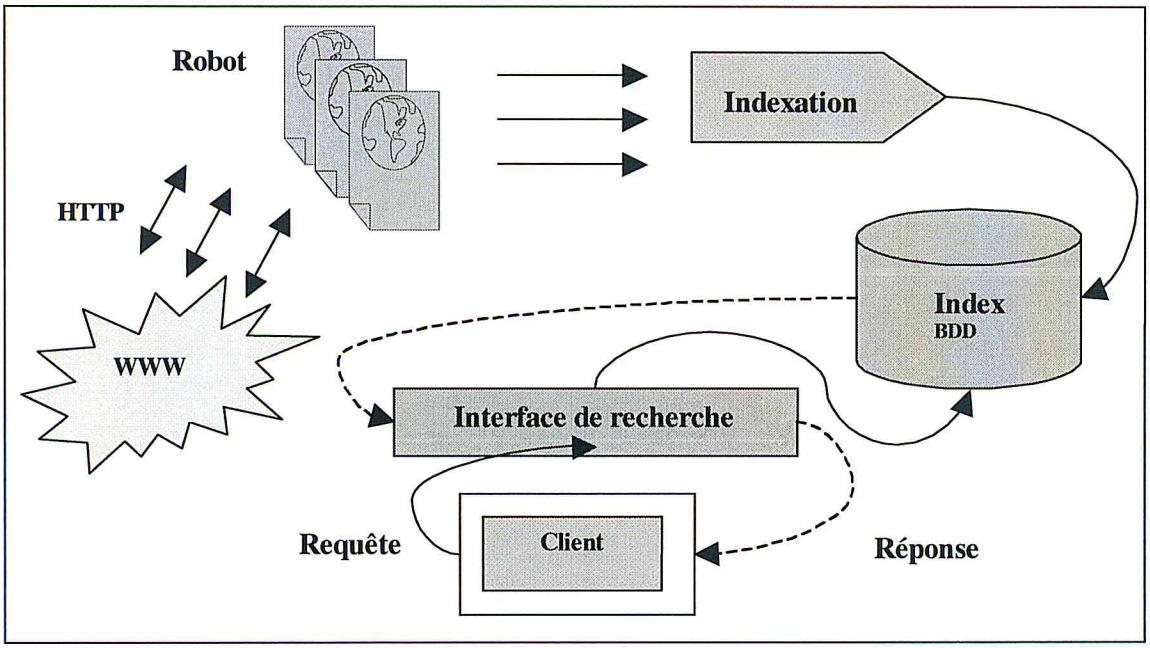


Figure I.2 Architecture générale d'un moteur de recherche

I.3.2 Fonctionnement d'un moteur de recherche

Le mode de fonctionnement d'un moteur de recherche peut être schématisé de la manière suivante **[MICH 03]** :

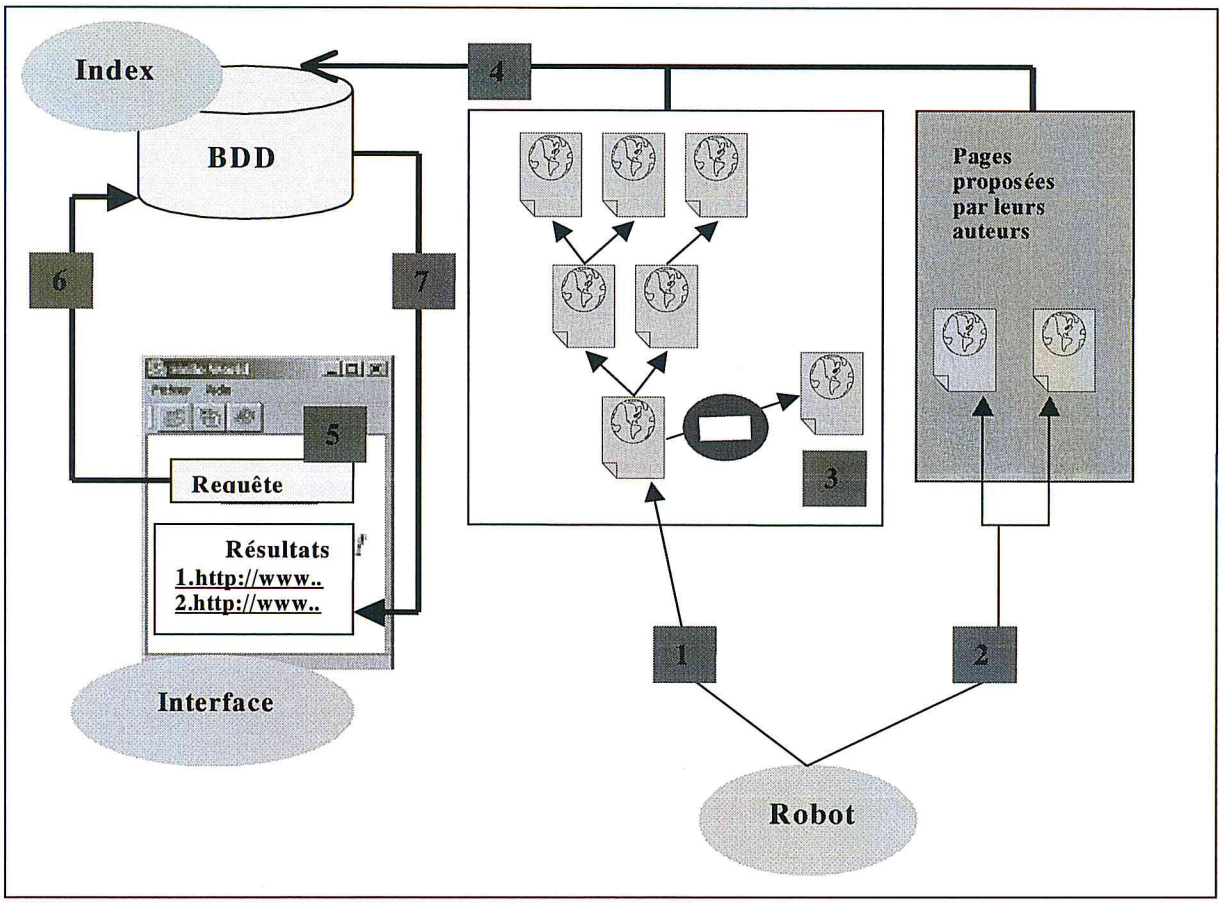


Figure I.3 Fonctionnement d'un moteur de recherche

Le robot repère les pages Web par suivi récursif des liens appartenant aux pages présentes sur le réseau **(1)** ou proposées par les auteurs **(2).** Toutefois, les pages à accès réservé ne seront pas indexées **(3)** alors que les autres seront indexés au sein de la base de données **(4)**. La requête est émise par l'utilisateur à travers l'interface de recherche **(5)** dont les termes seront recherchés dans l'index (la base de données) **(6)**. Les résultats sont ensuite affichés à travers l'interface sous forme de liens hypertextes **(7)**.

I.3.3 Domaines d'utilisations des moteurs de recherche

Les moteurs de recherche sont utilisés dans toutes les applications où il est nécessaire d'accéder au contenu de l'information tant pour la recherche que pour le filtrage ou la diffusion sélective. Parmi ces applications, nous pouvons citer celles de gestion de bibliothèques et de centres de documentation, la recherche dans un ensemble de sources d'informations qui s'intéressent à un domaine précis, par exemple un petit moteur de recherche fournis avec un site web spécialisé (informatique, médecine, théories d'algèbre ... etc.), qui cherche dans le lot des documents appartenant au site un sujet précis que demande l'utilisateur.

I.3.4 Les restrictions d'accès à l'information

Une autre fonctionnalité en plus de celles décrites plus haut, est la confidentialité des informations retournées. Il faut, cependant, garder à l'esprit que toutes les informations ne doivent pas être systématiquement accessibles à tous. Certaines informations doivent même rester confidentielles ou être accessibles à un nombre restreint d'utilisateurs. D'une façon générale, l'entreprise établit des critères d'accessibilité à chaque information ou ensemble d'informations, tels que des documents, des bases de données... etc. La gestion de l'information sous forme électronique doit également tenir compte de ces paramètres.

I.3.5 Le Robot (Spider)

Le "spider" ou encore le robot est la partie la plus importante du moteur de recherche, car celui-ci effectue la recherche directement sur le réseau Intranet pour en extraire le plus grand nombre d'informations relatives aux documents présents et les indexer au sein de sa base de données.

Un robot est un programme qui traverse la structure hypertexte du réseau Intranet pour retrouver un nombre maximum de documents, par un suivi récursif des liens, en utilisant des protocoles standards du Web et tout en faisant abstraction des sites qui n'autorisent pas la visite automatique des spiders.

Afin d'appréhender le robot, nous le présentons comme suit:

- La composition du robot.
- Le fonctionnement général du robot.

1. Le robot est déclenché avec une plage d'adresses (adresses IP des machines appartenant au réseau Intranet), dont le but est la recherche d'éventuels serveurs (HTTP, FTP).
2. Le robot télécharge la page d'accueil de chaque serveur et l'insère dans sa base de données.
3. Chaque page est ensuite analysée afin d'en extraire les liens hypertextes (pour les ajouter dans la base pour une analyse ultérieure), et les informations caractérisant cette page.

I.4 Conclusion :

La conception d'un robot tourne autour de trois fonctions élémentaires : Le crawling, l'analyse et l'indexation. Cette dernière permet d'indexer des documents selon des méthodes d'indexation différentes mais néanmoins adaptées au contenu et à la structure de chacun des types précédemment cités. il est utile et nécessaire de maîtriser les avantages et les **inconvénients** de chaque méthode. Le chapitre suivant détaillera ces différentes techniques

Chapitre II

L'INDEXATION

Dans ce chapitre :

- ❖ Définition
- ❖ Les composants de l'indexation
- ❖ Les techniques d'indexation

II.1 DEFINITION

L'indexation est un processus qui permet de représenter un document sous un autre aspect pour le rendre manipulable et exploitable à une recherche ultérieure, et cela à partir d'une analyse lexicale, syntaxique ou sémantique, en sélectionnant les mots ou les concepts représentant le contenu sémantique du document.

L'indexation est l'opération centrale de tout système documentaire. Elle consiste à analyser, lors de l'organisation du fond documentaire, les documents afin de produire un ensemble de mots clés (ou descripteurs). Cet ensemble, constituant un langage d'indexation, est organisé dans un index.

II.2 LES COMPOSANTS DE L'INDEXATION

L'indexation dispose, en plus du document à indexer, d'un ensemble d'objets qu'elle manipule, nécessaires à la construction de l'index final. Parmi ces éléments, nous citons :

Les mots clés : Ce sont des mots simples, des expressions composées d'un ou de plusieurs mots (Uni-termes ou multi-termes) qui décrivent au mieux le contenu du texte pour pouvoir l'indexer ultérieurement.

L'index : L'index est une représentation synthétique de l'information relative à un document qui met en évidence sa sémantique en vue d'une recherche ultérieure par une requête. L'index comprend une entrée pour chaque mot clé utilisé dans l'ensemble des documents : à cette entrée sont associés les clés (références) des documents contenant le mot correspondant avec, éventuellement, un poids **[LELO 98].** Cet index doit être trié par ordre alphabétique pour permettre une lecture rapide. Le schéma suivant montre les caractéristiques de l'index utilisé par la méthode du texte intégral que nous allons décrire plus loin dans ce chapitre.

I.3.5.1 Composition du robot

Un robot est généralement composé de trois modules effectuant chacun une fonction déterminée :

- **Crawler**

Le crawler à pour rôle de télécharger des documents à partir d'adresses URL en utilisant plusieurs techniques pour télécharger le plus grand nombre de pages sur une période de temps minimale.

- **Analyseur**

L'analyseur a pour tâche l'extraction des liens hypertextes présents dans la page, ainsi que l'extraction des mots nécessaires à une bonne indexation selon le type de document à indexer (HTML, PDF, DOC...).

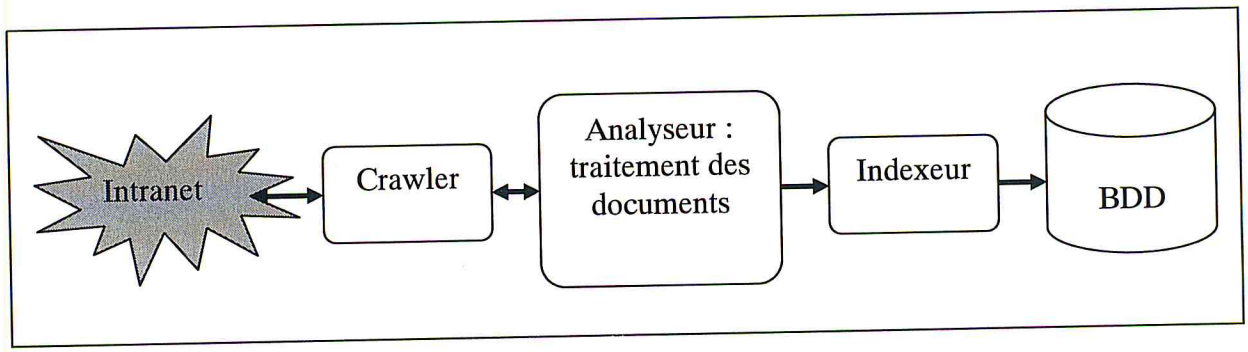
- **Indexeur**

L'indexeur finit par récupérer l'URL du document ainsi que le résultat de son analyse intégrale (ensemble de mots) afin de les insérer dans la base de données.

Le résultat final de notre robot sera une base de données qui peut être consultée par les utilisateurs à travers une interface d'interrogation.

I.3.5.2 Fonctionnement générale du robot

D'après les trois fonctions citées précédemment, une conception générale du robot peut-être donnée dans la figure suivante :



Fonctionnement générale du robot

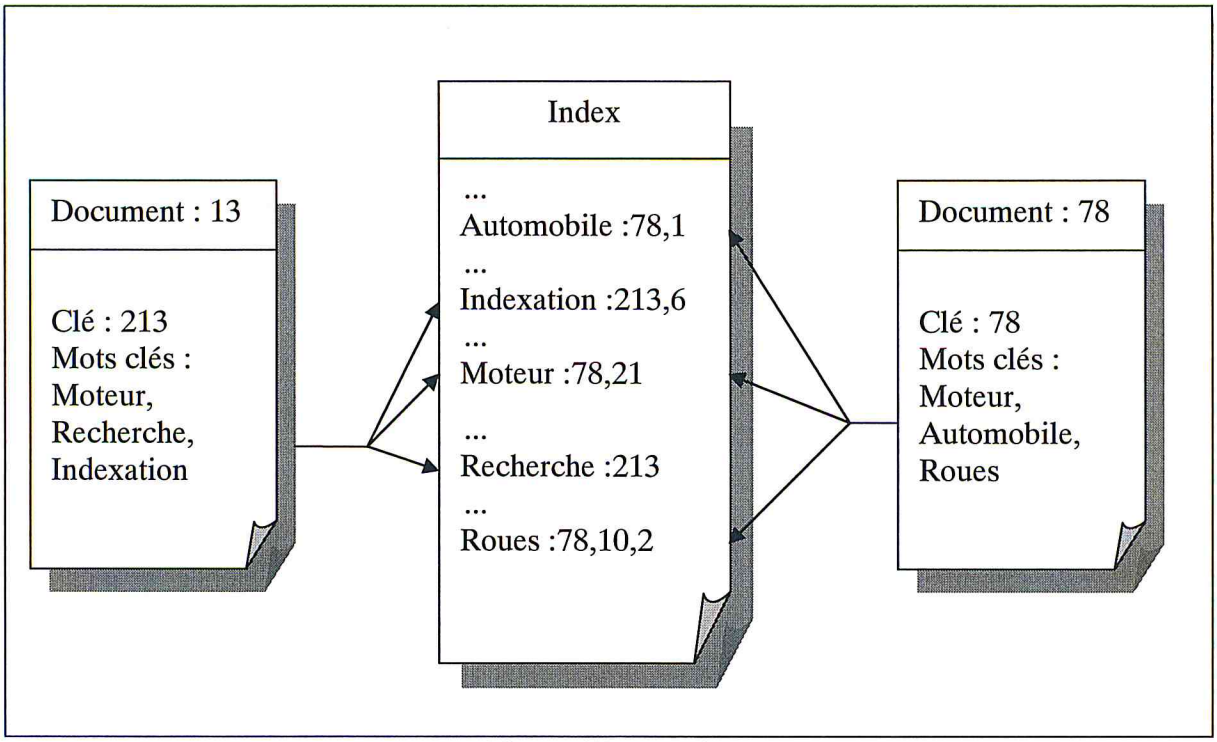


Figure II.1 l'Index

Lors de l'indexation, il faut tenir compte des jeux de caractéres de la langue avec laquelle le document est écrit et éliminer les mots vides qui n'ont aucun role symantique dans le texte.

❖ Les jeux de caractères de langues

La première caractéristique de l'indexation est évidemment le ou les jeux de caractères supportés. Les Anglo-saxons n'ont pas de difficultés d'écriture, car leurs textes sont pauvres en accents et ont très peu de caractères spéciaux, contrairement aux autres langues comme le Français (l'accent), l'Espagnol (le tilde) et d'autres : l'Arabe, le grec... etc. Plusieurs normes définissent les jeux de caractères, **l'ISO 8859.1** est la norme couramment utilisée aujourd'hui, notamment sur le WEB, par HTML2.0, HTML3.2 et dernièrement HTML 4.0 **[BENA 00]**.

❖ Les mots vides

Les mots vides ou ***stopwords*** sont des mots qui ne jouent qu'un rôle syntaxique, apportant peu de sens aux documents. Donc, il ne serait pas nécessaire de les indexer, car il en existe beaucoup dans les langues (ex : le, la, de, des, alors, lequel...), leur suppression ne modifie pas le concept sémantique du texte. Ils sont décomposés en deux classes :

- **Les mots outils** articles, prépositions, conjonctions de coordination ou de subordination, pronoms personnels, relatifs ou possessifs, etc.
- **Les mots ordinaires** tels que les mots très fréquents, les auxiliaires et leurs formes conjuguées, certains verbes, etc.

Lors de la construction des indexes, les mots vides sont éliminés pour deux raisons :

1. Minimiser la taille du fichier index (contrainte : espace)
2. Rendre la recherche et l'indexation plus rapides (contrainte : temps)

II.5 LES TECHNIQUES D'INDEXATION

Il existe aujourd'hui plusieurs familles pour la recherche par le contenu :

- La plus ancienne, est celle de la technique du fichier inverse ou texte intégral, qui se fait par une simple recherche de mots dans un texte, sans aucun traitement sémantique ou syntaxique sur ces mots.
- Celle, qui fait l'objet de nombreuses recherches depuis quelques années, grâce à la puissance des calculateurs, est la technique linguistique, qui s'appuie sur l'analyse du langage pour une recherche intelligente. En analysant non seulement des mots mais des termes, et parfois des expressions, voire des concepts contenus dans le texte.
- La troisième s'appuie sur des techniques qui ne sont pas spécifiques au sens du texte, mais sur une recherche de ressemblance entre la question posée et le corpus textuel disponible dans la base. Et ce sont les méthodes statistiques. Ces techniques reposent sur les statistiques pures, reconnaissance de formes, classification et algorithmes génériques.
- La dernière tire les avantages des deux techniques *linguistique* et *statistique*. On traite les documents par des méthodes linguistiques ensuite les résultats seront filtrés statistiquement.

Quelle est la meilleure technique ?

Il n'y a pas vraiment une technique meilleure que les autres, car chacune a son propre champ d'application et traite un aspect dans l'évolution de l'indexation d'un document. Les méthodes d'indexations sont le sujet de nombreuses recherches. En effet, compte tenu de la masse d'informations brutes fournies par le Web, l'absence totale de structure et le multilinguisme, les outils ainsi que les méthodes actuelles atteignent leurs limites. Il y a donc une recherche florissante autour de ce thème.

II.5.1 Texte intégral ou fichier inverse

Elle est la plus utilisée parmi toutes les méthodes, ceci est dû à sa simplicité de conception et de réalisation. Elle consiste à employer pour l'indexation tous les mots lexicaux

(noms, adjectifs, verbes, ...etc.) par élimination des mots inutiles ou vides (un anti-dictionnaire est indispensable) **[LELO 98]**.

Dans ce mode d'indexation, le fichier d'index comprend, en plus des mots, leurs positions dans le texte. Cette position peut être gérée soit lors de l'indexation et calculée par rapport au début du texte, soit en fonction de la structure du texte. Par exemple, le texte est structuré en section, paragraphe et phrase. Ce mode d'indexation permet des recherches plus évoluées (ex : si on veut s'informer sur les vaches folles, on peut préciser qu'on souhaite trouver des textes où vaches et folles sont l'un à côté de l'autre).

L'inversion (fichier inverse) réside dans le fait qu'on manipule les documents par leurs mots clés. Cette méthode n'est pas coûteuse, mais inconsistante car elle ne couvre pas tous les documents en question, du fait qu'elle ne cherche pas l'équivalence entre les termes. Les avantages de cette méthode sont principalement : la rapidité de la recherche et la simplicité de l'implémentation. Quant aux inconvénients, on peut citer la taille énorme prise par l'index (elle peut dépasser de trois fois celle des documents) et le temps de réorganisation de l'index surtout si l'environnement est dynamique **[LATO 98]**.

II.5.2 Sémantique ou Linguistique (science des langues)

La méthode linguistique fait appel à des techniques d'analyses reposant sur les connaissances actuelles de la langue et de sa structure, elle repose sur une analyse complète de la phrase et ses constituants. Cette méthode est concrétisée dans les travaux de **David** et **Plante (1990)** dans l'élaboration d'un système d'acquisition automatique des termes appelé **NOMINO** pour le centre d'ATO de l'Université du Québec à Montréal, dans les travaux de **Didier Bourigault (1992)** et les travaux de **Voutilainen (1993)** sur l'anglais qui ont donné naissance au logiciel nommé **NPtool [ELHA 97]** .

C'est une méthode de recherche documentaire très prisée, elle offre une efficacité accrue même si elle est difficile à mettre en oeuvre et très gourmande en ressources. Le principal atout de cette méthode est sa capacité à détecter les structures sémantiques contenues dans les documents, ce qui lui permet de retrouver des documents ne partageant aucun mot avec la question de l'utilisateur.

La localité sémantique des termes est une relation qui regroupe les termes qui sont utilisés pour exprimer une idée précise, cette technique intervient pour satisfaire les requêtes

d'utilisateurs qui visent en réalité les documents qui traitent une idée précise plutôt que des documents comportant des mots isolés.

Ces méthodes tiennent compte de la sémantique du texte, dont le but est d'indexer un document par son sens de telle façon que deux documents différents lexicalement mais ayant le même sens auront la même représentation.

Jusqu'ici nous n'avons parlé que des structures des mots, mais un mot peut avoir plusieurs sens différents (exemple : **bureau** peut être un local ou une table, et plus particulièrement le terme "or" qui est un mot vide ce qui nous empêche de trouver les documents sur l'**or** en tant que métal).

Le problème de la linguistique devient grave lorsque les documents sont écrits dans plusieurs langues. Il est donc important d'analyser le contenu textuel par la connaissance du **vocabulaire**, **la grammaire** et **la sémantique** des documents.

II.5.2.1 Principes des techniques linguistiques

a) Vocabulaire : Le vocabulaire comprend les mots de la langue, nous différencions deux types: **[BENA 00]**

- ✓ **Les mots lexiques ou mots sources :** les mots dans une forme normalisée, au singulier pour un nom et à l'infinitif pour un verbe.
- ✓ **Les formes lexicales ou formes fléchies :** les variantes de ces mots selon leurs contextes : pluriels, formes conjuguées, formes substantives...

b) Grammaire : Elle va décrire les catégories grammaticales : nom, verbe, article, adjectif, adverbe..., et les règles de structures de la langue, par exemple une structure comme: «article nom verbe adjectif» est acceptable en français.

c) Etude sémantique ou littéraire du texte : Le vocabulaire et la grammaire permettent de décrire les énoncés ou les phrases. Toutefois, pour comprendre un texte, il est nécessaire de faire aussi une analyse littéraire du texte.

1) Les dictionnaires électroniques : ils sont de 10 à 1000 fois plus grands qu'un dictionnaire traditionnel car ils contiennent toutes les formes fléchies des mots, par exemple pour le mot cheval on trouvera dans ces dictionnaires deux entrées cheval et chevaux, et pour un verbe on obtiendra toutes ses formes conjuguées. Cependant un terme peut avoir plusieurs significations comme "peigne" peut être un nom ou une forme conjuguée des verbes peindre et peigner. Il faut aussi

prendre en compte les mots composés et les expressions idiomatiques comme chemin de fer, cheval d'arçon.

Exemple d'un dictionnaire électronique :

chevaux : *nom*	forme canonique : **cheval**
avions : *nom*	forme canonique : **avion**
avions : *verbe*	forme canonique : **avoir**
cheval d'arçon : *mot composé*	forme canonique : **cheval d'arçon** domaine sémantique : sport définition : Appareil de gymnastique
...	

Tableau II.1 Dictionnaire électronique

2) **Les réseaux sémantiques :** c'est le fait d'associer des concepts aux mots ou aux expressions afin de traiter automatiquement le langage. Ces réseaux reposent sur des relations entre concepts qui dépassent largement la structure d'un ***thesaurus*** du fait que leurs relations sont identifiées et pondérées selon leurs raretés ou leurs particularités dans le langage.

Finalement, nous pouvons dire qu'à partir des dictionnaires électroniques et des réseaux sémantiques il va donc être possible d'effectuer une analyse du langage.

II.5.2.2 Les différentes étapes de traitement des documents

- **Analyse morphologique :** C'est la première étape de l'analyse textuelle, elle consiste à reconnaître les mots d'un texte et à associer à une forme fléchie la forme canonique d'un mot ou d'une expression idiomatique lorsque cela est possible, on appelle cette opération la **lemmatisation**.

Cette analyse parcourt donc essentiellement les dictionnaires électroniques. Or pour certains termes sous forme fléchie, il existe plusieurs formes canoniques (ex : Table est soit du verbe Tabler, soit un nom), le problème se complique lorsque le document contient des noms propres (ex : Monsieur Gâteaux).

L'analyse morphologique prend en compte toutes les formes typographiques : les majuscules, tirets, apostrophes... etc. Même les mots vides sont pris en compte. On peut finalement dire que cette analyse peut être aussi utilisée pour la correction orthographique si on utilise les dictionnaires.

- **Analyse syntaxique :** Elle s'appuie sur la structure grammaticale de la langue, elle analyse les phrases ou les extraits pour identifier, lorsque c'est ambigu, le rôle des différents termes : nom, verbe, adverbe, adjectif..., ainsi que leurs caractéristiques.

Le but de cette analyse est de lever les ambiguïtés en analysant les dépendances syntaxiques entre termes, ainsi dans la phrase: "poses ça sur la table!" le terme Table ne peut être qu'un nom et non une forme conjuguée du verbe Tabler.

- **Analyse sémantique :** C'est l'étape la plus importante dans le traitement sémantique, car elle met en œuvre tous les principes linguistiques pour analyser les mots et les phrases d'un texte, en employant les dictionnaires électroniques et les réseaux sémantiques, afin de déterminer le contenu sémantique (le sens) du texte.

A première vue, c'est l'approche idéale pour l'indexation d'information, car en théorie elle permettrait d'avoir une compréhension totale et globale des documents à indexer et la correspondance entre les différents documents et la question de l'utilisateur serait d'une efficacité accrue. Cependant, la mise en oeuvre d'une telle méthode requiert des algorithmes très puissants de traitement du langage naturel, des analyseurs lexicaux, des analyseurs syntaxiques ultra perfectionnés et un temps de traitement non négligeable.

Les méthodes basées sur une étude sémantique sont particulièrement efficaces dans des domaines restreints (d'un vocabulaire restreint), mais cette efficacité est moindre quand il s'agit d'une généralisation du domaine d'application.

En définitif, les méthodes sémantiques utilisent des outils très complexes et de ce fait sont difficiles à manipuler.

II.5.3 Les techniques statistiques

II.5.3.1 Définitions

Ce sont des méthodes qui mettent en oeuvre des concepts statistiques. Ainsi elles se basent dans leur analyse sur la fréquence (ou la présence) des mots constituant la question et les documents. Ces techniques reposent essentiellement sur la relation entre la fréquence d'un terme dans un document et la pertinence de ce document, et la relation inversement proportionnelle entre l'importance d'un terme et le nombre de documents le contenant. De ce fait, et en partant du principe que plus le terme est rare plus il est discriminant, les termes rares seront privilégiés, et inversement **[LATO 98]**.

L'approche statistique offre, dans certaines circonstances, des avantages indéniables puisqu'elle permet de s'attaquer à des ensembles de données d'une taille imposante qu'il serait tout à fait impensable de traiter manuellement. Elle permet aussi de traiter des ensembles textuels pour lesquels des dictionnaires électroniques n'ont pas été élaborés en vue d'un traitement linguistique. L'initiateur des méthodes d'indexation automatique reste sans aucun doute **H.P. Luhn** avec son célèbre article ***The automatic creation of literature abstracts*** paru en 1958 dans le ***Journal of Research and Development*** d'IBM **[ELHA 97]** .

Ces techniques tiennent compte de la pondération des termes des documents. Autrement dit, elles donnent un poids aux mots selon leurs fréquences et leurs positions dans le texte. Elles se basent sur les principes suivants :

- Il existe une relation entre la fréquence d'un terme à l'intérieur d'un document et son importance pour la représentation du document.
- Il existe aussi une relation inversement proportionnelle entre l'importance d'un terme et le nombre total de documents contenant ce terme.
- Le calcul du poids se fait selon la position du mot dans le texte (le mot aura un poids supérieur s'il fait partie du titre) ainsi que selon son occurrence.

Le principe utilisé consiste à calculer une distance entre les documents trouvés dans le lot de réponse, pour distinguer ceux qui se ressemblent et ressemblent le plus à la question. Les termes se divisent en trois catégories :

- Les termes très fréquents, peu spécifiques qui ***attirent*** les documents vers eux.
- Les termes peu fréquents, qui ***regroupent*** les documents relativement semblables.
- Les termes très rares, bons pour une recherche précise, du fait qu'ils sont discriminants.

Les méthodes statistiques, ou encore méthodes par extraction, sont basés sur deux types de traitement : le premier est basé sur le calcul de fréquences statistiques, le deuxième est construit sur une recherche de voisinage (que l'on appelle également méthodes par co-occurrence) avec ou sans élimination de polysémies ou avec le calcul de la distance moyenne

La méthode statistique est, en fait, basée sur le mot plein (le lexique). En effet, une fois que tous les mots vides, ceux qui ne portent pas de sens en soi (mots grammaticaux, articles, etc.), sont éliminés, il ne reste que les mots pleins. On tient compte du fait que plus un mot plein est présent dans un texte plus il est significatif et servira ainsi de descripteur et pourra apparaître lors d'une interrogation **[DROU 03]** .

II.5.3.2 Les étapes d'une indexation statistique

Toute méthode statistique doit comporter les phases suivantes :

- Génération du texte réduit par élimination des mots vides.
- Sélection des mots du titre et leur attribuer un poids préliminaire.
- Calcul des poids sémantiques des mots selon des critères.

II.5.3.3 La pondération des termes

Elle consiste à affecter à l'indexation comme à la recherche un coefficient de pondération à chaque descripteur, coefficient calculé selon son importance dans le document et dans la question. Ce poids détermine le pouvoir sémantique et discriminatoi re d'un mot clé dans la représentation du contenu d'un document.

La base de détermination d'un poids sémantique d'un mot-clé est sa fréquence d'apparition. Il existe plusieurs formules mathématiques pour calculer le poids de chaque terme dans le texte, en se basant sur un ou plusieurs critères (position du terme, sa fréquence, sa rareté, ...)

- **Fréquence d'apparition** : Les travaux de recherche documentaire s'appuient sur le fait que la fréquence d'apparition des mots dans des textes est significative de leur importance. On distingue deux types de fréquences :

1) **Fréquence relative** : La fréquence relative d'un mot est calculée par rapport à un document. Elle est égale au rapport entre le nombre de fois qu'on trouve ce mot dans le document et le nombre total de mots dans ce dernier. Ce qui donne :

$$fi = \frac{Log(Nci+1)}{Log(N+1)}$$

Où : ***fi*** : fréquence du terme i
Nci : nombre d'occurrences du terme i
N : nombre de termes dans le document

2) **La fréquence inverse du terme** : La fréquence inverse d'un terme est calculée par rapport à une collection de documents. C'est l'inverse du nombre de documents contenant le terme par rapport au nombre de documents D du corpus. Du fait que plus le mot est rare plus il est important.

$$fIi = 1 - \frac{\log(di + 1)}{\log(D + 1)}$$

Où : *fIi* : représente la fréquence inverse du terme *i*
di : représente le nombre de documents contenant le terme *i*.
D : représente le nombre total de documents dans le corpus.

- **La distribution des termes dans le texte :** Un autre critère de l'importance d'un mot est sa distribution et sa répartition dans le texte. En partant du principe : quand un mot est réparti sur un paragraphe, il est important seulement au niveau de cette partie du texte, mais s'il est réparti sur tout le document alors il est pertinent par rapport à ce dernier.

  La distribution nécessite d'attribuer à chaque mot une adresse dans le document de la forme : (N° paragraphe, N° phrase, N° proposition, N° mot). Le calcul de cette distribution, est une variance de chaque champ des adresses des occurrences de chaque mot, si cette variance est petite alors on peut conclure que le mot en question est local à une partie du texte, sinon il est bien distribué sur l'ensemble du document.

- **L'appartenance au titre** : Les concepteurs de moteurs de recherche, basés sur les méthodes statistiques, attribuent un poids supérieur aux mots du titre, vu que ce dernier est plus significatif que le contenu du document.

II.5.3.4 Les critères employés par les méthodes statistiques

Lors de la conception d'un moteur de recherche reposant sur une méthode statistique, les cinq critères suivants doivent être employés **[BENA 00]**:

- La fréquence des mots extraits de la question dans les documents trouvés. Plus un mot de la question est présent dans le texte du document trouvé, plus il est significatif pour le texte.
- Le deuxième critère étend le premier, car non seulement il tient compte de la fréquence des mots de la question, mais aussi celle des mots associés dans un dictionnaire de synonymes ou un thesaurus.
- Quant au troisième, il traite de la proximité des termes de la question dans les documents trouvés. Allant du principe, que plus les mots cherchés sont proches, plus le document est pertinent.

- La rareté des termes de la requête dans l'index. Ceci dit, plus la question est pointue, plus il ne lui correspond que peu de documents pertinents.
- La position du terme dans le texte. Si le mot fait partie du titre, on peut conclure qu'il représente mieux le contenu sémantique du texte, par opposition s'il était dans un paragraphe quelconque. De ce fait, la pondération change selon l'apparition du terme dans le document

Ces principes restent des suppositions mathématiques pouvant apporter une nuance dans la représentation du modèle sémantique d'un document. Par conséquent, ils peuvent aboutir à des résultats erronés, en donnant à un mot une importance qu'il ne mérite pas, ou inversement. Parmi les avantages de ces méthodes leur indépendance vis-à-vis de la langue utilisée.

II.5.4 Modèles hybrides

Les modèles hybrides sont, comme leur nom l'indique, à mi-chemin entre les modèles linguistiques et les modèles statistiques. En effet, certains auteurs préfèrent commencer le traitement des corpus par une analyse linguistique dont les résultats sont filtrés à l'aide de techniques statistiques alors que d'autres procèdent inversement **[DROU 03]** .

Les approches hybrides constituent un compromis entre les deux grandes tendances de base et s'en approprient donc les avantages et les inconvénients. En effet, leur puissance de traitement, reposant principalement sur l'adoption de modèles traitant de l'information sous forme numérique plutôt que linguistique, permet de s'attaquer plus facilement à des corpus de taille imposante. Cette caractéristique les sert bien puisque, de façon à obtenir des résultats de qualité et à minimiser le niveau de bruit obtenu, ces algorithmes doivent avoir accès à un volume de données important **[ELHA 97]** .

II.5.5 Comparaison entre les techniques d'indexations

Ces méthodes ne sont pas comparables car chacune vise un domaine d'application bien défini. Donc on va citer seulement les inconvénients et les avantages de chaque technique.

II.5.5.1 Le texte intégral

C'est la méthode la plus utilisée, du fait de sa simplicité et rapidité. Elle se contente d'indexer les mots du texte sans aucun traitement sémantique ou syntaxique. Enfin, on peut

dire que cette méthode peut être la base pour d'autres méthodes comme la statistique, afin de pallier à ses inconvénients.

II.5.5.2 La sémantique

Mis à part sa lourdeur et sa complexité, cette méthode est idéaliste et s'applique dans des langages restreints (scientifique, médical,...) et des domaines spécifiques (comme l'astronomie, les mathématiques, la physique,..).

II.5.5.3 Les méthodes statistiques

Ces méthodes présentent beaucoup d'avantages surtout lors de la recherche en classifiant les mots selon leur rareté et fréquence. Ceci permettra de classer les documents résultats et leur indépendance vis-à-vis des langues.

Leur seul inconvénient est qu'elles ne tiennent pas compte des relations entre les termes, par conséquent ceci limitera la recherche. En pratique ce problème peut être évité en associant un thesaurus contenant les termes les plus fréquents et leurs synonymes.

Finalement, nous pouvons dire que ces méthodes ne sont pas concurrentes mais complémentaires. Un bon moteur d'indexation et de recherche ne se contente pas d'une seule, car chacune a ses avantages et ses inconvénients, mais en les combinant nous les améliorons.

II.5.5.4 Les méthodes hybrides

L'approche hybride exploite aussi la systématicité, la rapidité et l'indépendance par rapport au domaine des algorithmes statistiques. Cette indépendance se manifeste aussi par l'absence du besoin de dictionnaires et de grammaires spécialisées. Il s'agit là d'un avantage indéniable étant donné que ces techniques ont généralement pour but d'assister l'humain dans l'élaboration de dictionnaires **[ELHA 97]** .

Les méthodes hybrides permettent aussi l'injection de connaissances par le linguiste. Cette intervention humaine permet de modeler les résultats sur les intuitions du linguiste et de mettre de côté les aspects plus «froids» des résultats obtenus par les approches statistiques. De résultats purement statistiques, l'intervention des connaissances linguistiques permet l'obtention de résultats plus satisfaisants pour le linguiste en fonction des phénomènes qu'il cherche à observer et à décrire.

Chapitre III

RECHERCHE ET PRESENTATION

Dans ce chapitre :

- La recherche
- La présentation des résultats

III.1 LA RECHERCHE

C'est la partie frontale pour l'utilisateur, là où il peut poser sa question et lancer sa requête. Cette action déclenche une recherche dans la base, à la suite une page (WEB ou une fenêtre) sera affichée, intégrant les réponses sous forme d'une liste de liens vers les documents classés par ordre de pertinence.

III.1.1 Interrogation documentaire

La finalité d'un moteur de recherche est la recherche des documents pour les mettre à la disposition des utilisateurs. Il existe de nombreuses méthodes de recherche d'informations dans un fond documentaire, nous citons les suivantes **[DELA 99]**:

III.1.1.1 Interrogation en langage courant (naturel ou quasi naturel)

De loin la mieux adaptée au texte, elle permet à l'utilisateur de formuler une question totalement libre (en langage naturel = le langage commun de "tous les jours") pour effectuer sa recherche. Une telle recherche nécessite une indexation et une recherche "intelligente" mettant en oeuvre des modules de traitements linguistiques élaborés. Elle rend la recherche aisée et accessible à tout utilisateur en lui offrant un langage d'interrogation proche de son langage courant. L'utilisateur formule une requête en texte libre, le moteur de recherche analyse son contenu et le convertit en éléments du langage d'indexation. Le système, après comparaison des éléments de la requête avec ceux des documents, détermine les degrés de ressemblances de ces derniers avec la requête et sélectionne ceux qui ont un degré de ressemblance supérieur à un seuil donné.

Du fait de sa complexité (utilisation de la sémantique), aucun système de recherche sur Internet ne dispose à l'heure actuelle d'une telle puissance de traitement du langage.

III.1.1.2 Interrogation par mots clés

L'utilisateur doit formuler sa question en utilisant des mots reconnus par le système. L'introduction d'un mot clé engendre la sélection des documents contenant ce mot clé ou ses synonymes, si un thesaurus est utilisé. Ce mode d'interrogation est une simple comparaison entre l'index et les mots de la requête, le résultat de cette concordance est l'ensemble des documents contenant les mots de la question.

Le schéma suivant donne les étapes suivies par un moteur de recherche pour effectuer la recherche des documents pertinents par rapport à une requête donnée.

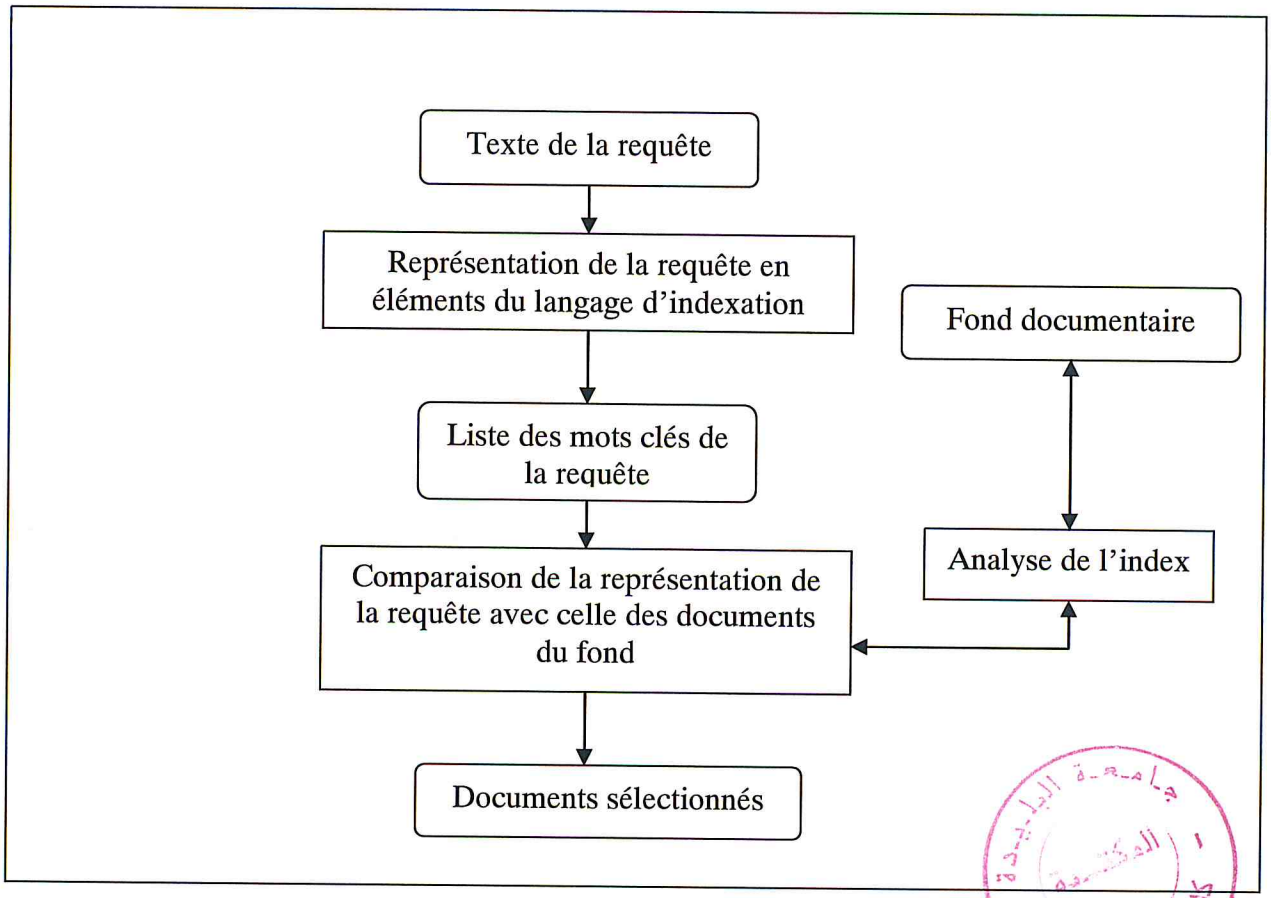


Figure III.1 Les étapes de la recherche des documents

III.1.2 Les techniques de recherche

Les techniques de recherches permettent de formuler et de transformer les requêtes des utilisateurs pour leur correspondance avec la base d'information. Les plus importantes seront présentées ci-dessous **[LELO 98]**.

D'abord, il faut savoir que la qualité de l'information obtenue est étroitement liée à la qualité de la formulation de la requête. Ainsi la recherche peut faire l'objet de quelques risques :

- ***Le Bruit*** : le premier risque d'une recherche non efficace est d'engendrer trop de documents, c'est le phénomène de ***bruit*** qui correspond aux documents nommés en réponse, mais qui ne sont pas pertinents par rapport à la question posée.
- ***Le Silence*** : a contrario, le second risque est le manque de résultat qui se traduit par un ***silence total*** correspondant aux documents pertinents qui n'apparaissent pas dans le résultat de la recherche.

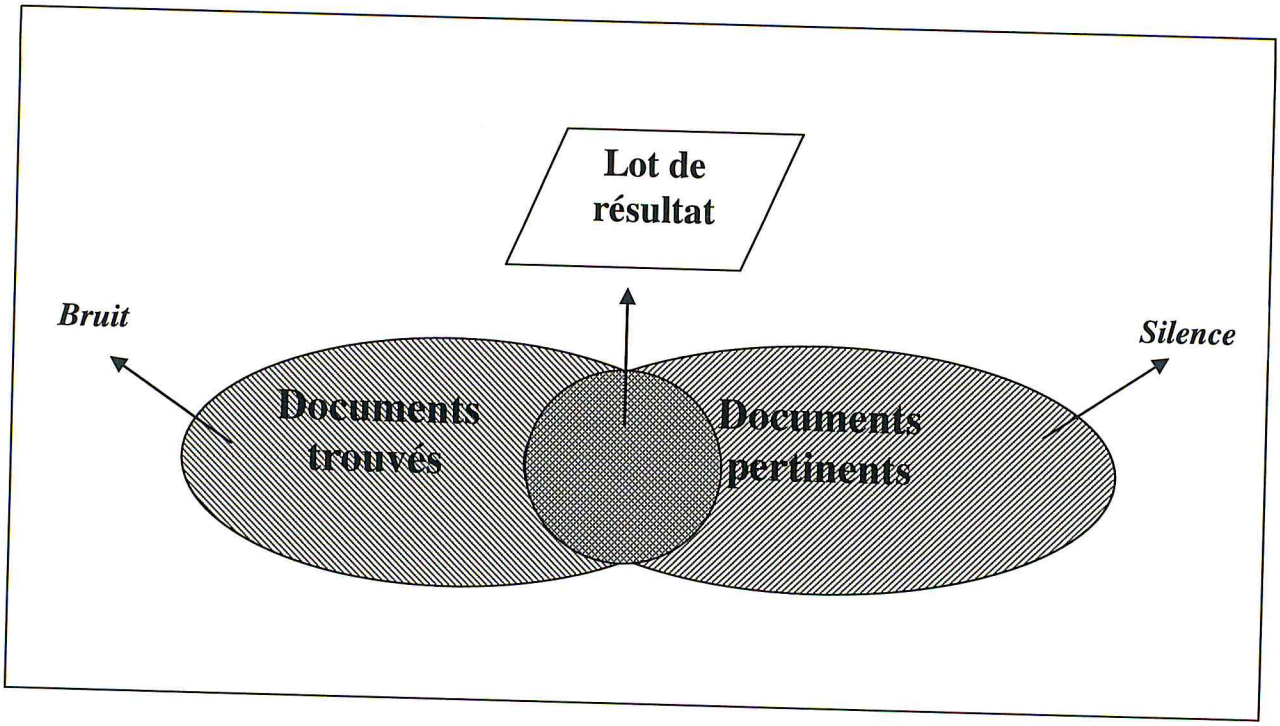


Figure III.2 Bruit et silence

Evidemment, l'idéal est de minimiser le taux de bruit et de silence, mais dans la réalité, il s'agit de trouver un compromis.

La recherche par le contenu est fondée sur une comparaison entre la question de l'utilisateur et le contenu de la base d'information. Cette comparaison implique une marge d'erreur et d'ambiguïté. De ce fait, l'évaluation de la qualité de la recherche documentaire se fait par ces quatre critères **[BENA 00]**:

$$\textit{Rappel} = \frac{\text{le nombre de documents pertinents parmi les réponses pour l' utilisateur}}{\text{le nombre de documents pertinents pour le système}}$$

$$\textit{Précision} = \frac{\text{le nombre de documents pertinents parmi les réponses pour l' utilisateur}}{\text{le nombre total de documents réponses pour l' utilisateur}}$$

SILENCE = 1 – RAPPEL ***BRUIT*** = 1 – PRECISION.

Remarque: pour un moteur de recherche idéal, il faudra que RAPPEL = PRECISION = 1.

III.1.2.1 Recherche booléenne

Elle se base sur la logique binaire pour décider qu'un document est pertinent pour une requête, cette dernière est une expression logique de termes connectés par la conjonction, disjonction ou la négation...Il s'agit d'introduire des opérateurs booléens à la requête. Les opérateurs les plus utilisés sont : Et, Ou, Sauf. Plus la requête est compliquée, plus elle devient difficile à déchiffrer. Ce modèle est simple sur le plan performance quantitative, mais du coté qualité cette technique est très limitée.

III.1.2.2 Recherche textuelle

Les opérateurs utilisés dans ce mode de recherche sont la troncature et le masque pour les indexes mots clés, et les opérateurs de proximité pour les indexes texte intégral.

a. La troncature

La troncature est un signe qui remplace zéro ou plusieurs lettres d'un mot. En recherche dans une banque de données de textes libres et non structurés comme en contient le web, l'opérateur de troncature est capital. Dans ce contexte général, cet opérateur sert à trouver des mots ayant des variantes, les fautes d'orthographe étant les plus répandues en français.

En général, la troncature est représentée par l'astérisque «*» ou le point d'interrogation «?» .Selon les auteurs, la troncature est appelée «troncation», «joker» ou «wildcard».Spécifiquement, il y a trois formes de troncatures : la troncature à gauche, la troncature milieu et la troncature à droite **[URL01]**.

- **La troncature à droite** C'est la plus utilisée et la plus intéressante, elle permet d'effectuer une recherche en utilisant le début d'un mot afin d'obtenir les différentes formes dérivées du mot. Par exemple, si vous recherchez des documents sur les chats, vous pouvez saisir la requête "chat*" afin d'obtenir les documents contenant les termes chat, chats, chatte, chattes, chaton, chatons, ... Le résultat va cependant être surprenant, puisque les documents contenant les termes français chatoiement, chatouille, chatterton vous seront également retournés. Mais ça ne s'arrête pas là, puisque vous obtiendrez de nombreux documents de la langue anglaise contenant les mots chat (bavardage), qui n'a rien à voir avec votre recherche.

- **Les troncatures à gauche et interne** Fonctionnent de la même manière, elles permettent respectivement de rechercher des chaînes sans spécifier le début du mot, ou une partie interne du mot.

Cependant, la troncature doit être utilisée avec discernement. Il n'est pas recommandé de faire une troncature à partir d'un mot composé d'une racine, surtout si ce mot est le seul terme de la recherche. Par exemple, le mot «talon » est une racine d'où dérivent «talonnade, talonnage, talonner, talonnette, talonneur et talonnière», déclinaison verbale et plurielle compris. Dans l'exemple ci-dessus, la troncature à «talon» n'aurait apporté aucun effet significatif : il aurait donné sensiblement le même résultat. Il en va autrement avec des troncatures comme «psycho*» qui est la racine d'une vingtaine de mots. Avant d'utiliser la troncature, il est bon de vérifier la famille du terme dans un dictionnaire.

b. Le masque

C'est un caractère spécial qui remplace un ou aucun caractère, par convention c'est le caractère '#'. Par exemple, soit la requête MICRO#INSTRUCTION nous trouvons les documents contenant les mots MICRO_INSTRUCTION ou MICROINSTRUCTION.

c. Les opérateurs de proximité

Étant donné que la recherche se fait dans un texte, les concepteurs des logiciels de recherche ont imaginé la notion de proximité. Cette notion permet d'unir plusieurs termes en contexte. Le contexte est soit la phrase soit le paragraphe. La notion de proximité se décline en proximité proprement dite et en adjacence. Selon les auteurs, ces deux notions sont interchangeables. Dans le cadre de cette étude, l'opérateur de proximité relie deux mots côte à côte et l'opérateur d'adjacence deux mots séparés par au moins un autre mot **[URL01]**.

1-L'opérateur de proximité : utilisé par tous les moteurs de recherche. Cet opérateur est représenté par les guillemets anglais « " " ». Par exemple, la recherche sur «talon d'Achille» s'écrirait ainsi : "talon d'Achille". Les guillemets anglais sont utiles pour les expressions et pour les noms.

2-L'opérateur d'adjacence : L'opérateur d'adjacence est réduit à unir deux mots séparés par plus d'un mot. En utilisant cet opérateur, la recherche se fera sur deux mots séparés d'au plus 10 mots en texte source. Cela signifie que la ponctuation et les tableaux interfèrent dans la reconnaissance du groupe de mots. Si le groupe de mots chevauche deux phrases, le moteur de recherche l'affichera. Il en est ainsi pour les tableaux. Si le groupe de mots est dans des cellules contiguës, il sera affiché aussi. Voici un exemple de la recherche avec l'opérateur NEAR : (Talon NEAR Achille) AND (grec OR grecque)→ cette requête a permis de trouver un texte qui avait cette phrase: « Pâris, aidé par Apollon, tuera Achille d'une flèche au talon. »

III.1.2.3 Recherche pondérée

Il s'agit d'affecter à chaque terme un poids ou pourcentage (entre 0 et 100%) de façon à privilégier les documents qui contiennent certains termes et pas d'autres. (Ex : Si nous nous intéressons aux voiture de sport, nous formulerons une recherche par (voiture? near sport) OU (sport near mécanique) OU (voiture?) ou(Rallye). Ce qui va influencer le calcul de pertinence pour les documents qui répondent à la question, et contenant le mot « voiture » ou « sport », du fait de son occurrence dans la requête (deux pour voiture et sport, et un pour mécanique et rallye).

III.1.3 Les modèles de recherche

La recherche d'information est la mise en correspondance des représentations sémantiques des documents du corpus et d'une représentation de la requête de l'utilisateur.

Les modèles de recherche sont l'ensemble formé par le modèle de document (la représentation du document), le modèle de requête (la représentation de la requête) et la fonction de correspondance. Parmi les différents modèles de recherche classiques existants, nous citons **[LELO 98] & [LATO 98]**:

III.1.3.1 Le modèle booléen

Il se base sur la logique binaire pour décider qu'un document est pertinent pour une requête, cette dernière est une expression logique de termes connectés par la conjonction (et), la disjonction (ou) et la négation (non). La recherche se fait en référençant tous les documents touchés par les termes de la requête à l'aide des opérateurs booléens (ou, et, sauf) entre les mots clés. Ce modèle est simple sur le plan performance quantitative, mais du coté qualité ce modèle est très limité (le taux de bruit est très élevé).

III.1.3.2 Le modèle vectoriel

Proposé par George Salton de Cornell University qui a introduit les outils mathématiques (matrice...) afin de réduire les inconvénients du modèle précédent.

Le modèle vectoriel est basé sur la relation entre l'apparition d'un terme dans un document et son importance. Chaque document est représenté par un vecteur de la taille du nombre de termes dans le système ou le corpus, et chaque élément de ce vecteur représente le nombre d'apparitions du terme dans le document. Quant à la recherche, on représente la requête du client par un vecteur selon le même principe, et on compare ce vecteur Vq avec

tous les vecteurs documents de la base Vd en se basant sur la comparaison du cosinus de l'angle entre les deux vecteurs avec un seuil prédéfini.

Ce modèle est cependant limité si le nombre de documents est important et si un nouveau terme se présente, l'espace vectoriel augmente et le calcul de tous les vecteurs documents est obligatoire.

III.1.3.3 Le modèle probabiliste

Proposé par K.J Van Rijsberen, il est fondé sur l'estimation de la probabilité de pertinence d'un document par rapport à une requête et la classification automatique des documents. Ce modèle possède un inconvénient qui est la complexité du calcul de la valeur théorique de la probabilité.

III.1.3.4 Le modèle des réseaux sémantiques

C'est le modèle le plus complexe, car il est basé sur la sémantique des concepts organisée sous forme d'un graphe de connaissance, subdivisée en niveaux hiérarchiques contenant des thèmes et domaines. Ainsi pour la recherche, il faut déterminer le sujet et le domaine qui intéressent l'utilisateur et qui sont en concordance avec sa question.

III.2 LA PRESENTATION DES RESULTATS

C'est par le biais de la présentation des résultats que l'utilisateur fait une première évaluation des documents qui lui semblent intéressants. Les informations affichées doivent permettre deux choses essentielles :

- Comprendre globalement comment le moteur de recherche fonctionne (les mots considérés comme vides, les mots retenus, les mots inférés par le système et les mots retrouvés par le système dans les documents-réponses),

- Evaluer facilement la pertinence des documents-réponses sans avoir besoin de les consulter tous pour vérifier.

Une présentation des résultats avec de telles qualités, couplée à un système de recherche efficace doit permettre de créer un climat de confiance entre l'utilisateur et le système qu'il utilise. C'est en créant une telle relation que la recherche devient réellement efficace. Les critères concernant les informations délivrées par le système sur son fonctionnement et sur les documents ainsi que l'organisation des documents-réponses apportent une valeur ajoutée à la liste des documents.

III.2.1 Les informations à afficher par le système

Cet aspect doit être vu suivant deux angles différents : Les informations générales concernant la recherche et les informations portant sur chaque document-réponse.

III.2.1.1 Informations générales

Au niveau des informations générales concernant la recherche, nous pouvons citer **[BENA 00]** :

- **Le nombre de documents retrouvés par le système pour la recherche**

Cette information va permettre à l'utilisateur d'évaluer la pertinence de sa question. Si ce nombre est trop élevé, il va essayer de préciser sa question et de relancer une recherche. En revanche, s'il est relativement faible, il va certainement reformuler sa question de manière un peu plus générale afin de ne pas "passer à côté" de documents intéressants.

- **Listes des termes reconnus, non reconnus et ignorés par le système**

Ces termes permettent à l'utilisateur de mieux comprendre comment sa question a été traitée et donc de mesurer l'adéquation entre sa question telle qu'il l'a formulée et telle qu'elle a été interprétée par le système. Pour pouvoir distinguer les termes reconnus des termes non reconnus dans le lexique de la langue ou du domaine, le système doit fonctionner avec un dictionnaire (ou thesaurus), ce qui n'est pas le cas des systèmes actuels qui considèrent toutes les chaînes de caractères comme étant valides. Ils ne sont donc qu'en mesure de fournir une liste des termes présents ou absents de la base. Or, comme les moteurs indexent pour la plupart tous les mots, il est très rare que des mots ne soient pas reconnus, et l'information "termes reconnus/non reconnus" perd son intérêt pour évaluer l'adéquation entre le vocabulaire de l'utilisateur et celui utilisé dans la base. Il y a donc très peu de moteurs de recherche qui fournissent de telles listes.

III.2.1.2 Informations propres à chaque document

Au niveau des informations propres à chaque document, il y a quelques critères importants qu'un bon moteur de recherche doit retenir pour une bonne évaluation du contenu des documents et de leur pertinence :

- **Un lien hypertexte ou l'url du document d'origine**

Le lien hypertexte vers le document est indispensable afin de pouvoir le consulter. L'ensemble des outils disponibles affiche un lien vers le document trouvé.

- **Titre du document d'origine**

Tout comme le lien vers le document d'origine, le titre de ce dernier paraît être une information vitale pour un minimum de souplesse d'utilisation.

- **Extrait du document**

Il s'agit bien généralement des deux ou trois premières lignes du document, ce qui nous donne un texte souvent très peu informationnel. Un "résumé automatique" se révèle être une collecte de trois ou quatre phrases prises aléatoirement dans le document. Les seuls outils fournissant un résumé intéressant sont bien entendu ceux pour lesquels ce dernier est rédigé manuellement, mais en contre partie, très peu de documents sont disponibles avec un résumé.

- **La taille du fichier**

Ce n'est pas une donnée vitale, elle est fournie par la majorité des moteurs de recherche. Elle permet tout de même de se rendre compte si le document a plutôt la taille d'un résumé ou d'un livre entier (bien que cette taille soit toujours exprimée en Octets ou Kilo Octets, ce qui n'est pas forcément très significatif pour un non informaticien).

- **Listes des mots importants de chaque document**

C'est la liste des mots pertinents du document, indépendamment de ceux de la question. Cette liste permet à l'utilisateur d'avoir une idée globale sur le contenu du document. Ainsi lui donner la possibilité de juger par lui-même l'importance du document à consulter.

- **La date de création du document**

Elle donne à l'utilisateur une bonne idée sur la validité des données contenues dans le document à consulter.

- **La date de dernière mise à jour du document**

Cette information qui est liée à la précédente permet de connaître la date de la dernière modification du document. C'est une information clé dans un souci de veille informationnelle. En effet, on peut très bien imaginer de créer une interface permettant d'effectuer des recherches mais ne fournissant à l'utilisateur que les documents nouveaux ou ceux ayant changés récemment.

- **Une mesure de pertinence**

Elle permet d'évaluer approximativement la pertinence d'un document par rapport à un autre. Il faut néanmoins se méfier de ces valeurs souvent très subjectives, car parfois elles sont mesurées de manière différentes selon les moteurs.

- **La mise en évidence des mots de la question dans les documents-réponses**

C'est le moyen le plus commode de se rendre compte de la pertinence d'un document. En effet, si le système vous indique que tel document contient quatre des mots de la question, alors qu'un autre n'en contient qu'un seul, le premier semble beaucoup plus proche de ce que vous recherchez.

III.2.2 L'organisation des réponses

Le classement des documents-réponses fournis par les systèmes de recherche est un point important pour l'utilisateur. En effet, si ce classement est performant, l'utilisateur trouvera dans les premiers documents retournés par le système un nombre important de documents pertinents. Il sera alors amené à faire confiance au système et il gagnera énormément de temps en ne consultant que les premières réponses du système **[LELO 98]**.

Bien sur, il n'existe pas de critères absolus pour définir la pertinence d'un document et deux utilisateurs différents ne jugeront pas les mêmes documents comme pertinents pour une même question, même si le nombre de documents communs est relativement important. Les moteurs de recherche disponibles sur Internet/Intranet utilisent tous des heuristiques plus ou moins simplistes. En effet, les critères retenus pour donner du poids à un mot de la question sont :

- Le nombre d'occurrences du mot dans le document,
- La présence du mot dans le titre du document,
- La présence du mot dans les premières lignes du document,
- L'adjacence des mots de la question dans le document,
- La distribution du mot dans le document,
- Le nombre faible d'occurrences du mot dans la base. Ce critère provient du postulat que plus un mot est rare plus il est informationnel,
- La popularité de l'auteur : c'est le nombre de documents écrit par ce dernier.

Comme nous le voyons, les méthodes de classement varient énormément d'un moteur à l'autre.

III.2.3 Les informations à valeur ajoutée à fournir à l'utilisateur

Un certain nombre de recherches porte actuellement sur les systèmes graphiques de navigation sur Internet. De telles représentations graphiques devraient permettre de synthétiser un grand nombre d'informations sur les documents et surtout sur leur contexte, permettant ainsi à l'utilisateur d'avoir une vision claire des résultats obtenus.

La possibilité d'effectuer de nouvelles recherches à partir des documents-réponses sélectionnés est une fonctionnalité que l'on trouve dans certains moteurs de recherche. Son but est de pouvoir, à partir d'un ou plusieurs documents obtenus en résultat, de lancer une nouvelle recherche afin de trouver les documents similaires. Il s'avère que les résultats sont actuellement très peu convaincants. La recherche à partir de documents entiers provoque des résultats énormément bruités et donc inexploitables **[DELA 99]**.

Chapitre IV

CONCEPTION

Dans ce chapitre :

- Introduction
- La methode de conception OMT
- Le Spider
- L'interface web
- Persistance des données
- Architecture

IV.1 INTRODUCTION

De nos jours, les réseaux Intranet connaissent une croissance exponentielle en terme de documents stockés, d'informations échangées et d'utilisateurs en quête d'information. Cette dernière est souvent disponible mais inaccessible. Afin de palier a ce problème, les systèmes d'informations doivent être dotés d'outils de recherche documentaire performants et puissants.

Dans ce sens, nous travaillons sur l'élaboration d'un moteur de recherche Intranet effectuant les taches suivantes :

- Collecter les documents présents au sein du réseau pour une recherche, une consultation et un accès efficaces par les utilisateurs. Notre système parcours le réseau pour indexer son contenu documentaire et ceci dans le but d'éviter aux utilisateurs le calvaire de la recherche séquentielle qui les oblige à parcourir toute la base d'information.
- Permettre aux utilisateurs de publier des documents sur le réseau, en leur fournissant les moyens nécessaires pour la publication et la gestion de leurs documents.
- Prendre en considération la confidentialité des documents car certaines informations ne doivent être délivrées qu'à un nombre restreint d'utilisateurs.
- Permettre aux utilisateurs d'effectuer des recherches documentaires rapides et efficaces à l'aide d'interfaces faciles à utiliser.

IV.2 LA METHODE DE CONCEPTION OMT

La méthode OMT (Object Modeling Technique) a été inventée dans le centre de recherche et de développement de Général Electric a la fin des années 80. Le principal ouvrage d'écrivant la méthode est paru en 1991. OMT emploie trois modèles différents pour décrire un système : le modèle objet décrivant les objets et leurs relations dans le système ; modèle dynamique décrivant l'interaction entre les objets dans le système ; le modèle fonctionnel décrivant la transformation des données du système. Chaque modèle est applicable pendant toutes les phases du développement et agrège des détails d'implémentation au fur et à mesure que le développement progresse. La secription complete

UML (Langage de Modélisation Unifié) est une fusion des notations de Booch, OMT, OOSE et d'autres notations. La méthodologie UML est conçue pour être lisible sur des supports très variés. Les concepteurs de la notation ont recherché avant tout la simplicité; UML est intuitive, homogène et cohérente. Les symboles embrouillés, redondants ou superflus ont été éliminés en faveur d'un meilleur rendu visuel **[MULL 00]**.

La méthodologie UML se concentre sur la description des artefacts du développement de logiciel, plutôt que sur la formalisation du processus de développement lui-même, elle peut ainsi être utilisée pour décrire les éléments logiciels obtenus par l'application de différents processus de développement. UML n'est pas une notation fermée : elle est générique, extensible et configurable par l'utilisateur. UML ne cherche pas la spécification à outrance : c'est à dire qu'il n'y a pas une représentation graphique pour tous les concepts imaginables. En cas de besoins particuliers, des précisions peuvent être apportées au moyen de mécanismes d'extension et de commentaires textuels. Une grande liberté est donnée aux outils pour le filtrage et la visualisation d'information **[MULL 00]**.

IV.2.1 Les trois modèles

La méthode OMT emploie trois modèles différents pour décrire un système :

IV.2.1.1 Le modèle statique

- **Les diagrammes de classes :** qui représentent la structure statique en terme de classes et de relations.
- **Les diagrammes d'objets :** qui représentent les objets et leurs relations et correspondent à des diagrammes de collaboration simplifiés, sans représentation d'envoi de messages.

IV.2.1.2 Le modèle dynamique & fonctionnelle

- **Les diagrammes d'activités :** qui représentent le comportement d'une opération en termes d'actions.
- **Les diagrammes de cas d'utilisation :** qui représentent les fonctions du système du point de vue utilisateur.
- **Les diagrammes de collaboration :** qui sont une représentation spatiale des objets, des liens et des interactions.
- **Les diagrammes d'états – transitions :** qui représentent le comportement d'une classe en terme d'états.
- **Les diagrammes de séquences :** qui sont une représentation temporelle des objets et de leurs interactions

IV.2.2 implementation

- **Les diagrammes de composants :** qui représentent les composants physiques d'une application.
- **Les diagrammes de déploiement :** qui représentent le déploiement des composants sur les dispositifs matériels.

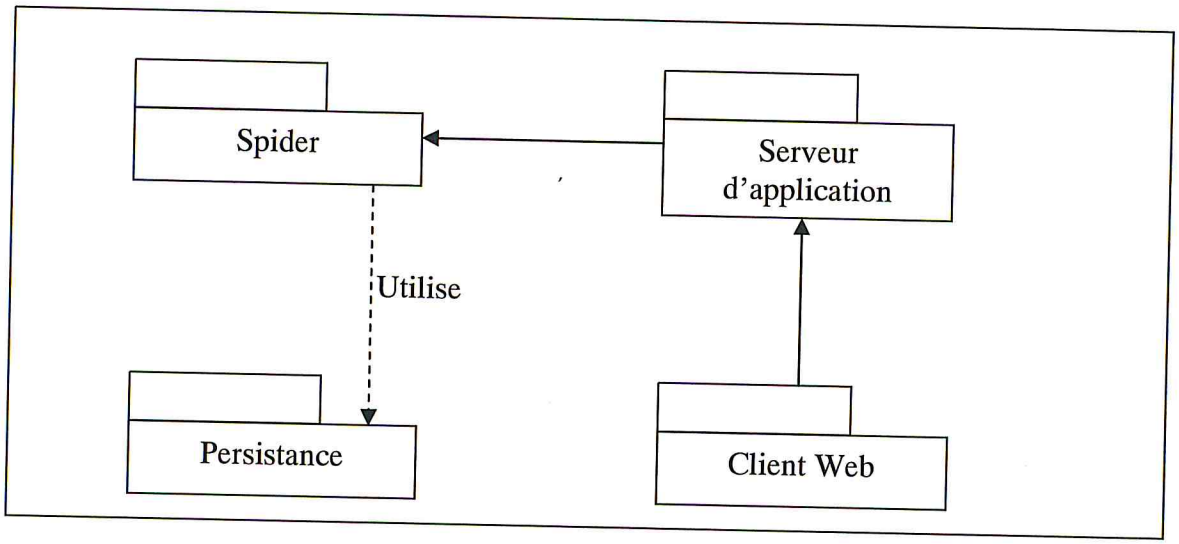


Figure IV.1 Paquetage du domaine Uni_Blida SEARCH

IV.3 DIAGRAMME DES CLASSES

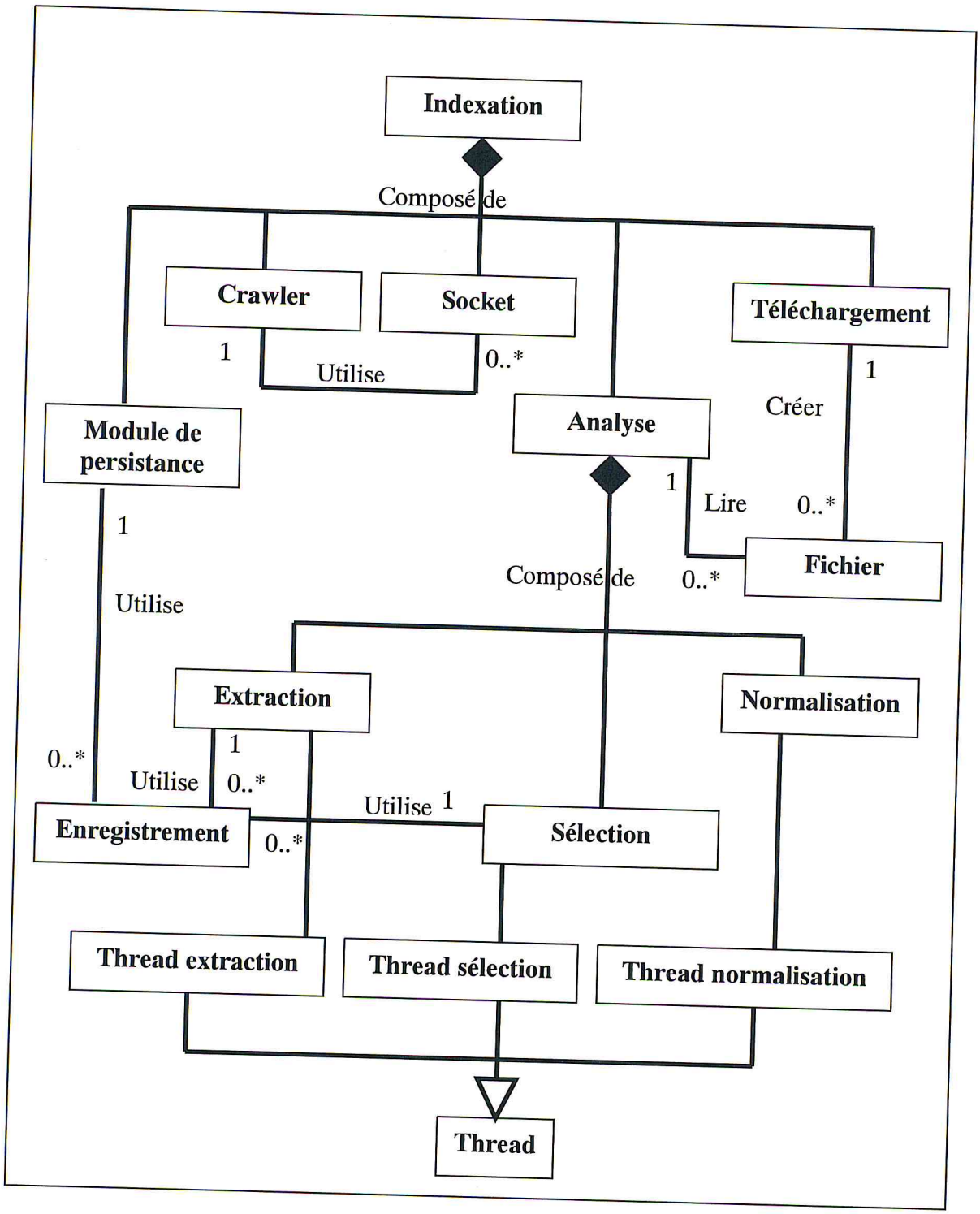


Figure IV.2 Diagramme de classes final du système Uni_Blida SEARCH

IV.4 LE SPIDER

IV.4.1 Détermination des cas d'utilisation

L'analyse débute par la recherche des acteurs du système. Un acteur représente un rôle joué par une personne ou autre qui interagit avec le système. Les acteurs se déterminent en observant les utilisateurs directs du système, ceux qui sont responsables de son exploitation ou de sa maintenance, ainsi que les autres systèmes qui interagissent avec le système en question **[MULL 00]**.

Un cas d'utilisation décrit un ensemble de séquences d'actions, y compris les variantes, qu'un système exécute pour produire des résultats tangibles pour un acteur.

Les acteurs du système sont :

- L'administrateur : acteur manipulant le système en indexant les documents contenus dans le réseau ainsi qu'en ajoutant et supprimant les utilisateurs ou les groupes.

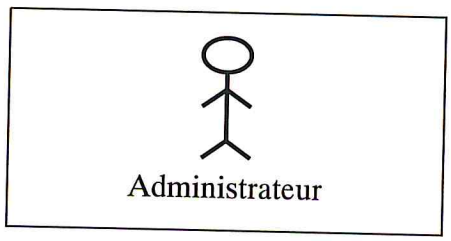


Figure IV.3 Les acteurs du système SPIDER

Après avoir défini les acteurs interagissant avec le système, nous determinons ses cas d'utilisations qui représentent le comportement du système en réponse à une interaction avec un acteur. Après analyse, nous pouvons associer à chaque acteur un ou plusieurs cas d'utilisations du système.

Acteur	Cas d'utilisation
Administrateur	• Ajout d'un utilisateur • Suppression d'un utilisateur • Modification des informations d'un utilisateur • Ajout d'un groupe • Suppression d'un groupe • Modification des utilisateurs d'un groupe • Lancement de l'indexation automatique • Gestion des documents.

Tableau IV.1 Cas d'utilisations du système SPIDER

IV.4.2 Diagramme des cas d'utilisations

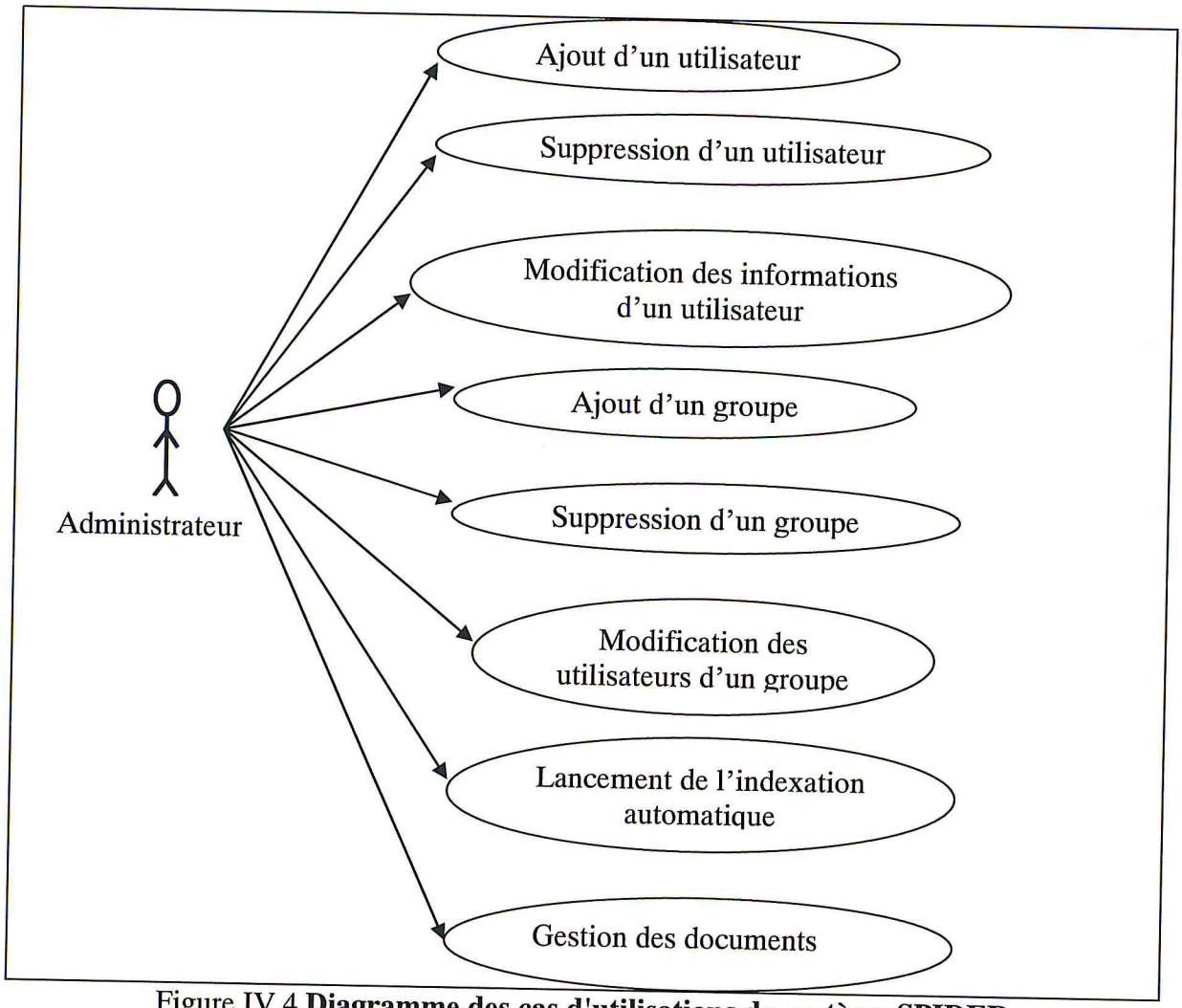


Figure IV.4 **Diagramme des cas d'utilisations du système SPIDER**

IV.4.3 Description des cas d'utilisations

Dans cette section nous décrivons chaque cas d'utilisation par un diagramme de séquence.

- **Ajout d'un utilisateur**
  - L'administrateur déclenche l'opération « ajout d'un utilisateur ».
  - Le système lui répond par un formulaire de saisie.
  - L'administrateur saisit les différents champs : le nom, le prénom, le login, le mot de passe (et sa confirmation) et enfin le nom du (des) groupe(s).
  - Le système vérifie que l'utilisateur n'est pas déjà inscrit.
  - Le système met à jour la base s'il s'agit d'un nouvel utilisateur.

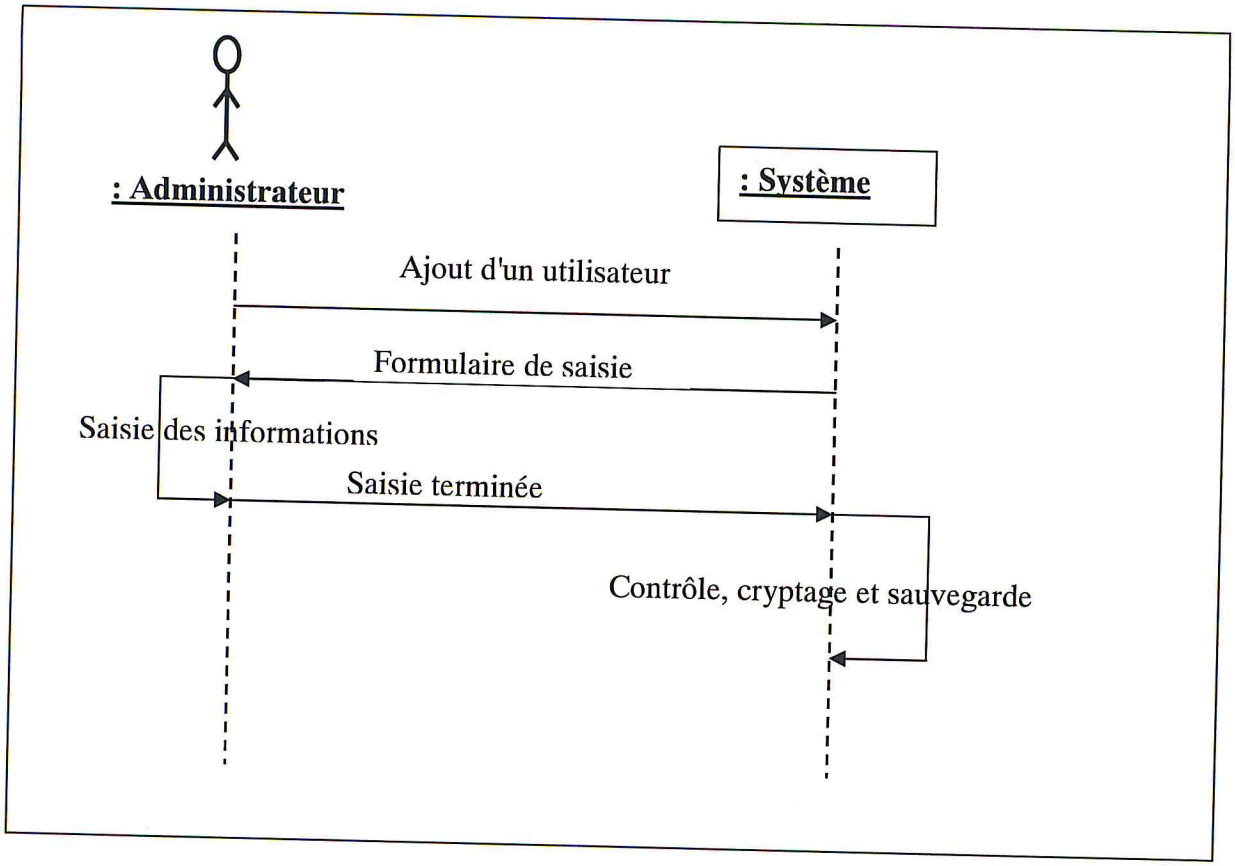


Figure IV.5 Cas d'utilisation «Ajout d'un utilisateur»

Suppression d'un utilisateur

- L'administrateur sélectionne l'utilisateur à supprimer dans la liste des utilisateurs.
- Le système enregistre les informations à supprimer et demande la confirmation de suppression.
- L'administrateur déclenche l'opération de suppression.
- Le système supprime l'utilisateur de la base.
- Le système met à jour l'affichage de la liste des utilisateurs.

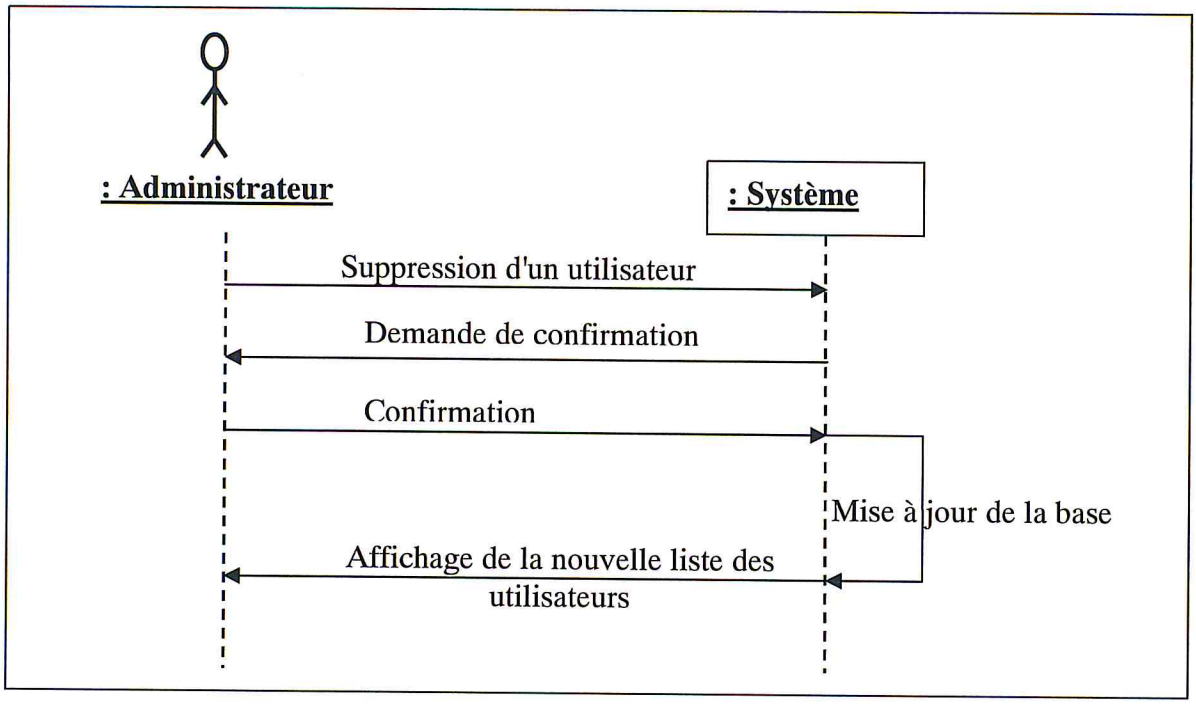


Figure IV.6 Cas d'utilisation «Suppression d'un utilisateur»

➢ **Modification des informations d'un utilisateur**

- L'administrateur sélectionne l'utilisateur.
- Le système affiche les informations relatives à l'utilisateur.
- L'administrateur modifie le login et/ou le groupe et valide.
- Le système vérifie que le login n'est pas déjà utilisé et met à jour la base.

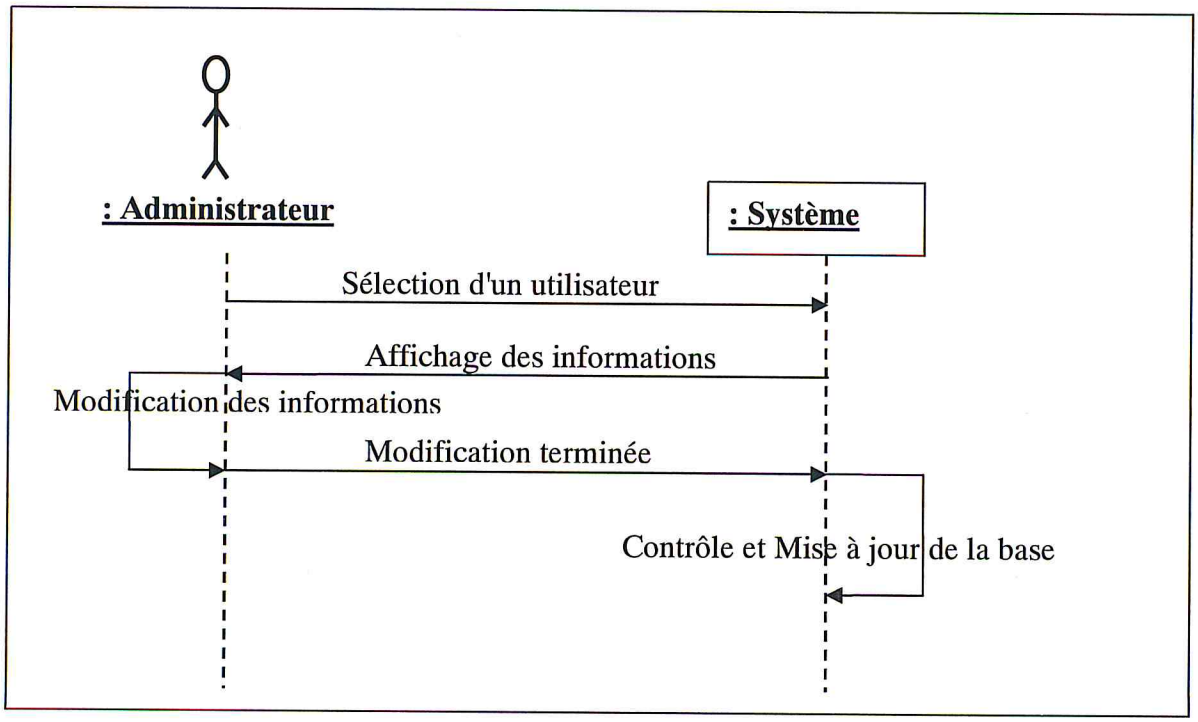


Figure IV.7 Cas d'utilisation «Modification des informations d'un utilisateur»

➢ **Ajout d'un groupe**

- L'administrateur déclenche l'opération ajout d'un groupe.
- Le système répond par un formulaire.
- L'administrateur saisi le nouveau groupe.
- Le système vérifie si le groupe existe déjà.
- Le système met à jour la base.

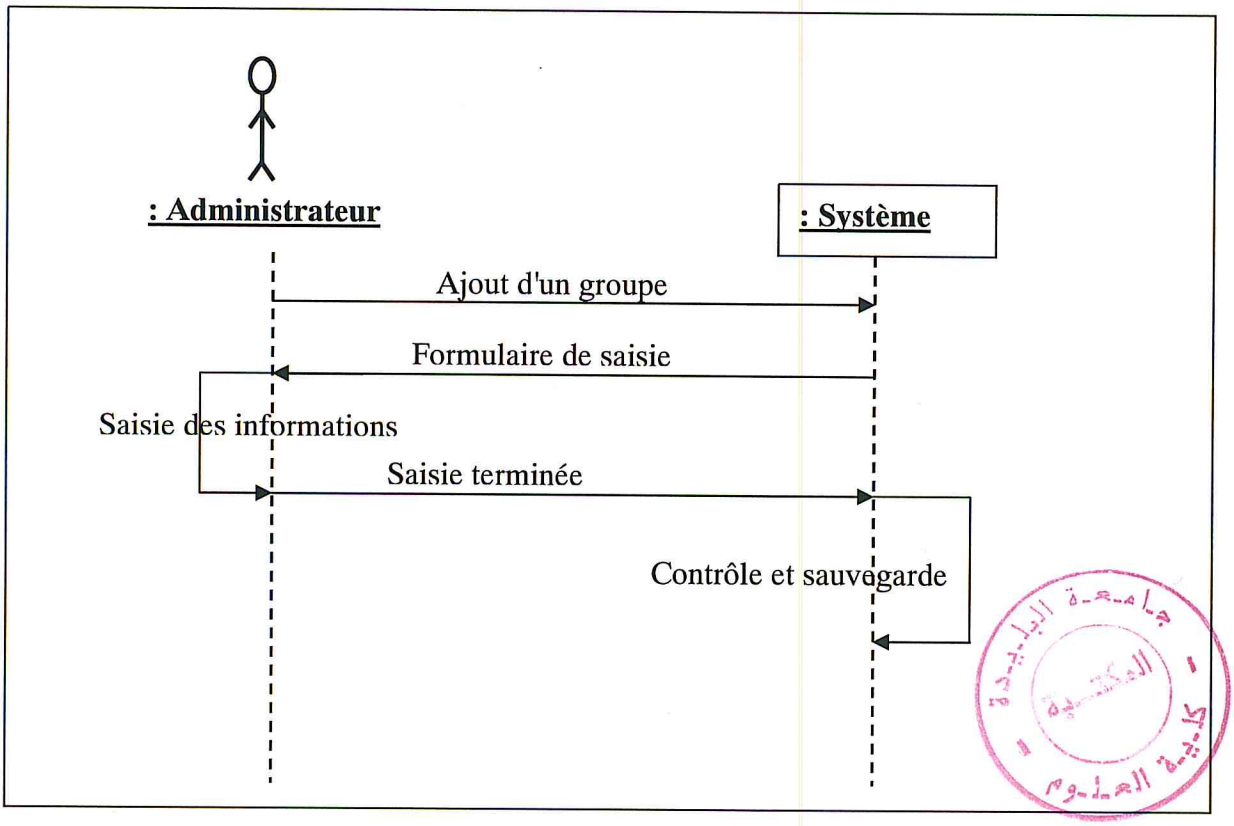


Figure IV.8 Cas d'utilisation «Ajout d'un groupe»

➢ **Suppression d'un groupe**

- L'administrateur sélectionne le groupe à supprimer de la liste des groupes.
- Le système enregistre les informations à supprimer et demande la confirmation de suppression.
- L'administrateur déclenche l'opération de suppression.
- Le système met à jour la base et affiche la nouvelle liste des groupes.

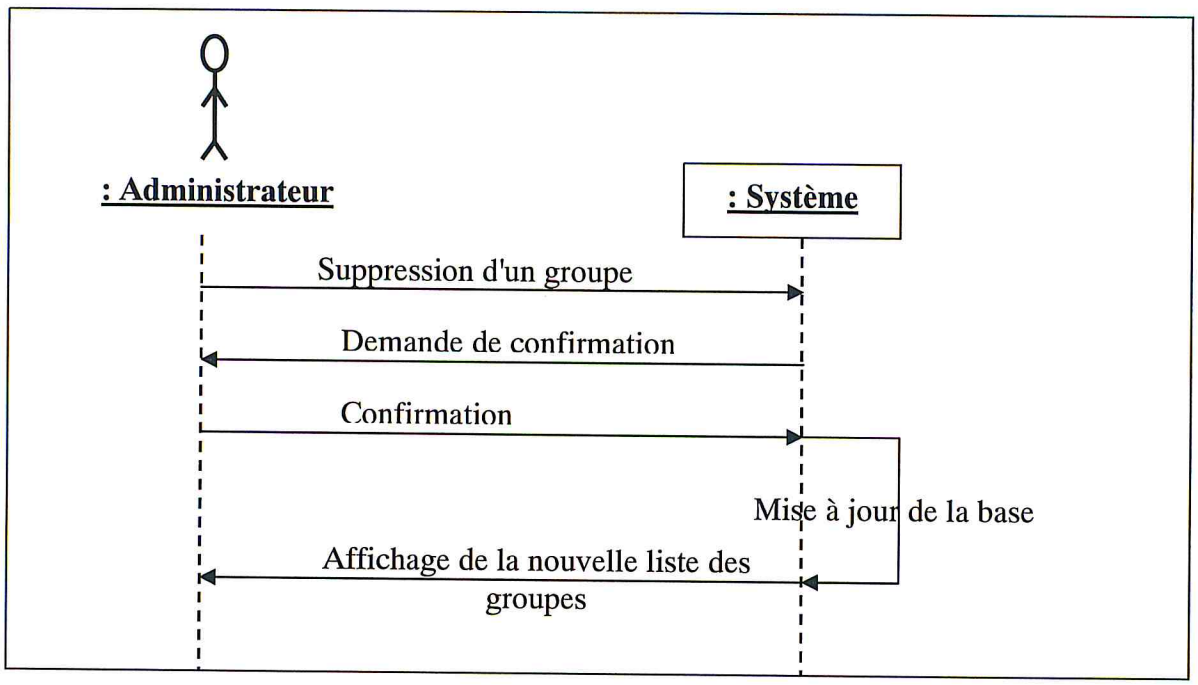


Figure IV.9 Cas d'utilisation «Suppression d'un groupe»

➢ **Modification des membres d'un groupe**

- L'administrateur sélectionne le groupe à modifier de la liste des groupes.
- Le système affiche la liste des membres du groupe.
- L'administrateur sélectionne un membre et déclenche l'opération de suppression/ajout.
- Le système met à jour la base et affiche la nouvelle liste.

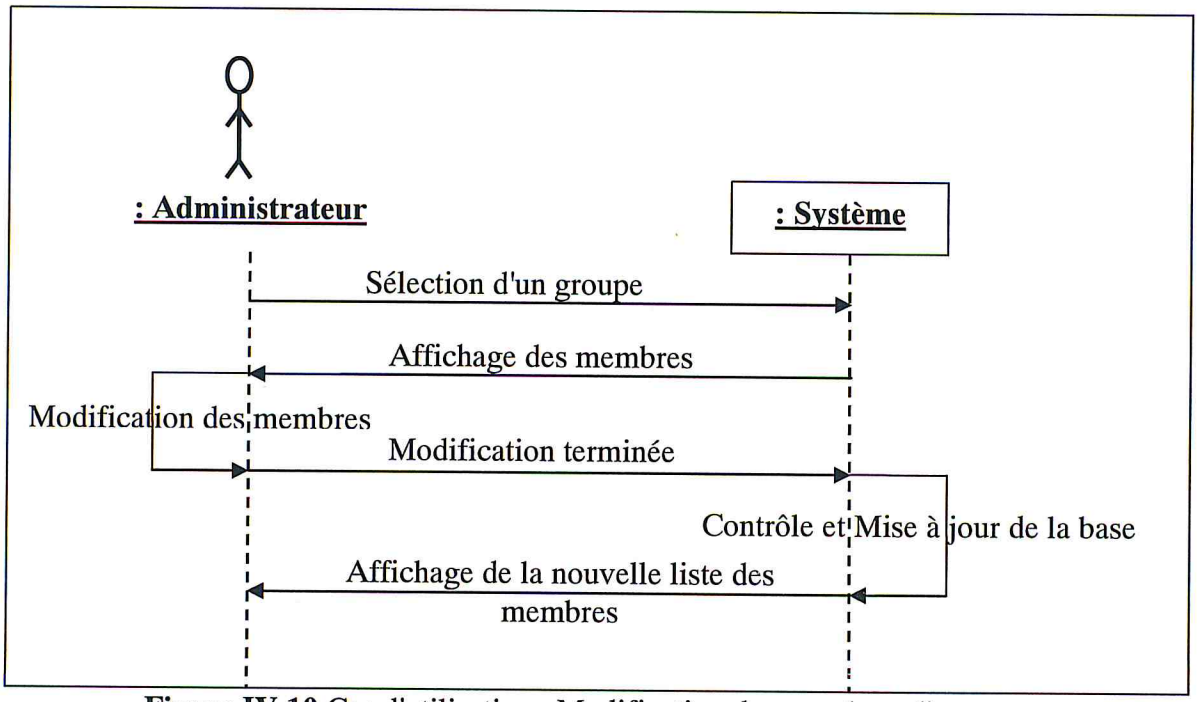


Figure IV.10 Cas d'utilisation «Modification des membres d'un groupe»

➢ **Lancement de l'indexation automatique**

- L'administrateur déclenche l'opération d'indexation automatique.
- Le système répond par une interface.
- L'administrateur saisi l'intervalle des adresses IP et lance l'indexation.
- Le système effectue l'opération d'indexation.

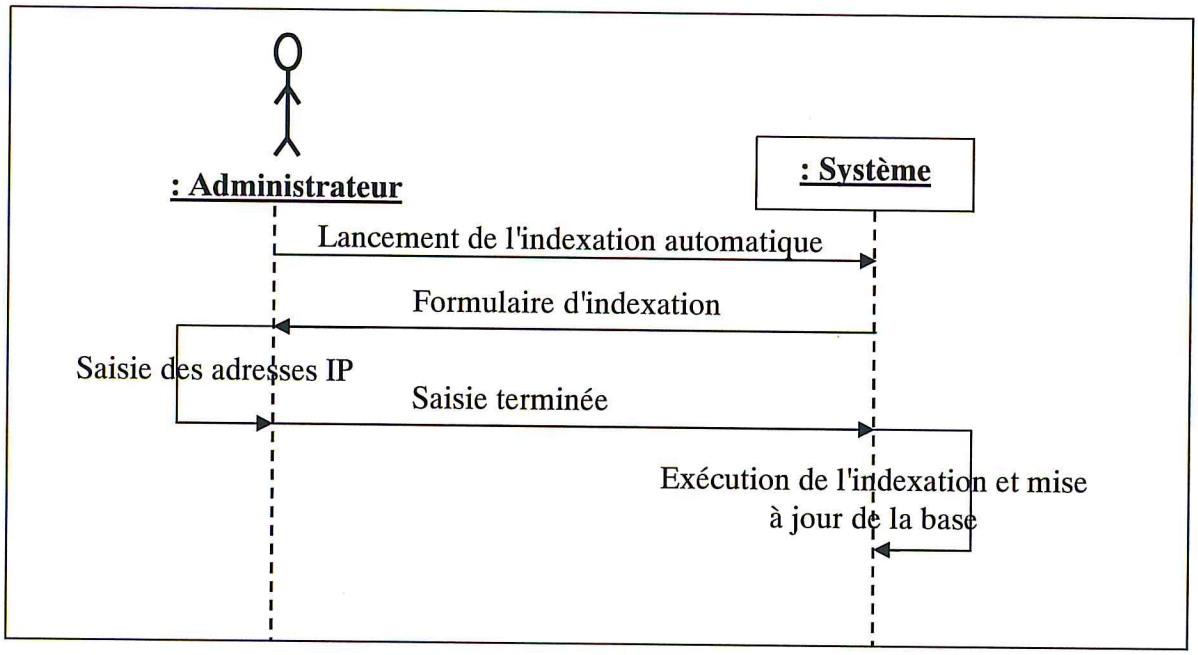


Figure IV.11 Cas d'utilisation «Lancement de l'indexation automatique»

IV.4.4 Description des collaborations

Les fonctions décrites par des cas d'utilisations sont réalisées par la collaboration des objets du domaine. La réalisation de ces cas fait intervenir des objets supplémentaires qui n'appartiennent pas au domaine d'application mais qui sont nécessaire à son fonctionnement, et qui assurent généralement l'interfaçage entre le système et ses acteurs. **[MULL 00]**

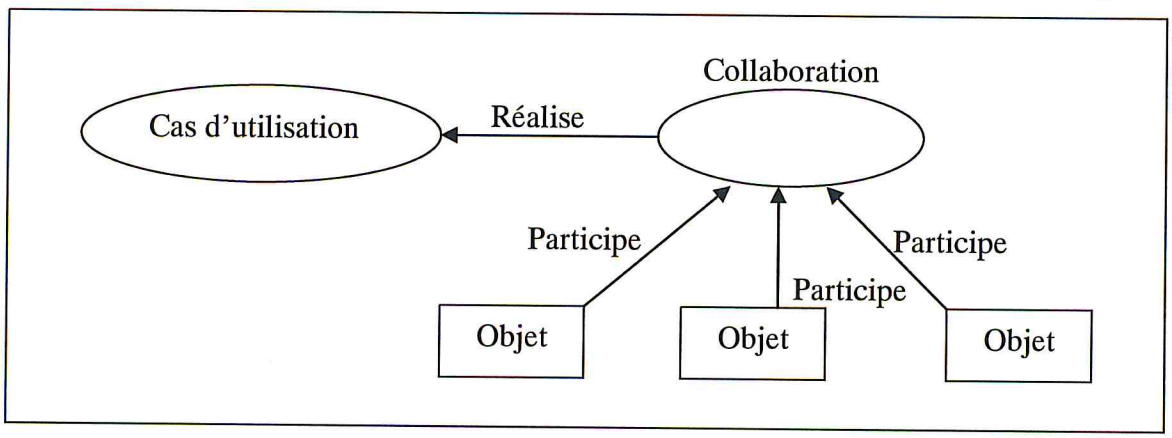


Figure IV.12 Réalisation d'un cas d'utilisation par collaboration d'objets du domaine

Les interfaces utilisateur peuvent être décrites au moyen des classes qui représentent les différentes fenêtres. Par convention les objets de l'interface sont préfixés par la lettre I.

- **Ajout d'un utilisateur**

L'administrateur déclenche l'opération « Ajout d'un utilisateur », par le biais de l'interface « Administration » dans le menu « Utilisateurs », la fenêtre « Ajouter un utilisateur» s'affiche afin que l'administrateur puisse saisir les différents champs (Nom, Prénom, Login, Mot de passe, groupe), à la fin de la saisie l'administrateur valide l'opération, revenant ainsi à l'interface « Ajouter utilisateur » où le nouvel utilisateur est affiché et la table des utilisateurs sera mise à jour.

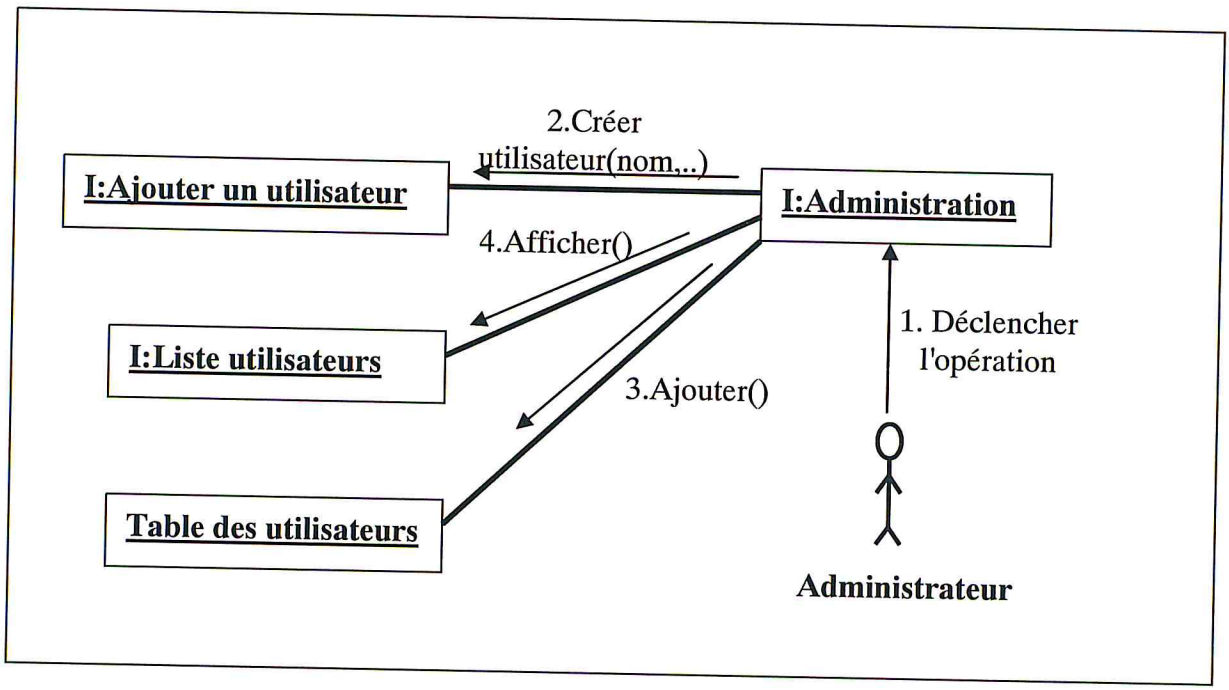


Figure IV.13 Réalisation de l'ajout d'un utilisateur par collaboration entre objets

- **Suppression d'un utilisateur**

L'administrateur déclenche l'opération « Suppression d'un utilisateur », par le biais de l'interface « Administration » dans le menu « Utilisateurs », la fenêtre « Supprimer un utilisateur » s'affiche afin que l'administrateur puisse sélectionner l'utilisateur à supprimer. L'administrateur valide la suppression revenant ainsi à l'interface « Supprimer utilisateur» où la liste et la table des utilisateurs seront mis à jour.

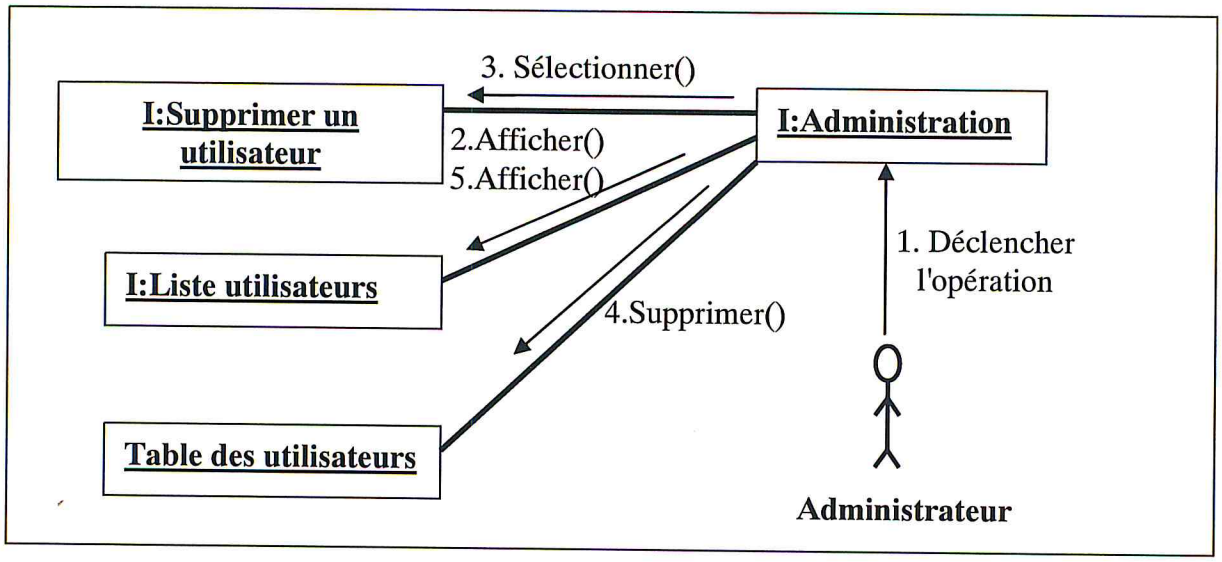


Figure IV.14 Réalisation de la suppression d'un utilisateur par collaboration entre objets

➢ **Modification des informations d'un utilisateur**

L'administrateur déclenche l'opération par le biais de l'interface«Administration» dans le menu « Utilisateurs », la fenêtre « Modifier un compte utilisateur » s'affiche afin que l'administrateur puisse sélectionner un utilisateur et modifier son login et/ou son groupe. A la fin de la saisie l'administrateur valide l'opération revenant ainsi a l'interface « Modifier un compte utilisateur » où la liste et la table des utilisateurs seront mis à jour.

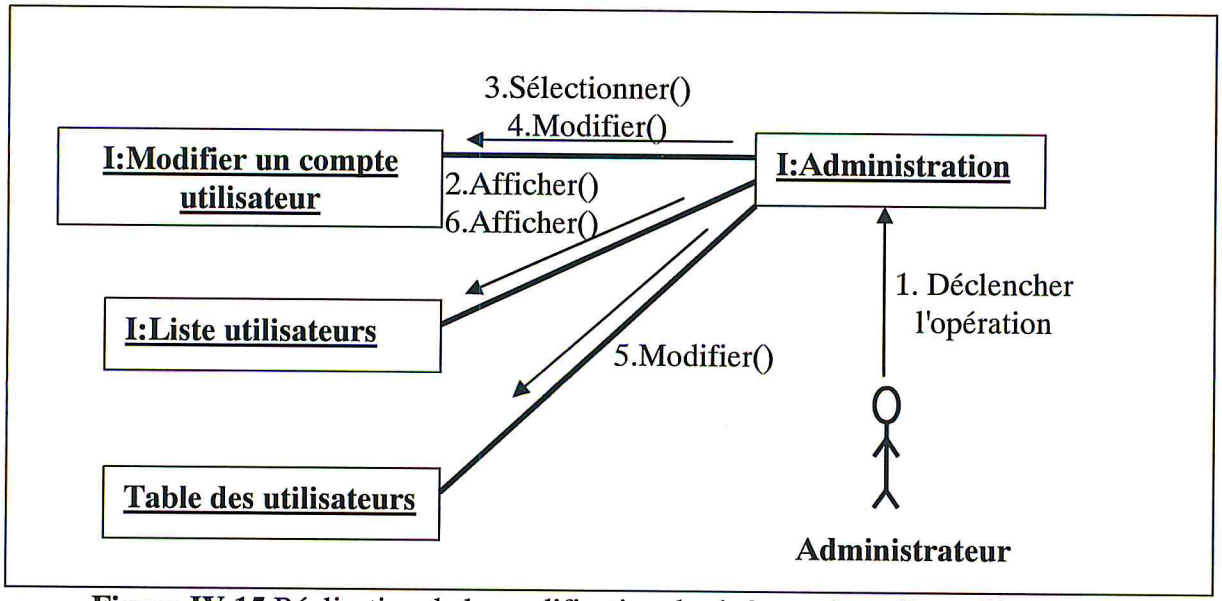


Figure IV.15 Réalisation de la modification des informations d'un utilisateur par collaboration entre objets

➢ **Ajout d'un groupe**

L'administrateur déclenche l'opération par le biais de l'interface «Administration» dans le menu « Groupes », la fenêtre « Ajouter un groupe » s'affiche afin que l'administrateur puisse saisir les différents champs. A la fin de la saisie l'administrateur valide l'opération,

revenant ainsi à l'interface « Ajouter un groupe » où le nouveau groupe est affiché et la liste des groupes sera mise à jour.

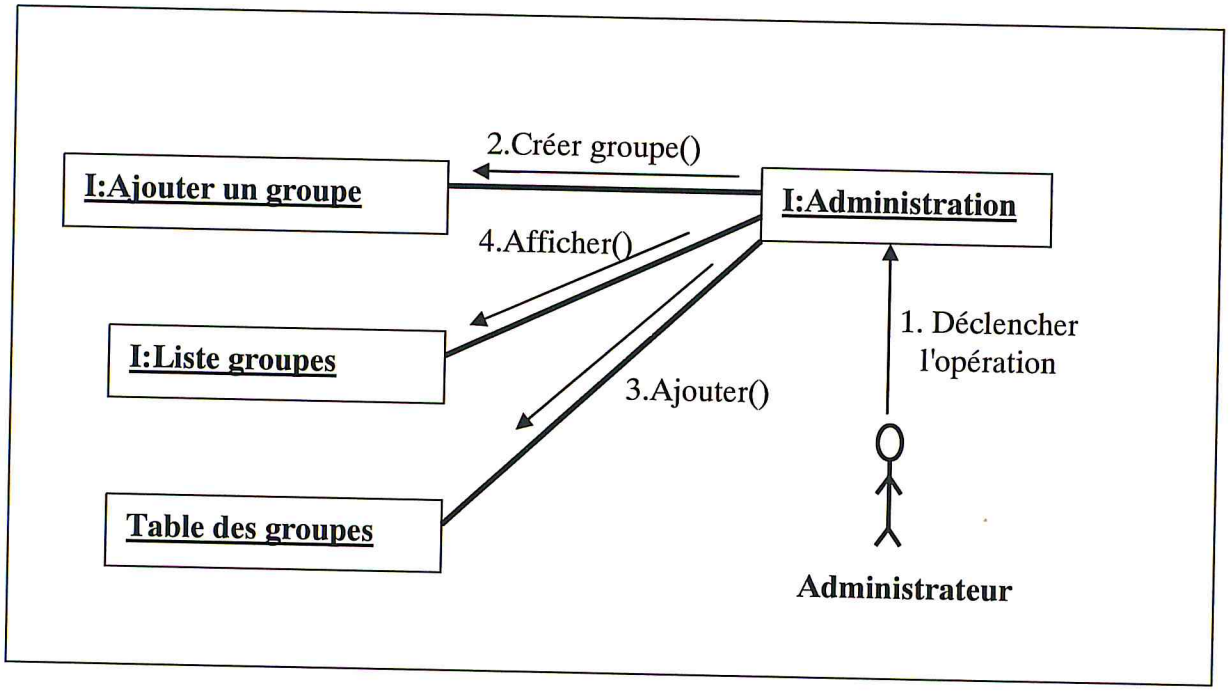


Figure IV.16 Réalisation de l'ajout d'un groupe par collaboration entre objets

- **Suppression d'un groupe**

L'administrateur déclenche l'opération par le biais de l'interface «Administration » dans le menu « Groupes », la fenêtre « Supprimer un groupe » s'affiche afin que l'administrateur puisse sélectionner le groupe à supprimer. L'administrateur valide la suppression revenant ainsi à l'interface « Supprimer un groupe » où la liste des groupes sera mise à jour.

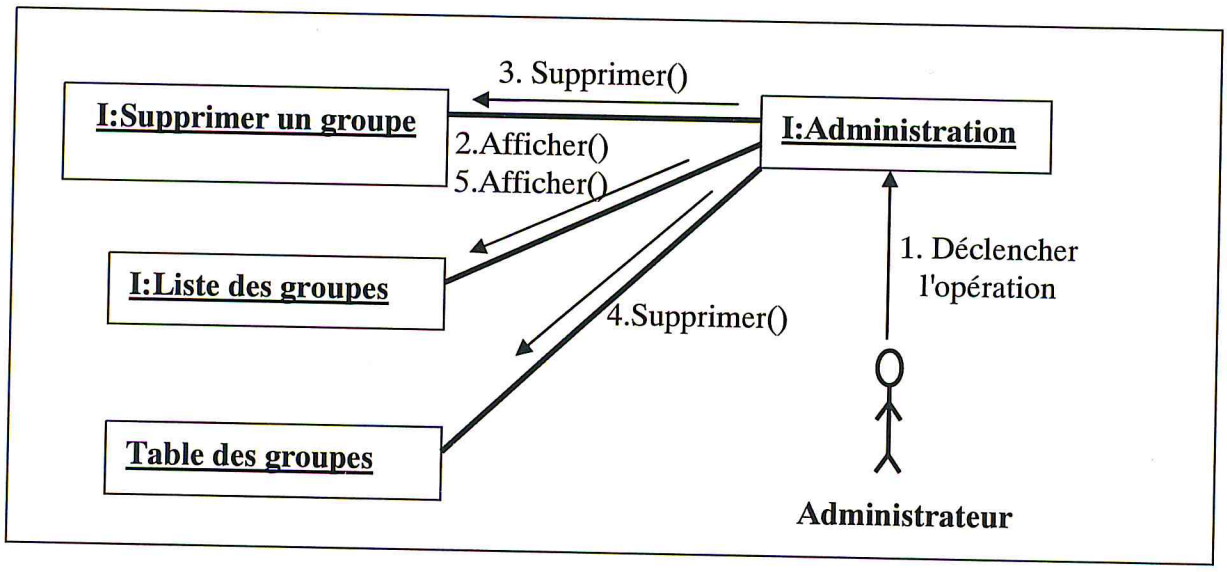


Figure IV.17 Réalisation de la suppression d'un groupe par collaboration entre objets

➢ **Modification des membres d'un groupe**

L'administrateur déclenche l'opération par le biais de l'interface «Administration » dans le menu « Groupes », la fenêtre « Groupes » s'affiche afin que l'administrateur puisse sélectionner le groupe à modifier. La liste des membres du groupe s'affiche, l'administrateur sélectionne le(s) membre(s) à supprimer, valide la suppression et met à jour la liste des membres affichée. Si l'administrateur veut ajouter des membres au même groupe, l'interface « Ajout membre» sera affichée (elle contient la liste des utilisateurs qui n'appartiennent pas au groupe) afin que l'administrateur puisse sélectionner les membres à ajouter. L'administrateur valide l'opération en mettant à jour la liste des membres du groupe.

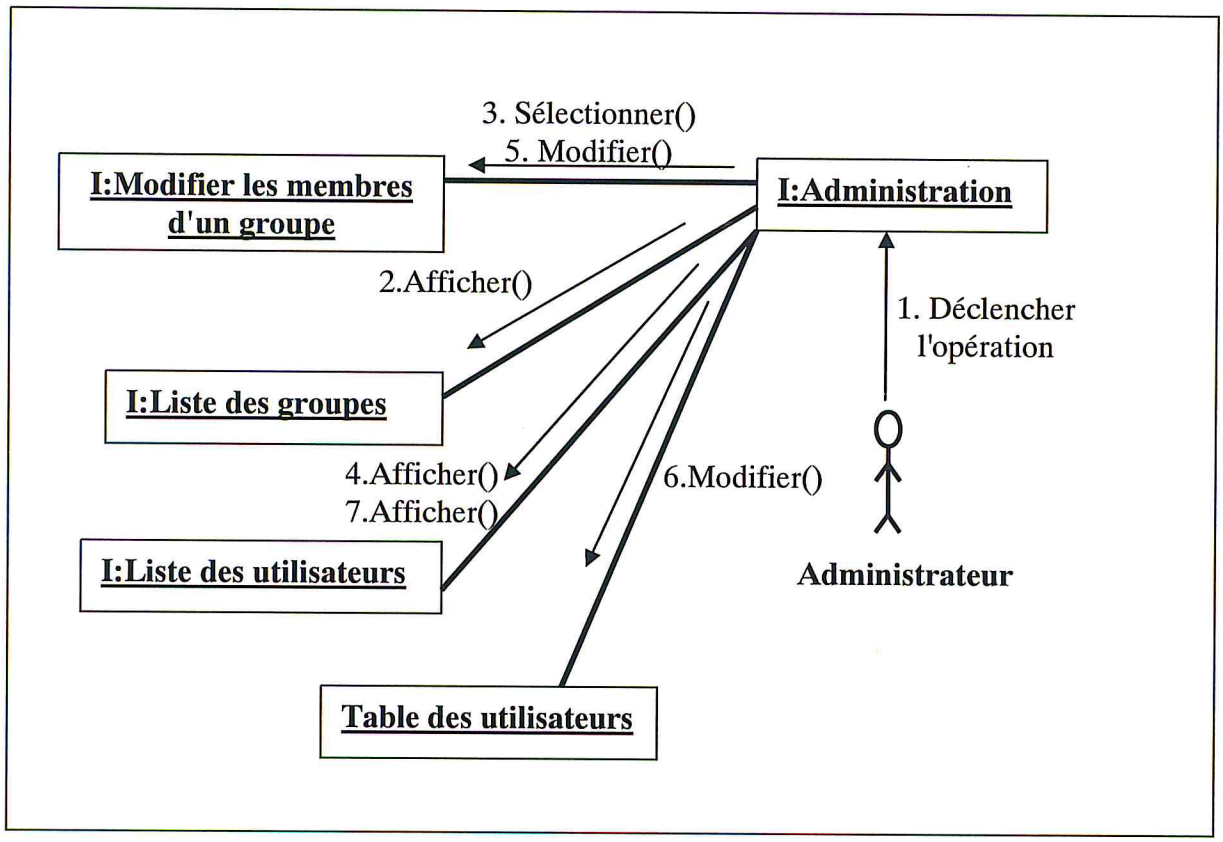


Figure IV.18 Réalisation de la modification des membres d'un groupe par collaboration entre objets

➢ **Lancement de l'indexation**

L'administrateur déclenche l'opération par le biais de l'interface «Administration » dans le menu « Indexation », la fenêtre « Indexation » s'affiche afin que l'administrateur puisse saisir l'intervalle des adresses IP et lancer l'indexation qui permet la mise à jour de l'index du système.

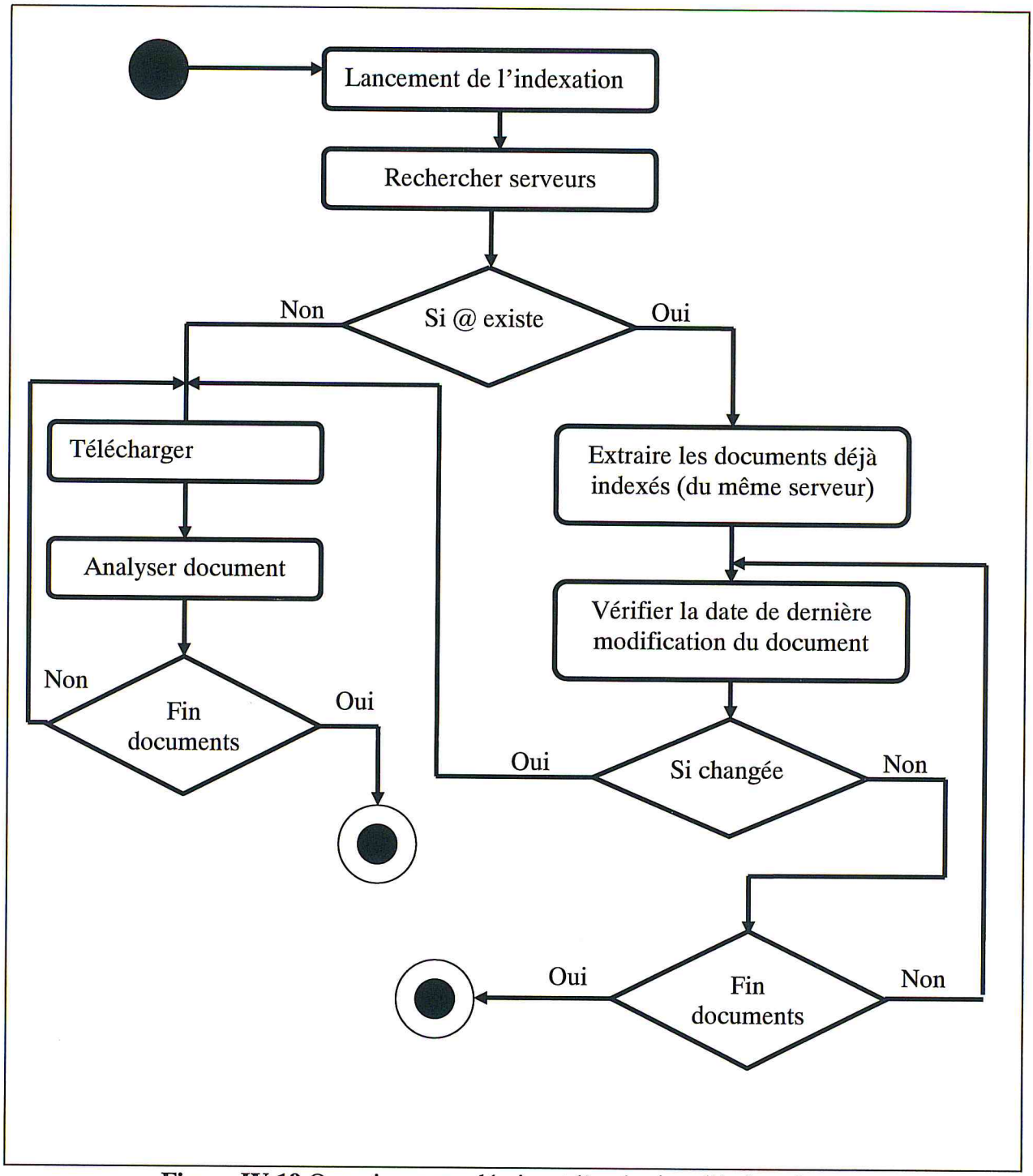


Figure IV.19 Organigramme décrivant l'opération d'indexation

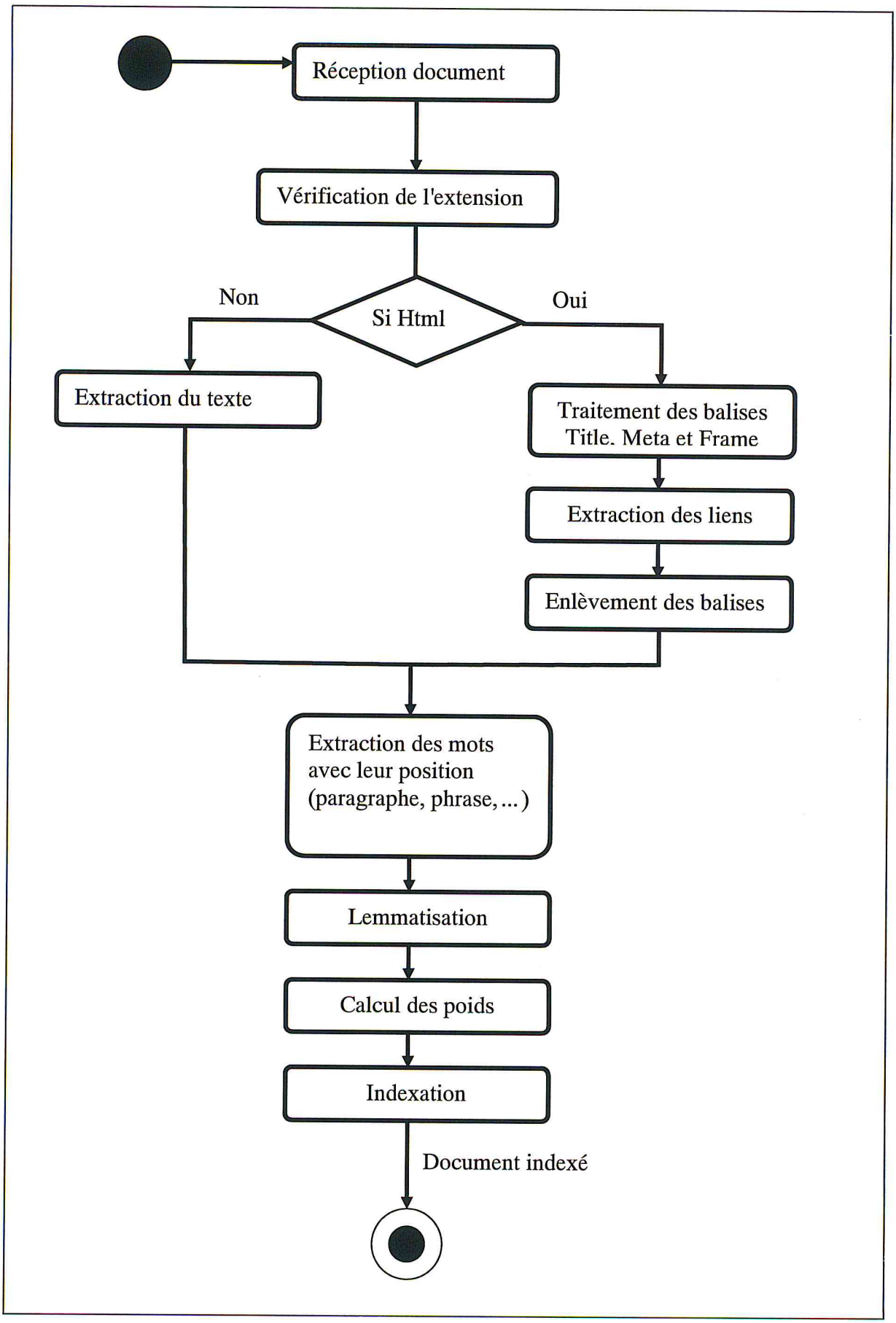


Figure IV.20 Description de la fonction "Analyse"

IV.5 L'INTERFACE WEB

IV.5.1 Détermination des cas d'utilisations

Les acteurs du système sont :

- Le client web : acteur manipulant le système en effectuant des recherches ou en publiant des documents.
- Le serveur d'application : La relation entre le serveur d'application et le client web se limite à la communication via des requêtes URL.
- Le spider : La relation entre le spider et le client web est l'indexation des documents publiés par ce dernier.

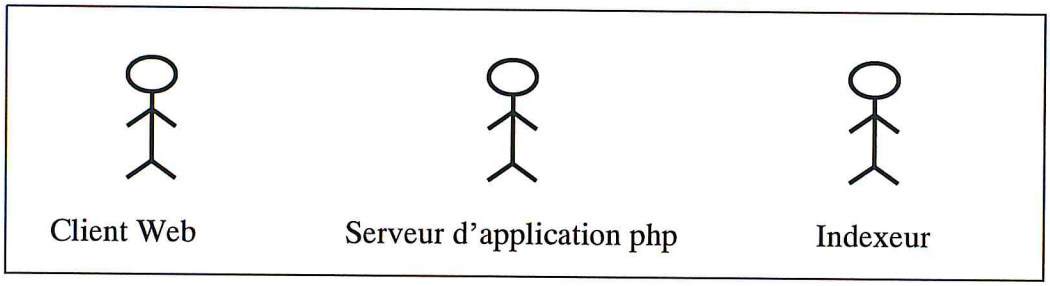


Figure IV.192 Les acteurs du système INTERFACE WEB

Acteur	Cas d'utilisation
Client web	• Identification. • Recherche documentaire publique. • Recherche documentaire privée. • Publication des documents. • Suppression des publications. • Modification des publications. • Modification du compte.

Tableau IV.2 Cas d'utilisations du système INTERFACE WEB

IV.5.2 Diagramme des cas d'utilisations

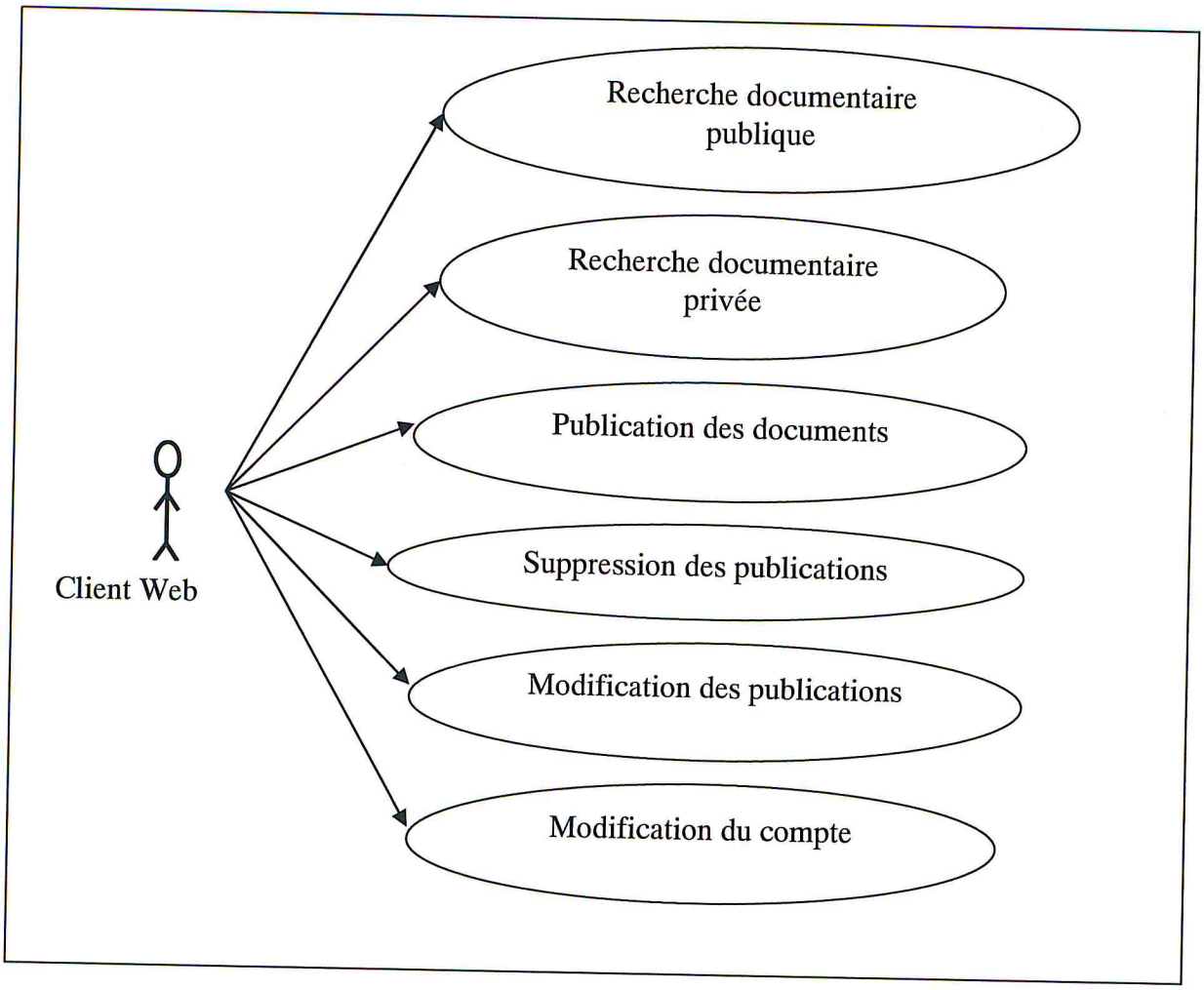


Figure IV.203 Diagramme des cas d'utilisations du système Interface Web

IV.5.3 Description des cas d'utilisations

- **Recherche documentaire publique**

  - Le client saisi la requête.
  - Le système traite la requête (élimination des blancs), l'analyse, recherche les documents correspondants, les trient et les affichent au client.

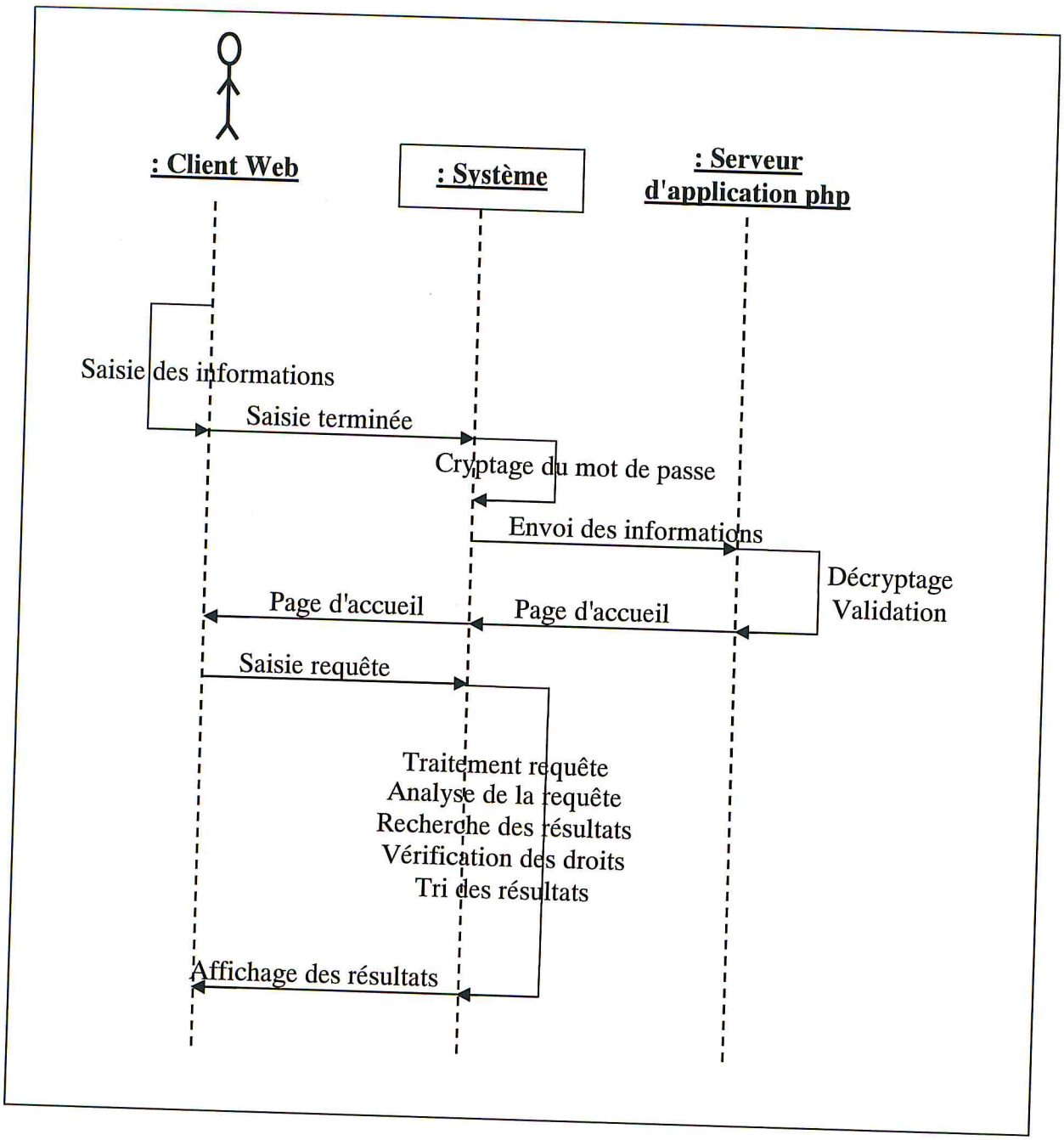


Figure IV.225 Cas d'utilisation «Recherche documentaire privée»

➢ **Publication documentaire**

- Le client demande la page de publication.
- Le serveur Web envoi la page au client.
- Le client sélectionne le mode d'indexation (manuel ou automatique).
- Le serveur web envoi la page correspondante.
- Le client saisi les informations nécessaires.
- Le client envoi les informations au serveur web.
- Le serveur envoi ces informations au spider.

- Le spider traite les informations, fait la mise à jour et envoi un accusé au serveur web.
- Le serveur web envoi un accusé au client.

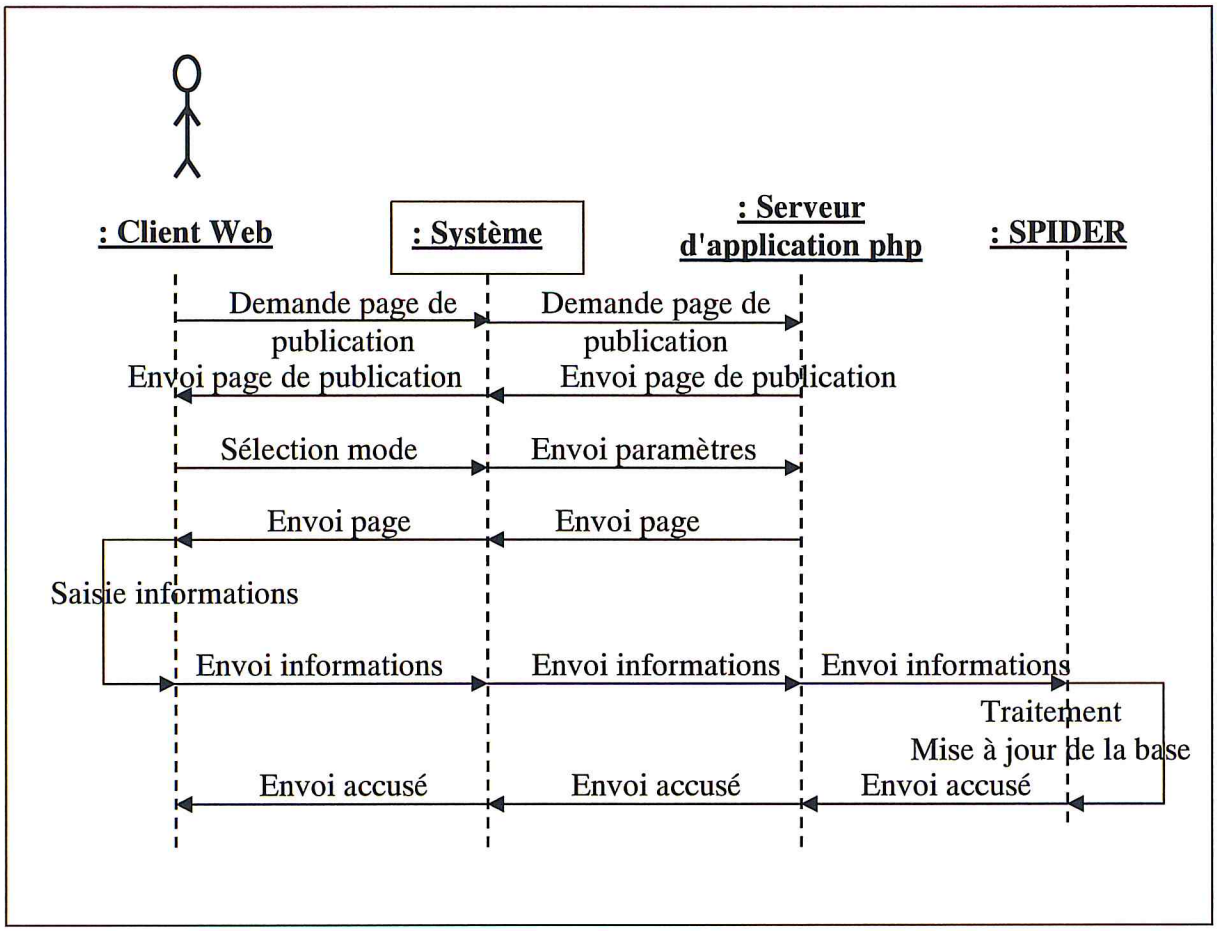


Figure IV.236 Cas d'utilisation «Publication documentaire»

➢ **Suppression des publications**

- Le client sélectionne le(s) document(s) à supprimer.
- Le système demande la confirmation.
- Le client déclenche l'opération de suppression.
- Le système met à jour la base.

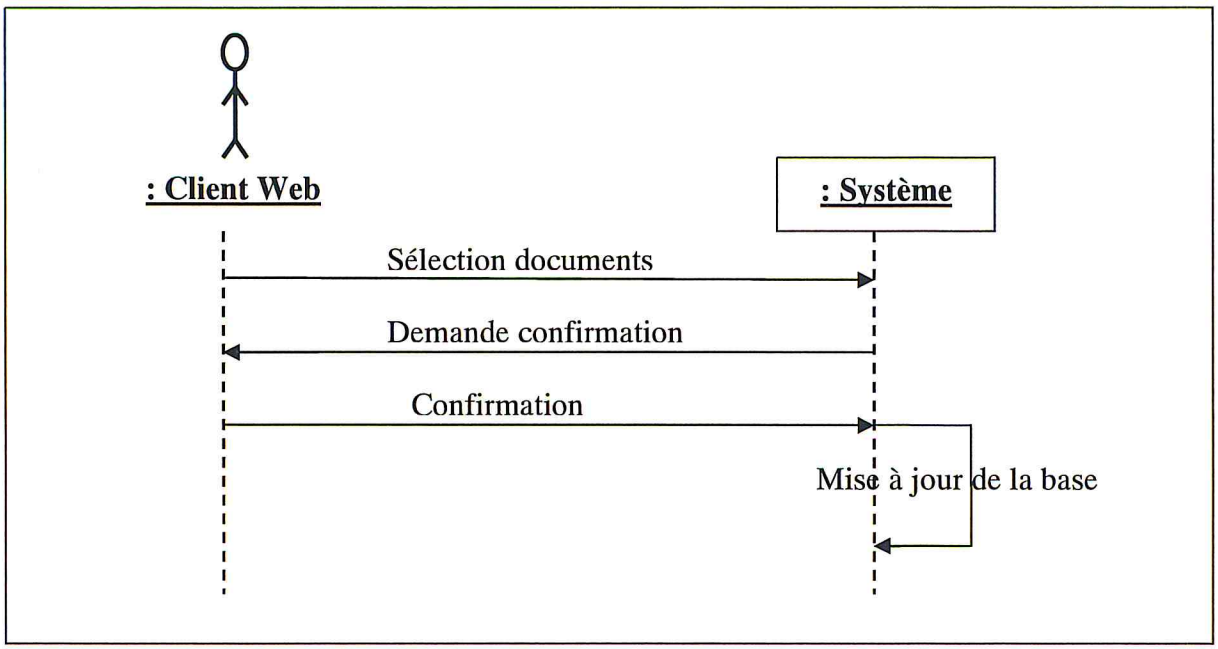


Figure IV.27 Cas d'utilisation «Suppression de publications»

➢ **modification des publications**

- Le client demande la page de modification.
- Le serveur web envoi la page de modification.
- Le client sélectionne le document à modifier.
- Le système affiche les informations correspondantes.
- Le client modifie les informations.
- Le système met à jour la base et affiche la liste des documents rafraîchie.

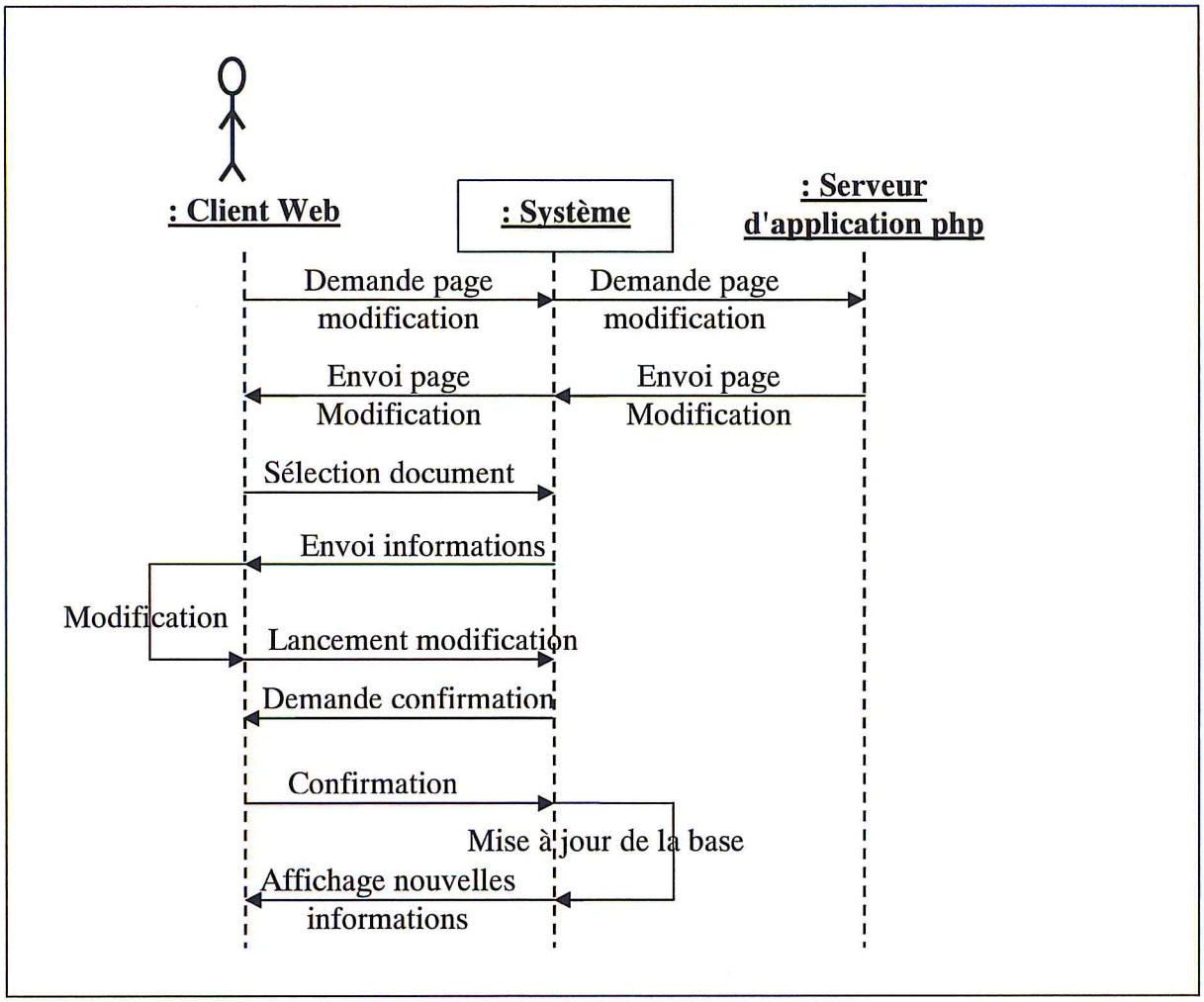


Figure IV.28 Cas d'utilisation «Modification des publications»

➢ **Modification du compte**

- Le client s'identifie.
- Le client demande la page de modification.
- Le serveur d'application envoi la page.
- Le client modifie ses informations (Login, Password).
- Le système vérifie si le Login n'existe pas.
- Si la vérification se termine avec succès, le système effectue la mise à jour de la base ainsi que les variables de session, sinon une page d'erreur est envoyée au client.

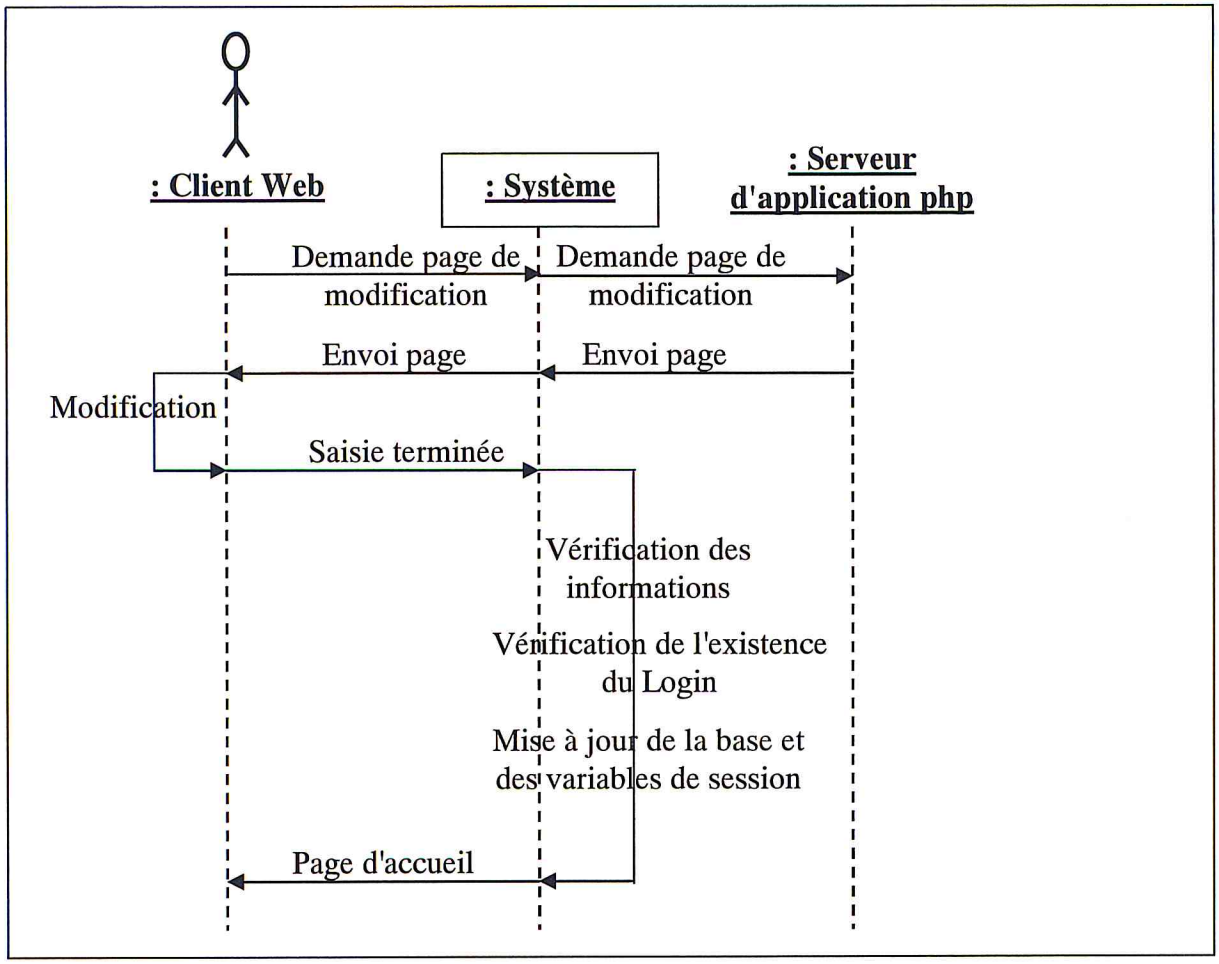


Figure IV.29 Cas d'utilisation «Modification du Login»

IV.5.4 Schéma global du site

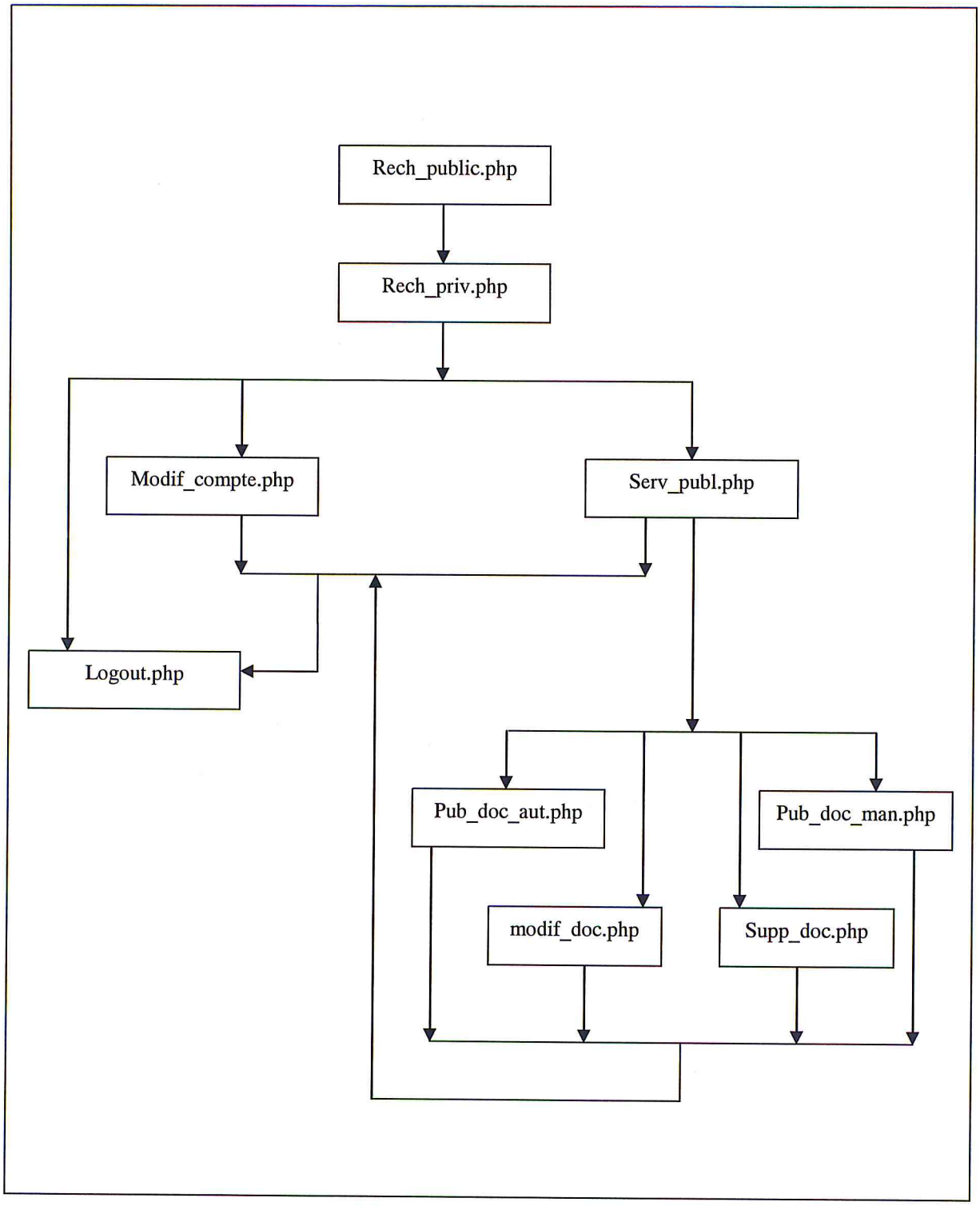


Figure IV.240 Schéma global du site Web

- **Les partages** : Stockés dans la table « Partage » selon les identifiants des documents.

Id_document	Id_groupe

Tableau IV.8 Table partage

- **Les droits d'accès** : Stockés dans la table « Appartient » selon le "login" de l'utilisateur.

Login	Id_groupe

Tableau IV.9 Table appartient

- **Les correspondances entre les documents et le mots clés (Index)** : Stockés dans la table « Contient_mot » selon les identifiants des documents.

Id_document	Mot	Position

Tableau IV.10 Table contient_mot

IV.7 DIAGRAMME DE FLUX DES DONNEES

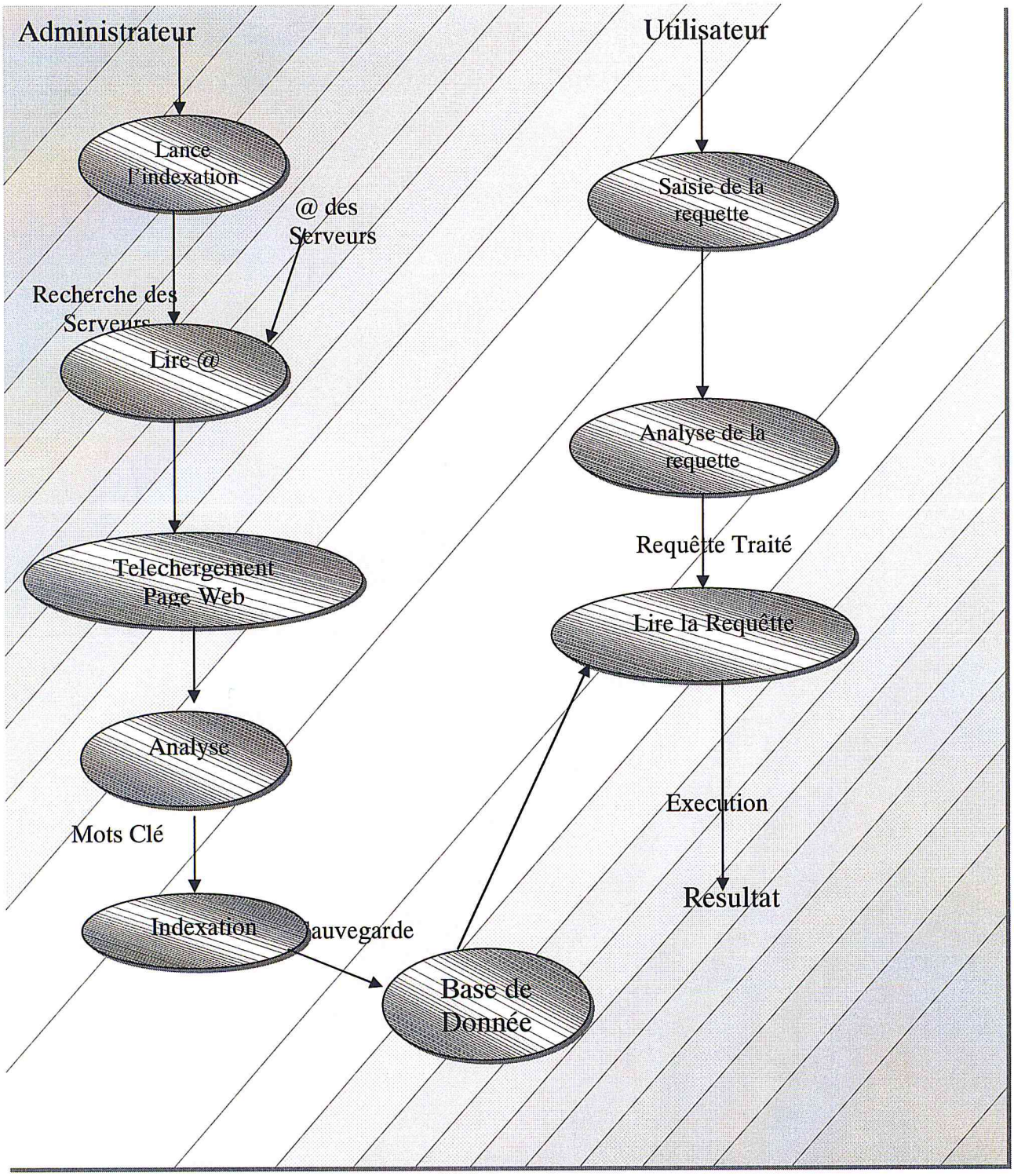


Figure IV.32 Diagramme de flux des Données

IV.6 PERSISTANCE DES DONNEES

Dans notre application, les données sont stockées dans les tables Oracle suivantes :

- **Les documents** : Stockés dans la table « Document ».

Id_document	Url	IP	Login	Taille	Dernière_modif

Tableau IV.3 Table document

- **Les serveurs** : Stockés selon dans la table « Serveur » suivant les adresses IP.

IP	Web	Ftp	Infos	Dns

Tableau IV.4 Table serveurs

- **Les utilisateurs** : Stockés dans la table « Utilisateur ».

Login	Password	Nom	Prénom

Tableau IV.5 Table utilisateur

- **Les groupes** : Stockés dans la table « Groupe ».

Id_groupe	Libelle_groupe

Tableau IV.6 Table groupe

- **Les mots clés** : Stockés dans la tale « Mot clé » qui contient une seule colonne.

Mot

Tableau IV.7 Table mot_ cles

IV.8 ARCHITECTURE

IV.3.1 Architecture logicielle

Les objets de domaines se regroupent en quatre principaux paquetages :

- Le spider.
- Le serveur d'application PHP.
- Le client Web.
- Le module de persistance qui représente les informations tirées des documents après leur indexation.

IV.3.2 Architecture matérielle

Les différents paquetages seront déployés comme suit :

- Le spider sur un poste de travail.
- Le serveur Web sur le même poste ou sur un poste à double réseau.

Les clients web sur des postes équipés d'un browser.

IV.6.1 Le modèle logique

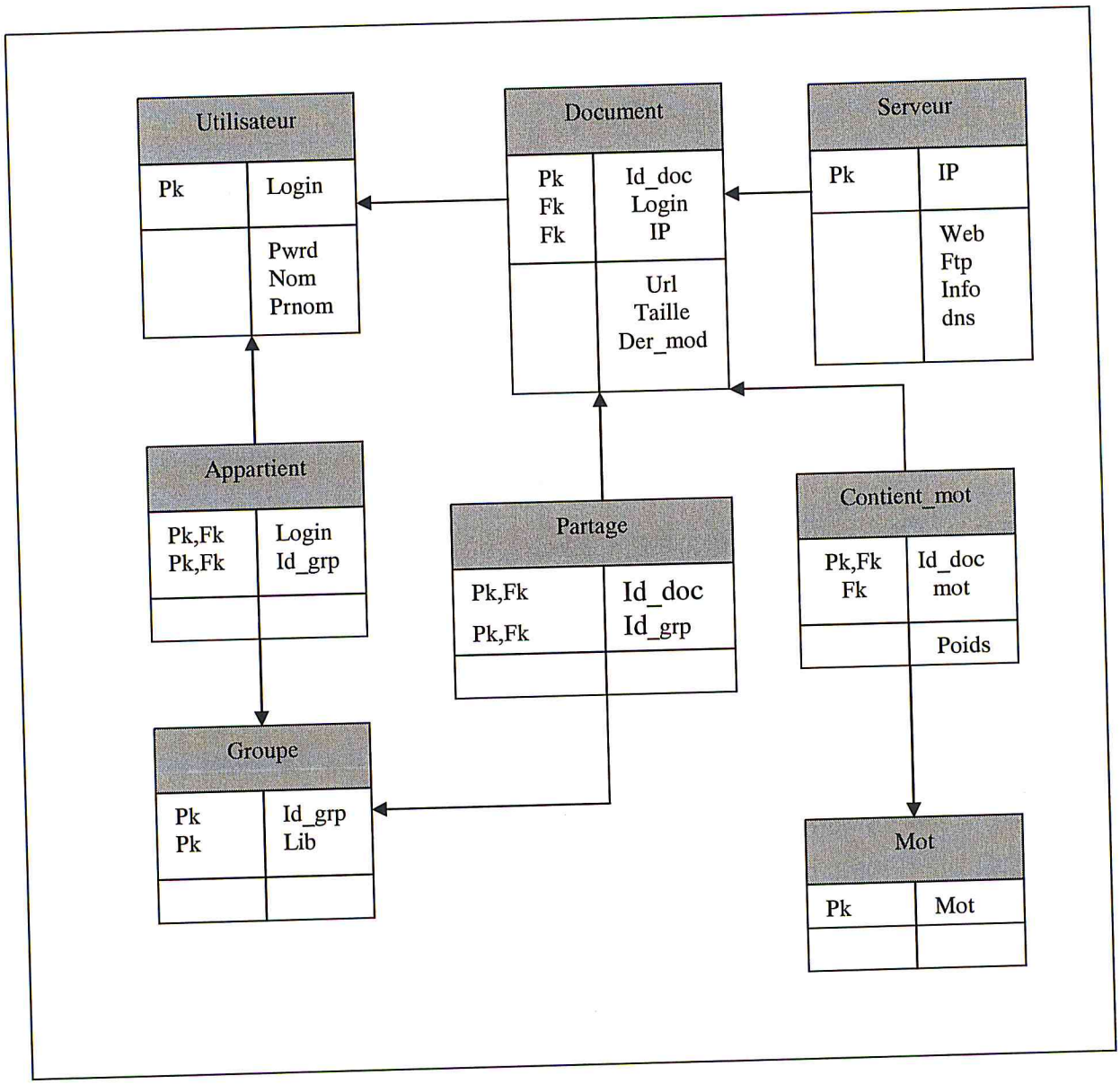


Figure IV.251 Le modèle logique

Chapitre V

MISE EN ŒUVRE

Dans ce chapitre :

- Environnement de développement
- Le système « Uni_Blida SEARCH »

V.1 ENVIRONNEMENT DE DEVELOPPEMENT

V.1.1 Environnement logiciels

Afin que le Robot puisse assurer les fonctionnalités requises, l'installation de la plate-forme .NET est indispensable sur le serveur qui hébergera l'application. Pour le moment, les applications développées par le Framework .NET fonctionnent exclusivement dans un environnement Windows. Cependant, MicroSoft envisage la possibilité d'immigration de ces applications vers d'autres systèmes (UNIX par exemple) en travaillant sur la portabilité du .NET.

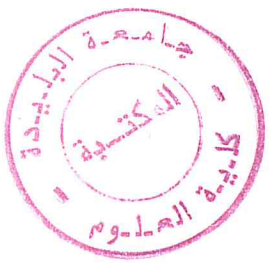
V.1.2 Langages de programmation

Pour l'implémentation de notre système, nous exploitons trois environnements logiciels, à savoir celui de la localisation et l'indexation des documents « Robot», celui de l'interface de recherche pour les utilisateurs « Site du moteur de recherche » et le troisième pour la base de données.

❖ **Choix du langage de programmation du Robot (Spider)**

Etant donné que le logiciel sera développé dans un environnement réseau, ses modules devront assurer :

- La manipulation des classes.
- La manipulation des threads.
- La manipulation des sockets.
- La manipulation des bases de données.
- La création des services windows.



Pour l'implémentation du SPIDER, le système de gestion des comptes utilisateurs et d'administration de la base de données, nous avons utilisé **VB.Net**, un langage de programmation élaboré par la firme MicroSoft et qui se base sur sa nouvelle plate-forme de développement des applications .NET. En gardant un peu de ressemblances avec le VB classique, MicroSoft a apporté beaucoup de changements dans la version .Net. Les modifications apportées sont telles qu'on peu parler d'un nouveau langage. VB.NET est maintenant un véritable langage orienté objet, avec un déploiement simplifié autour d'un environnement de développement cohérent qui permet le développement d'applications traditionnelles, client serveur, et web **[URL02]**.

VB.NET procure aux développeurs un langage facile et un outil productif pour créer rapidement des applications pour Microsoft Windows et le Web. VB.NET offre des concepteurs visuels améliorés, des performances accrues au niveau des applications et un puissant environnement de développement intégré (IDE) qui met le développeur sur la voie rapide du développement d'applications. Sa puissance réside dans les points suivants :

- Création facile d'applications Web : Il permet aux développeurs d'exploiter leurs compétences pour créer de véritables applications Web pour clients légers et les contrôles WebForms côté serveur permettent un fonctionnement correct des applications Web sur n'importe quel serveur, quelle que soit la plate-forme utilisée .
- Simplicité et flexibilité d'accès aux données : VB.Net permet un accès aux données d'une grande souplesse et hautement évolutif avec la classe ADO.Net qui permet de récupérer et de manipuler des données facilement et rapidement. Il est adapté à des applications web qui utilisent des accès à des bases de données en mode connecté et déconnecté.
- Une interface facile à utiliser et un grand nombre de fonctionnalités intégrées.
- VB.Net se base sur la plate-forme **.NET** qui lui fournit une très grande possibilité dans le développement des applications. Diverses méthodes sont désormais intégrées dans des lignes de code réduites, ce qui facilite la tache pour le développement en réduisant la taille du code ce qui permet une meilleur lisibilité.

❖ Choix du langage de programmation des pages du site Web

Etant donné que les pages du site devront être de type dynamique permettant ainsi un affichage dynamique des pages d'après le trafic échangé (requêtes et réponses) et accédant aux informations relatives aux documents indexés et aux utilisateurs de notre moteur de recherche, l'utilisation de langages de script exécutés du côté serveur tel que PHP et exécutés du côté client tel que JAVASCRIPT, s'est imposée.

➢ PHP

PHP a été créé en 1994 par Rasmus Lerdorf pour les besoins des pages web personnelles. A l'époque il s'appelait : Personnal Home Page. En 1997, son interpréteur est réécrit par Zeev Curaski et Andi Gutman et il devient ainsi "PHP: Hypertext Preprocessor".

C'est un langage de script HTML, exécuté coté serveur. L'essentiel de sa syntaxe est emprunté aux langages C, Java et Perl, avec des améliorations spécifiques. L'objet de ce langage est de permettre aux développeurs Web d'écrire des pages dynamiques rapidement.

Aujourd'hui, les capacités de PHP vont bien au-delà de la génération de pages HTML : PHP génère des documents PDF, des images ou même des animations Flash **[URL03]**.

- Atouts :
    - La simplicité d'écriture de scripts.
    - La possibilité d'inclure le script PHP au sein d'une page Html.
    - La simplicité d'interfaçage avec des bases de données.
    - L'intégration au sein de nombreux serveurs Web (Apache, Microsoft IIS).
    - La gratuité et la disponibilité du code source.

➢ **JavaScript**

JavaScript est un langage de scripts qui incorporé aux balises Html, permet d'améliorer la présentation et l'interactivité des pages Web. JavaScript est donc une extension du code Html des pages Web. Les scripts, qui s'ajoutent aux balises Html, peuvent en quelque sorte être comparés aux macros d'un traitement de texte. Ces scripts vont être gérés et exécutés par le browser lui-même sans faire appel aux ressources du serveur. Ces instructions seront donc traitées en direct et surtout sans retard par le navigateur.

JavaScript a été initialement développé par Netscape et s'appelait alors LiveScript. Adopté à la fin de l'année 1995, par la firme Sun, il prit alors son nom de JavaScript.

- Avantages :
    - Le JavaScript est plus simple à mettre en oeuvre car s'agit de code ajouté à une page.
    - écrite en html, avec par exemple un simple éditeur de texte.
    - Les codes JavaScript ne ralentissent pas le chargement d'une page.
- Inconvénients:
    - Le champ d'application de JavaScript est assez limité.
    - Comme le code JavaScript est inclus dans une page html, celui-ci est visible et peut être copié par tout le monde.

❖ **Choix du SGBD pour la base de données**

Pour l'environnement de la base de données, nous avons utilisé le SGBD « Oracle» qui est le leader mondial des systèmes de gestion de bases de données. Il supporte le modèle relationnel, et assure un haut niveau de sécurisation de la base ainsi que la gestion des accès concurrents tout en optimisant la taille, le temps d'accès et le temps d'exécution des requêtes

V.2 LE SYSTEME « UNI_BLIDA SEARCH »

Le moteur de recherche permet d'accéder à l'information de façon directe, il va rapporter tous les documents en sa connaissance qui parlent des termes de la requête de l'utilisateur. C'est un domaine de recherche très actif mais difficile à mettre en oeuvre. Le choix de la méthode d'indexation reste un facteur décisif pour la qualité des résultats obtenus.

Les travaux présentés dans ce mémoire ont pour objectif d'analyser ces diverses difficultés et adopter une méthode pour les processus d'indexation, de recherche et de présentation des résultats. Un utilisateur voudra toujours disposer d'une information où qu'il soit, où qu'elle soit et au plus tôt possible. Quand aux producteurs d'informations, ils préfèrent ne pas avoir à se soucier, lors de la rédaction de leurs documents, des contraintes liées au stockage, à la recherche et à la consultation. Les préoccupations des administrateurs des systèmes informatiques s'orientent vers des logiciels nécessitant une maintenance minimum, perturbant le moins possible les utilisateurs et les autres systèmes informatiques.

Afin de faciliter la recherche et la consultation d'informations au sein des réseaux Intranet du l'Université de Blida, nous proposons un système qui concilie la plupart des points évoqués ci-dessus. Il offre une recherche sur l'intégralité des textes contenus dans les documents. La recherche tient compte aussi de l'identité de l'utilisateur afin de ne lui présenter que les documents dont il a le droit de consulter (documents à accès publique ou publiés par d'autres utilisateurs autorisant l'accès à un nombre restreint d'utilisateurs). La consultation des documents se fait depuis n'importe quel poste de travail en réseau, disposant d'un client universel hypertexte (typiquement un navigateur Web). Des liens hypertextes vers les documents conformes à la requête de l'utilisateur, sont cités par ordre de pertinence, si l'utilisateur a le droit d'y accéder.

Le fonctionnement de notre système **Uni_Blida SEARCH** est basé sur les points suivants :

- **Gestion des utilisateurs** par la création de leurs comptes, la gestion des groupes et les éventuels partages de documents.
- **La localisation des documents dans l'intranet** par la recherche des serveurs qui contiennent de la matière informationnelle (serveurs web, ftp...) et offrir aux utilisateurs du système la possibilité de publier des documents qui seront indexés et référencés par notre moteur de recherche et dont les droits d'accès sont définis par les utilisateurs eux-mêmes.

- **L'indexation des documents collectés** permettant d'extraire l'information pertinente de ces derniers. Ceci devrait nécessiter, pour être réellement efficace, une extraction intelligente des structures du document (mots, phrases, titres,...) et une méthode d'analyse du texte et d'extraction des mots qui le représente le mieux.
- **Le traitement de la question, la recherche et le tri** des documents pertinents, qui, encore une fois pour être efficace, nécessite de mettre en oeuvre des outils de traitement des requêtes plus efficaces en proposant diverses options de recherche (recherche booléenne, recherche par troncatures...).
- **La présentation des résultats** la plus riche et synthétique possible, pour que l'utilisateur puisse facilement juger de l'intérêt des documents.

V.2.1 Système de gestion des comptes utilisateurs et des documents

L'exécution de l'application commence tout d'abord par la saisie du mot de passe de l'administrateur de la base de données (DBA). Il peut consulter toutes les informations concernant chaque compte utilisateur. Une fois le mot de passe validé, l'administrateur aura le droit d'apporter des modifications sur ces comptes, en ajouter d'autres, gérer les groupes d'utilisateurs et les documents publiés.

Une fois que l'administrateur s'authentifie, l'interface principale du système s'affiche. Elle comprend les menus suivant :

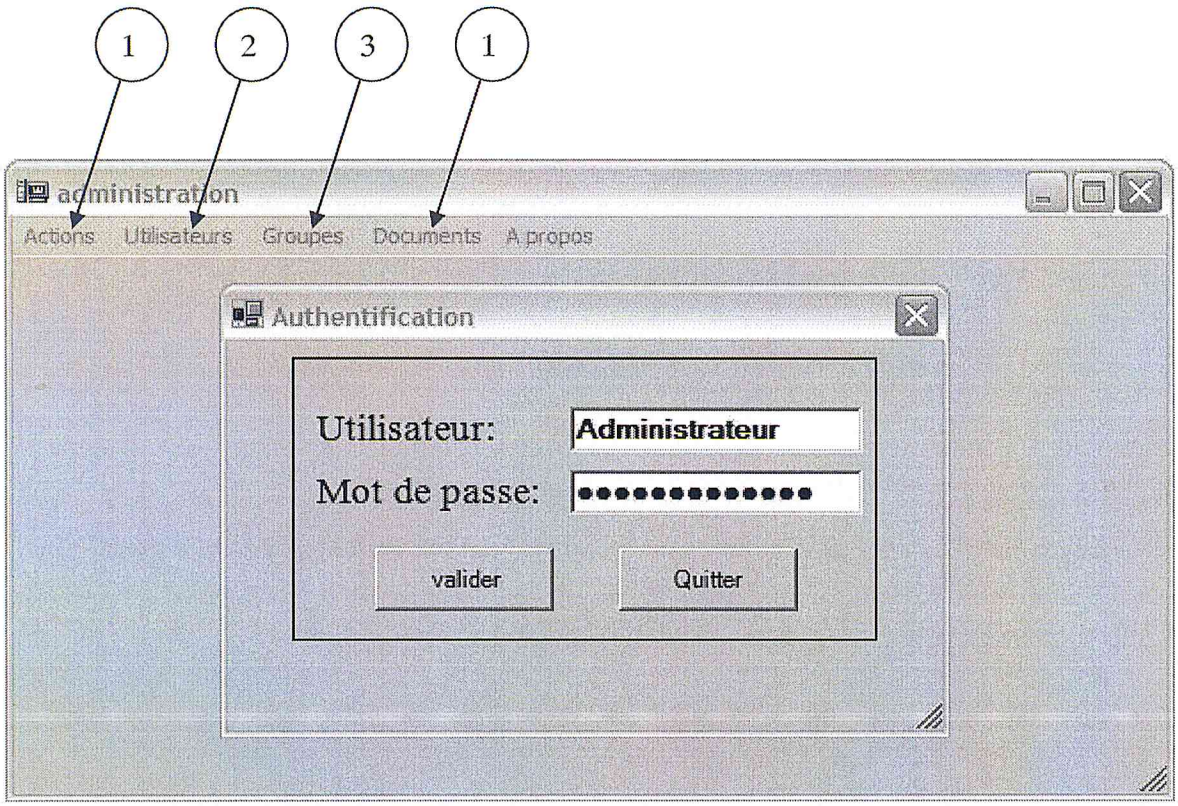


Figure V.1 Présentation générale

1) Le menu Actions permet d'effectuer des opérations comme lancer le formulaire d'authentification de l'administrateur, changer le mot de passe administrateur...
2) Le menu Utilisateurs correspond à la gestion des comptes utilisateurs : ajout d'un nouvel utilisateur, suppression et modification d'un compte utilisateur.
3) Le menu groupes correspond à la gestion des groupes d'utilisateurs : ajout, suppression et modification des membres des groupes existants.
4) Le menu document visualise les informations relatives aux différents documents publiés.

V.2.1.1 Gestion des comptes utilisateurs

L'administrateur peut créer un compte, supprimer ou modifier les informations relatives à un compte :

- **Ajout d'un nouvel utilisateur**

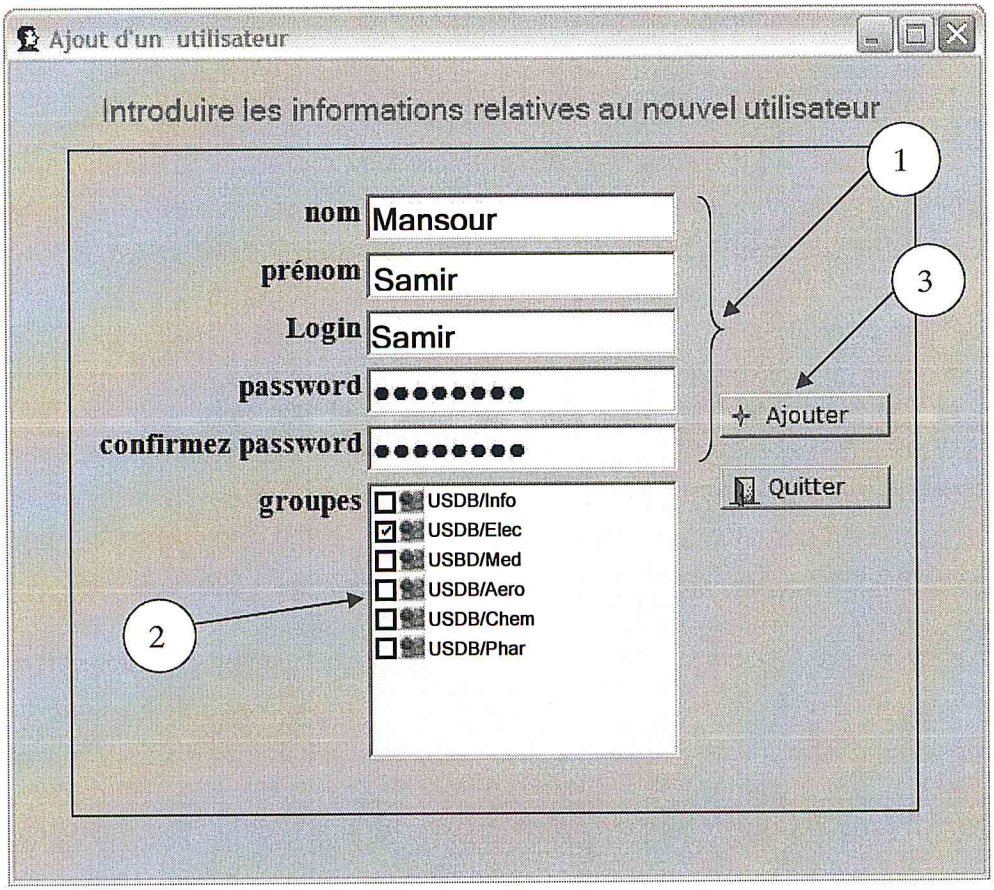


Figure V.2 Ajout d'un nouvel utilisateur

1) Informations relatives à l'utilisateur.
2) Liste des groupes existants avec des cases à cocher permettant la sélection des groupes dans lesquels l'utilisateur sera ajouté.
3) Contrôle et mise à jour de la base.

➢ **Suppression d'un utilisateur**

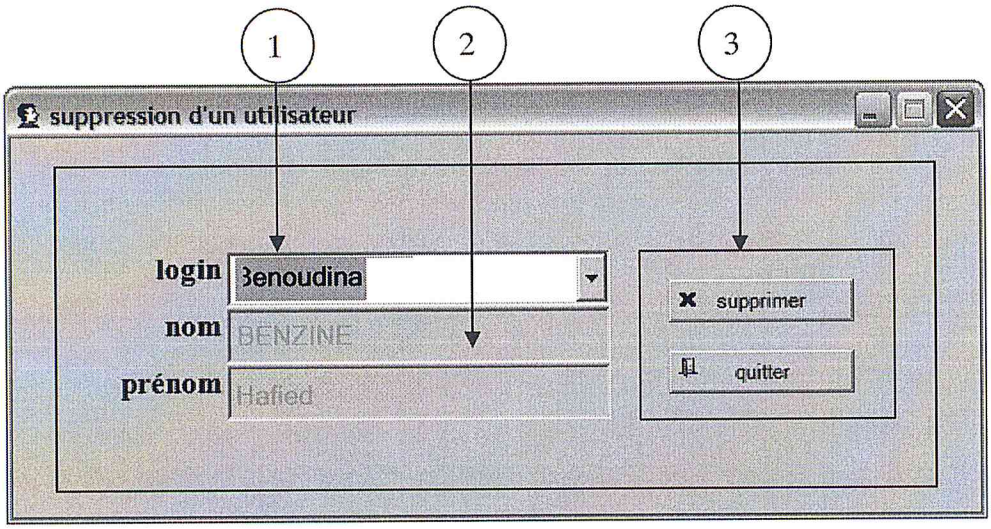


Figure V.3 Suppression d'un utilisateur

1) Liste de sélection permettant le choix de l'utilisateur.
2) Affichage des informations concernant l'utilisateur.
3) Supprimer.

➢ **Modification d'un compte utilisateur**

✓ Ajouter l'utilisateur à des groupes.

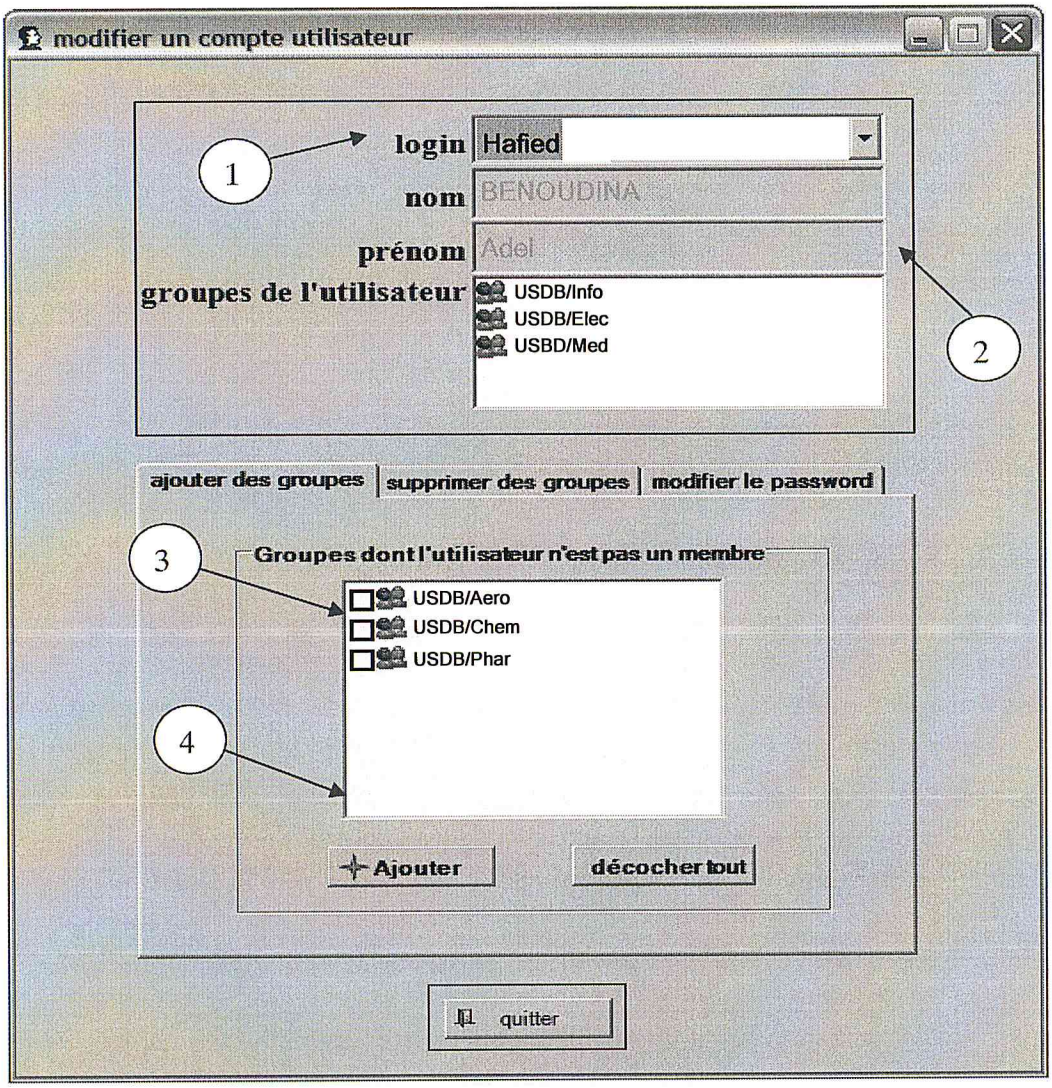


Figure V.4 Ajouter l'utilisateur à des groupes

1) Liste de sélection permettant le choix de l'utilisateur.
2) Affichage des informations qui concernent l'utilisateur
3) Cases à cocher permettant de sélectionner les groupes.
4) Ajouter groupe(s).

✓ Supprimer l'utilisateur du(des) groupe(s) où il est membre.

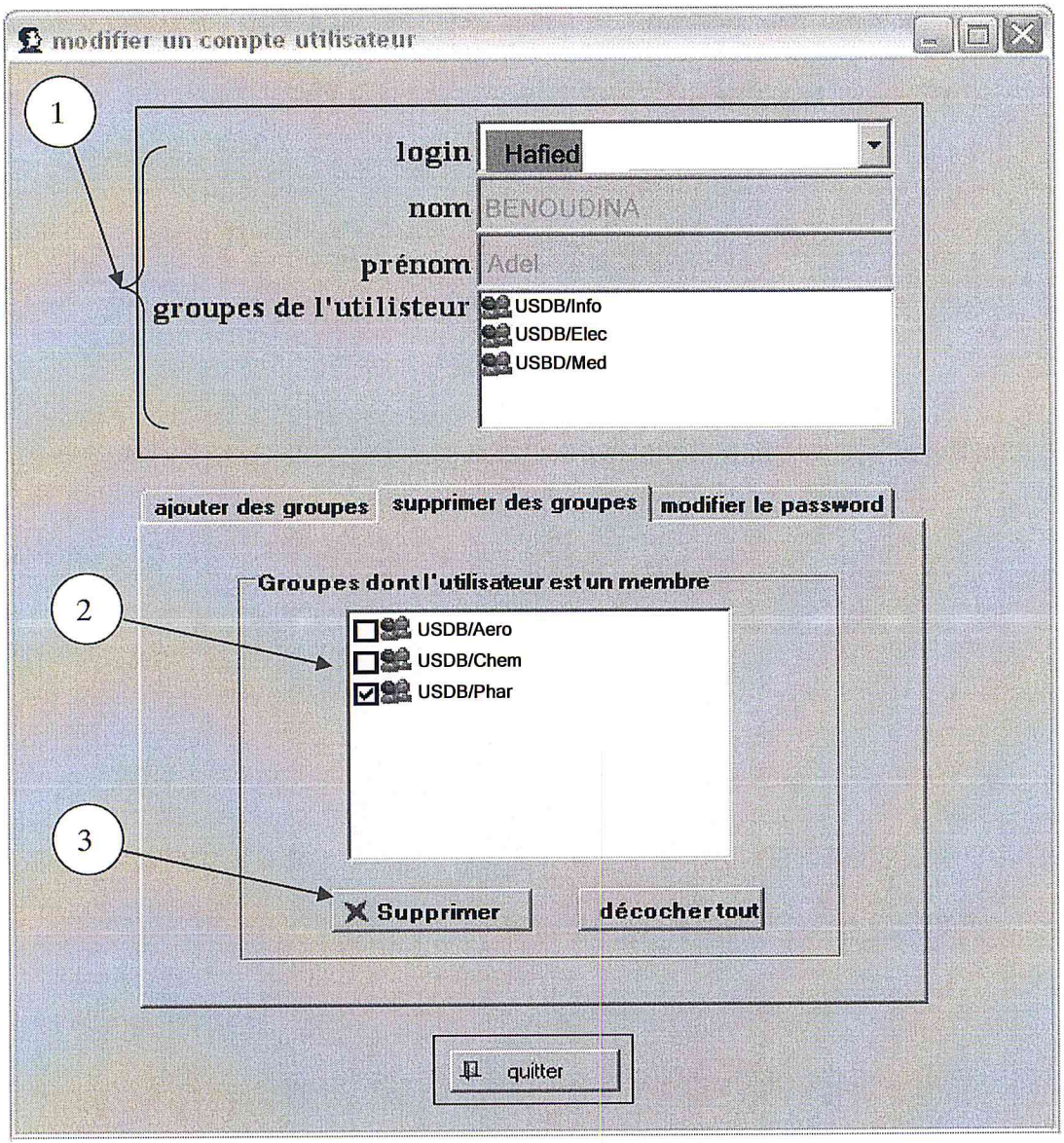


Figure V.5 Suppression des groupes pour un utilisateur

1) Cases à cocher qui permettent de sélectionner les groupes.
2) Liste des groupes dans lesquels l'utilisateur est membre avec des cases à cocher qui permettent la sélection des groupes à partir desquels l'utilisateur sera retiré.
3) Supprimer groupe(s).

V.2.1.2 Gestion des groupes

✓ Ajouter un groupe.

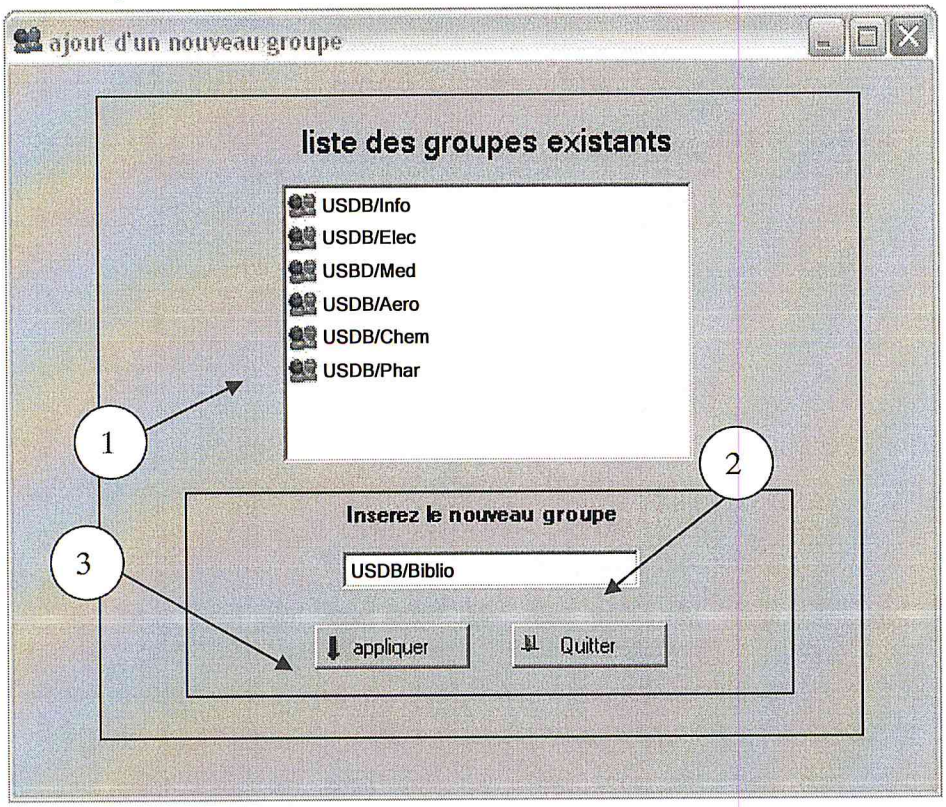


Figure V.6 Ajout d'un nouveau groupe

1) Liste des groupes existants.
2) Champ pour la saisie du nouveau groupe.
3) Ajouter groupe.

✓ Suppression d'un groupe.

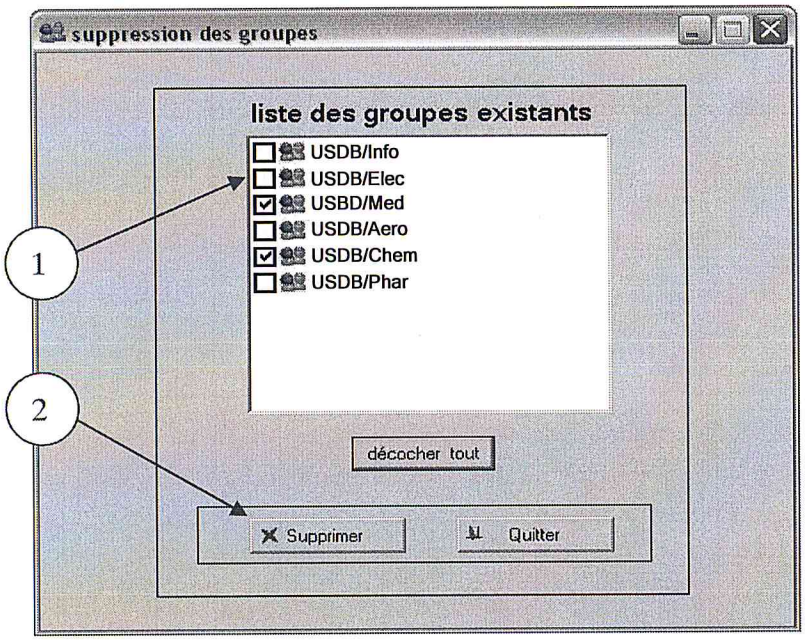


Figure V.7 Suppression des groupes

1) Cases à cocher qui permettent la sélection des groupes.
2) Supprimer groupe(s).

✓ Suppression des membres d'un groupe.

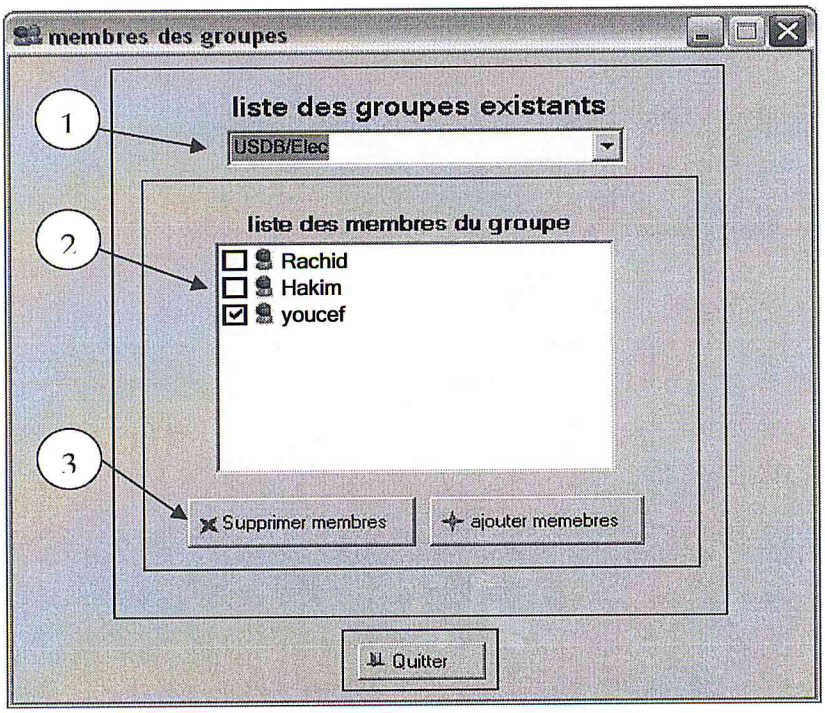


Figure V.8 Suppression des membres d'un groupe

1) Liste permettant la sélection du groupe.
2) Cases à cocher permettant la sélection des utilisateurs.
3) supprimer.
4) Ajouter des membres au groupe sélectionné.

- ✓ Ajout des membres a un groupe

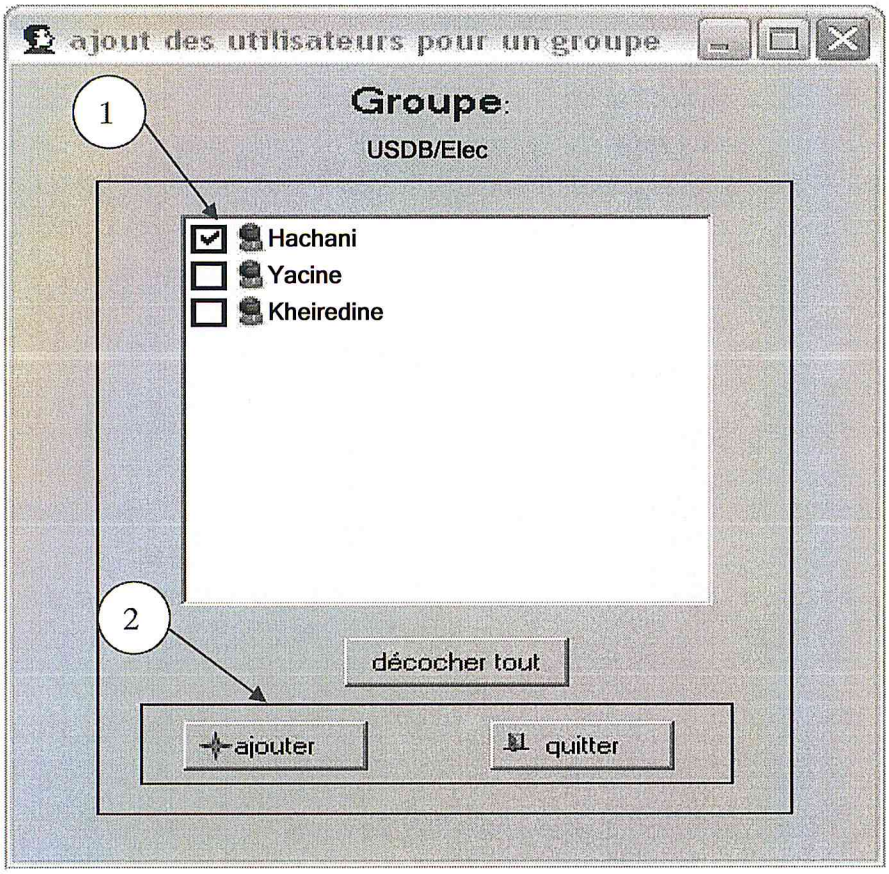


Figure V.9 Ajout des membres à un groupe

1) Cases à cocher qui permettent la sélection des utilisateurs.
2) Ajouter.

V.2.1.3 Gestion des documents

En ce qui concerne les documents publiés par les utilisateurs, l'administrateur ne peut faire que la consultation et la suppression de ces documents s'il juge que c'est nécessaire.

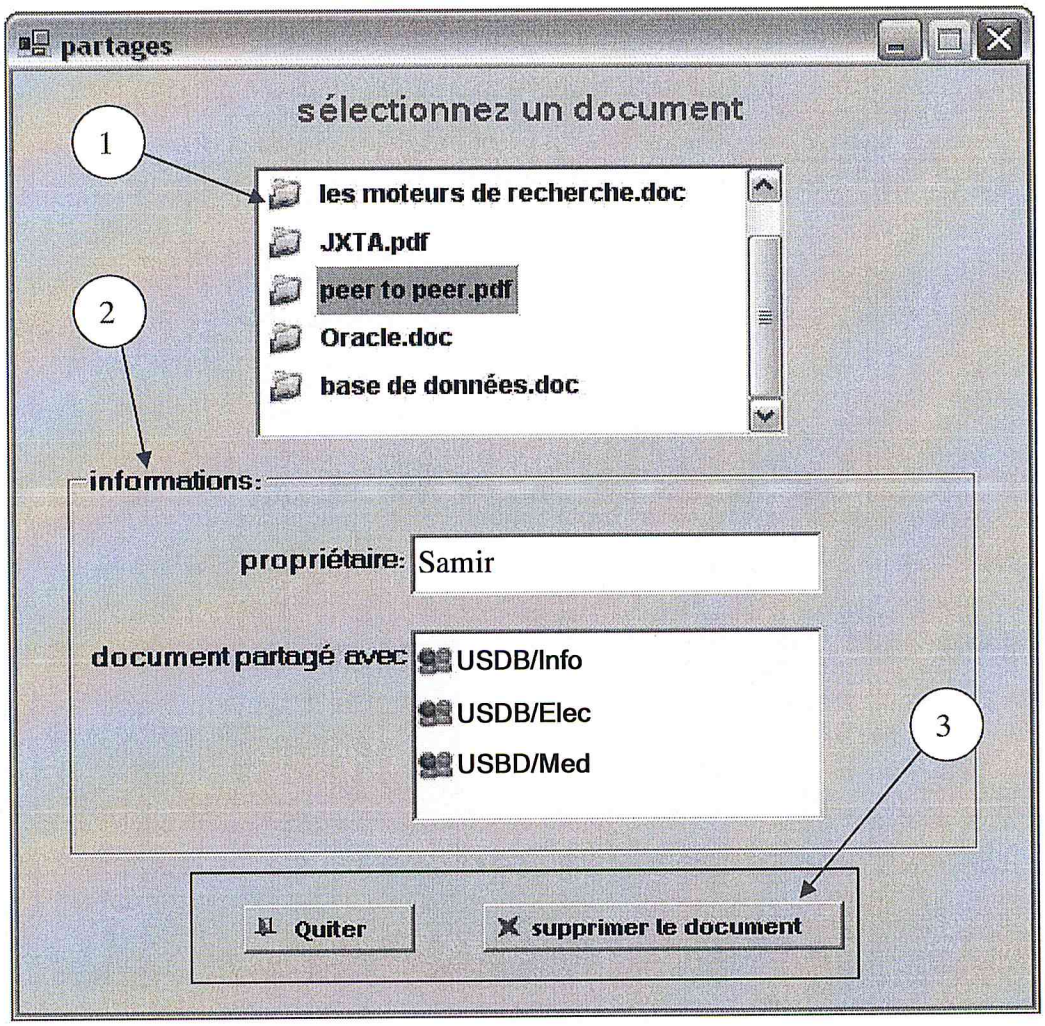


Figure V.10 Gestion des documents publiés

1) Liste permettant la sélection d'un document.
2) Informations sur le document sélectionné.
3) Supprimer le document.

V.2.2 Le Robot (Spider)

Etant donnée que le sujet de la réalisation des moteurs de recherche a été profondément traité et a fait l'objet de plusieurs études dans les laboratoires de recherches, une multitude de modèles ont vu le jour pour l'élaboration de tels outils. Chacune de ces méthodes propose une démarche différente pour aboutir à des résultats satisfaisantes qui garantissent une fiabilité du système et une performance élevée dans la qualité de l'indexation et de la recherche. Cependant, elles ont toutes suivi la même optique qui consiste à partager le système de recherche en deux grandes parties à savoir le Robot qui se charge de la localisation des documents et leur indexation, et l'interface de recherche qui utilise les résultats du Robot afin de satisfaire la requête d'un utilisateur.

D'après ce qui a été présenté lors de l'étude détaillée des moteurs de recherche dans les chapitres I, II, III. La partie la plus importante pour le fonctionnement d'un moteur de recherche est le Robot. Il représente la phase de la collection et d'indexation des documents qui se trouvent au sein du réseau. De ce fait, Une bonne réalisation de cette partie garantit la performance voulue de tout le système. Lors de l'implémentation de Robot pour le système Uni_Blida SEARCH, nous avons tenu compte du contexte dans lequel notre moteur de recherche va évoluer : un réseau Intranet.

V.2.2.1 Collection et indexation des documents

Les réseaux locaux se caractérisent par rapport aux réseaux étendus (WAN) par la limitation de la plage d'adresse IP des machines qui les constituent. En plus, l'information sur les lieux du stockage des documents peut être facilement obtenue. Alors, pour la mise en œuvre du Robot, l'indexation peut se faire de deux manières différentes. La première est automatique en injectant au système d'indexation une plage d'adresses qu'il doit scruter pour localiser les documents et les indexer. La deuxième est que l'administrateur, en cas de besoin, introduit une liste d'URL des sites et des documents pour que le système les indexe dans sa base de données.

En se basant sur les diverses possibilités et fonctionnalités offertes par le FrameWork **.NET**, la localisation des serveurs et le téléchargement des pages qu'ils contiennent se fait par des méthodes intégrées dans les classes Sockets et Webclient, alors que l'indexation se base sur des méthodes statistiques de traitement de texte. L'accès à la base de données Oracle est réalisé à l'aide des SDK développés pour l'architecture .NET

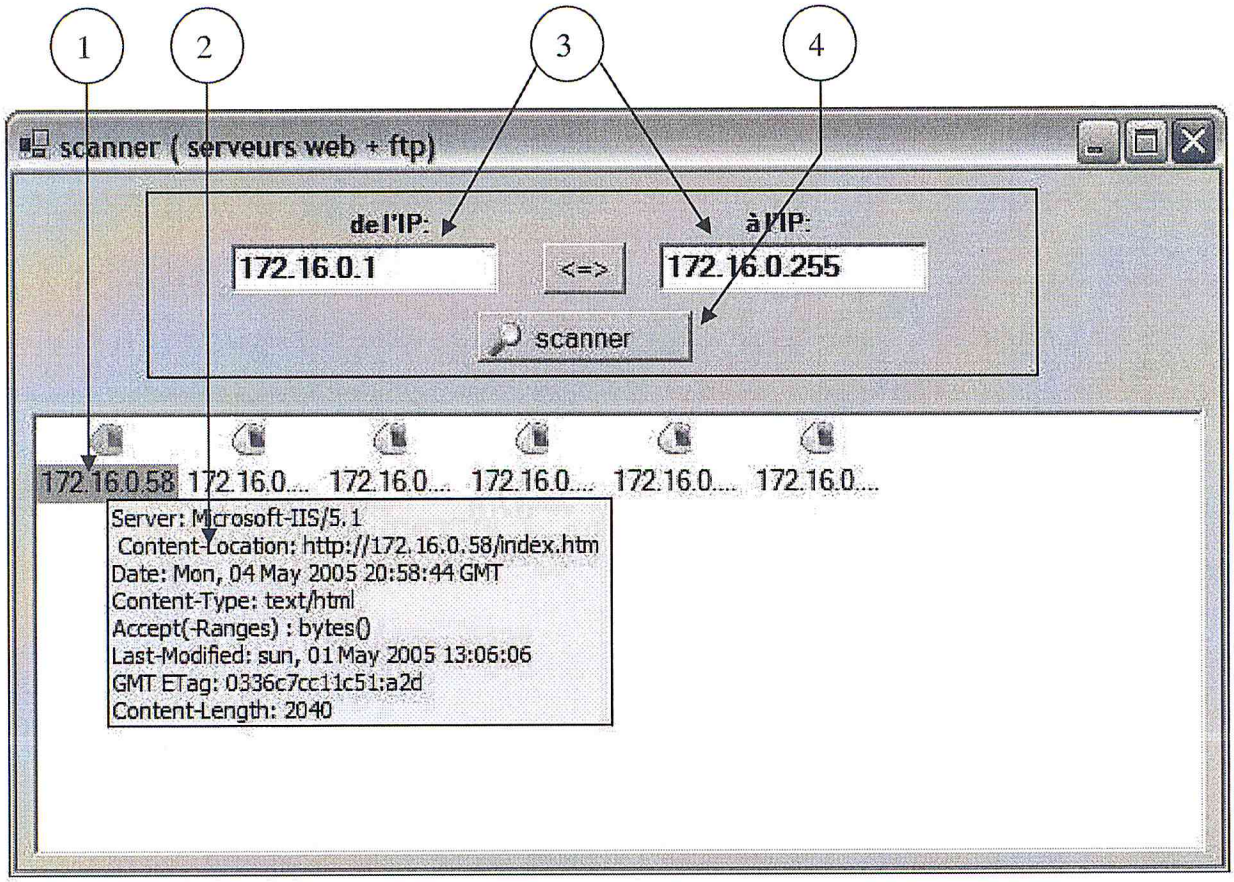


Figure V.11 Scanner de serveurs

1) Liste des serveurs trouvés.
2) Informations sur le serveur sélectionné.
3) La plage d'adresse à scanner.
4) Lancer le scan

V.2.3 Le site

V.2.3.1 Présentation générale

Cette partie de notre système représente les différentes interfaces par lesquelles passe le client pour effectuer les opérations :

- De recherche de documents ou d'informations.
- De publication de documents.
- De gestion des publications.

Parmi les différentes techniques de recherches citées dans le chapitre III nous avons retenus la recherche booléenne et la recherche textuelle. De ce fait, et lors de sa recherche, l'utilisateur peut utiliser les opérateurs suivants :

- Booléens :
  - Disjonction (OR).
  - Conjonction (AND).
  - Négation (NOT)
- De troncature ().
- De masque ().

Le système analyse la requête et génère une requête SQL finale suivant l'algorithme suivant :

DEBUT
Lire (Requête) ; // lecture de la requête tapée par l'utilisateur.
A← **ElimEsp** (Requête); // élimination des espaces (blancs) superflus de Requête
B ← **Minuscule** (A) ; // convertir la chaîne 'A' en minuscules dans la chaîne 'B'
vec ← **Extract_Mots** (B); //extraire les termes de la requête et les stocker dans un vecteur
req_fin ← '' ; // représente la requête SQL finale (initialement vide)
i ← 0 ; // l'indice de déplacement dans la chaîne 'B'
j ← 0 ; // l'indice de déplacement dans le vecteur 'vec'
TQ (i < **Taille** (B)) // parcourir toute la chaîne B
Si (B [i] = ' (' **ou** B [i] = ') ') **Alors** // ajouter le caractère courant à 'req_fin'
req_fin ← req_fin + B [i] ;
i ← i + 1 ;
Sinon
Switch (B [i])

```
                Case '+' : // il s'agit d'un opérateur ET ( l'intersection )
                        req_fin ← req_fin + . INTERSECT . ;
                        i ← i + 1 ;
                Case '-' : // un opérateur SAUF ( la différence )
                        req_fin ← req_fin + . EXCEPT . ;
                        i ← i + 1 ;
                Case ' ' : // un opérateur OU ( l'union )
                        req_fin ← req_fin + . UNION . ;
                        i ← i + 1 ;
                Default : // la rencontre d'un terme
                        tc ← vec [ j ] ; // lecture du terme courant à partir du vecteur
                        // 'vec'et le stocker dans la chaîne 'tc'
                        Si ( Exist-Tronc (tc1) ) Alors// 'tc1' contient un symbole de
                        // troncature (un caractère '*') ou de masque (le '?')
                        tc2 ← RemplaceTr ( tc1 ) ; // remplacer le caractère '*' par '%'
                        // et le caractère '?' par un '_'
                        req_fin ← req_fin + . SELECT d FROM doc_mot
                                WHERE mot LIKE tc2 . ;
                        Fsi
                        i ← i + Taille ( tc ) ;
        FinSwitch
    Fsi
FTQ
Résultats←Executer-SQL (req-fin ) ; // Exécution de la requête SQL et récupération des
//résultats.
FIN
```

Après la récupération des résultats, le système effectue un tri et un classement des documents selon un ordre de pertinence afin d'orienter l'utilisateur dans le choix des informations qu'il désire avoir sans être obligé de parcourir toutes les réponses.

La dernière étape à effectuer par notre système est la consultation du document choisi dans la liste des liens hypertextes par l'utilisateur. C'est par l'activation de l'ancre que le document est restitué. Ce dernier sera affiché par le navigateur web.

V.2.3.2 Présentation détaillée

❖ Page RECHERCHE PUBLIQUE

Cette page permet de faire des recherches publiques ou de s'authentifier pour faire des recherches privées.

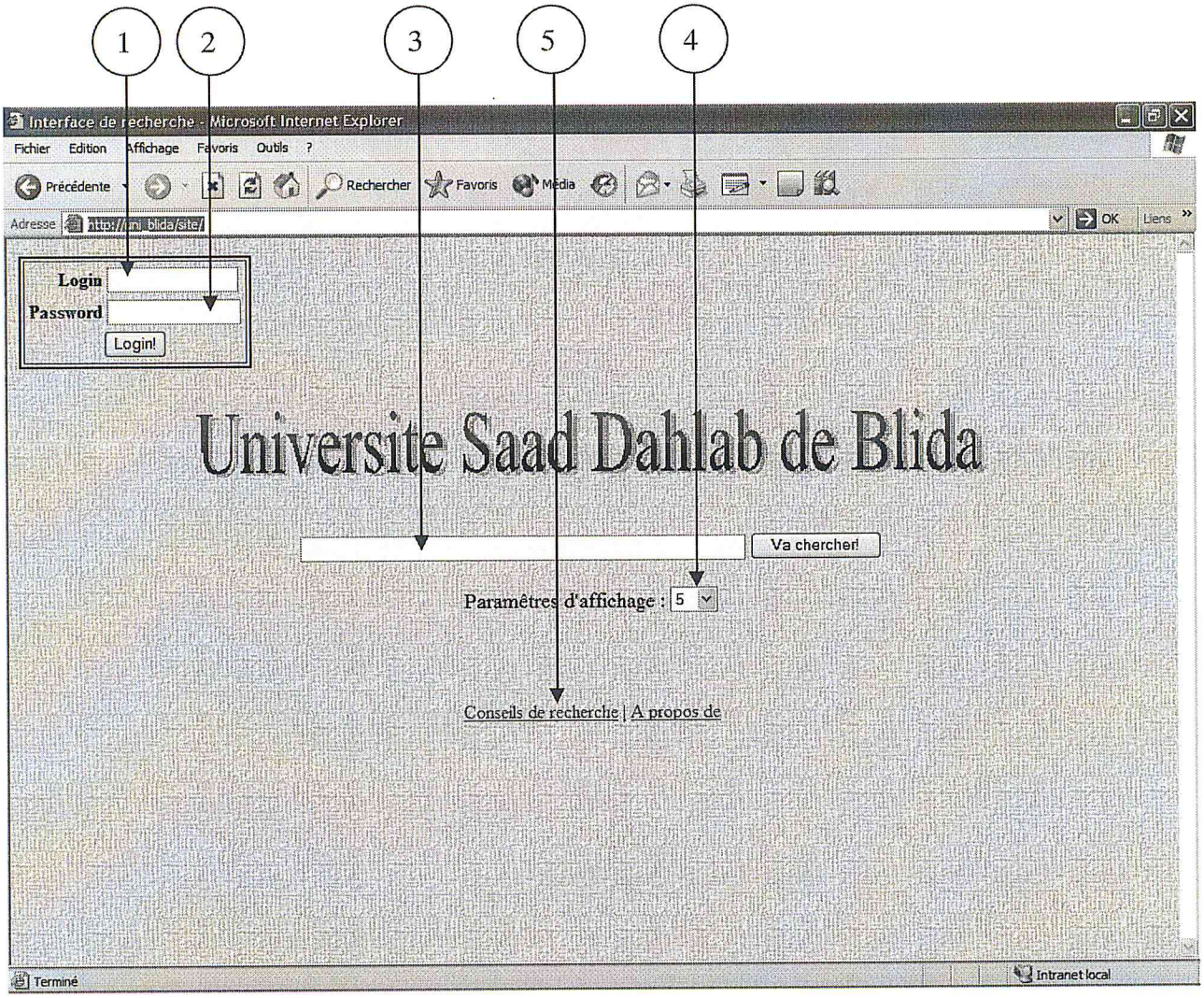


Figure V.12 Page «Recherche publique»

1) Login (de l'utilisateur).
2) Mot de passe.
3) Requête.
4) Nombre de documents à afficher.
5) Lien d'accès aux conseils pour la recherche.

❖ PAGE RECHERCHE PRIVEE

Cette page permet d'effectuer des recherches privées, de changer les paramètres personnels de l'utilisateur ou d'accéder au service de publication.

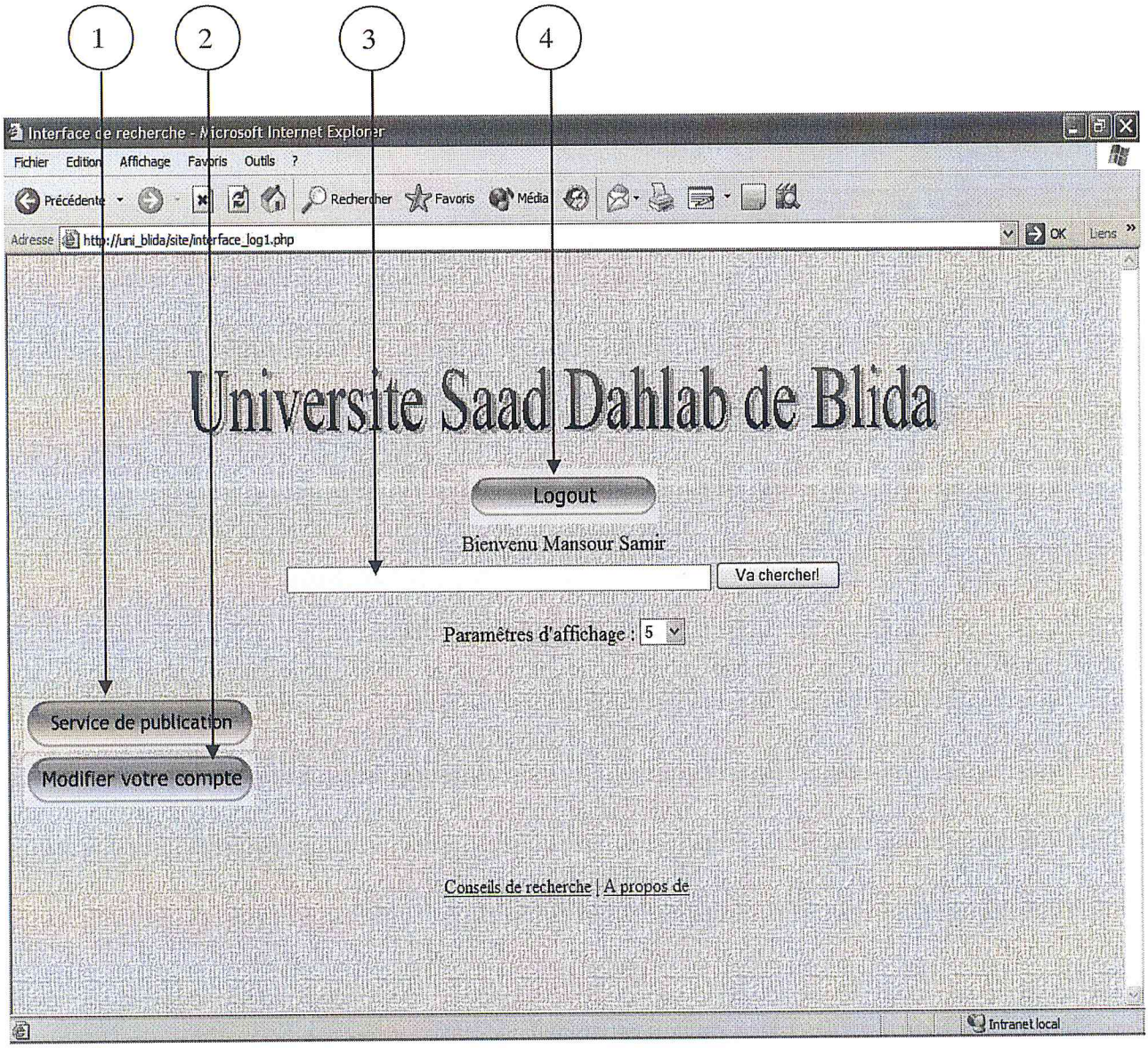


Figure V.13 Page «Recherche»

1) Lien qui permet d'accéder au service de publication documentaire.
2) Lien qui permet d'accéder à la page de modification du compte.
3) Requête.
4) Lien qui permet de se déconnecter.

Les résultats de la recherche seront affichés comme montré dans la figure ci-dessous:

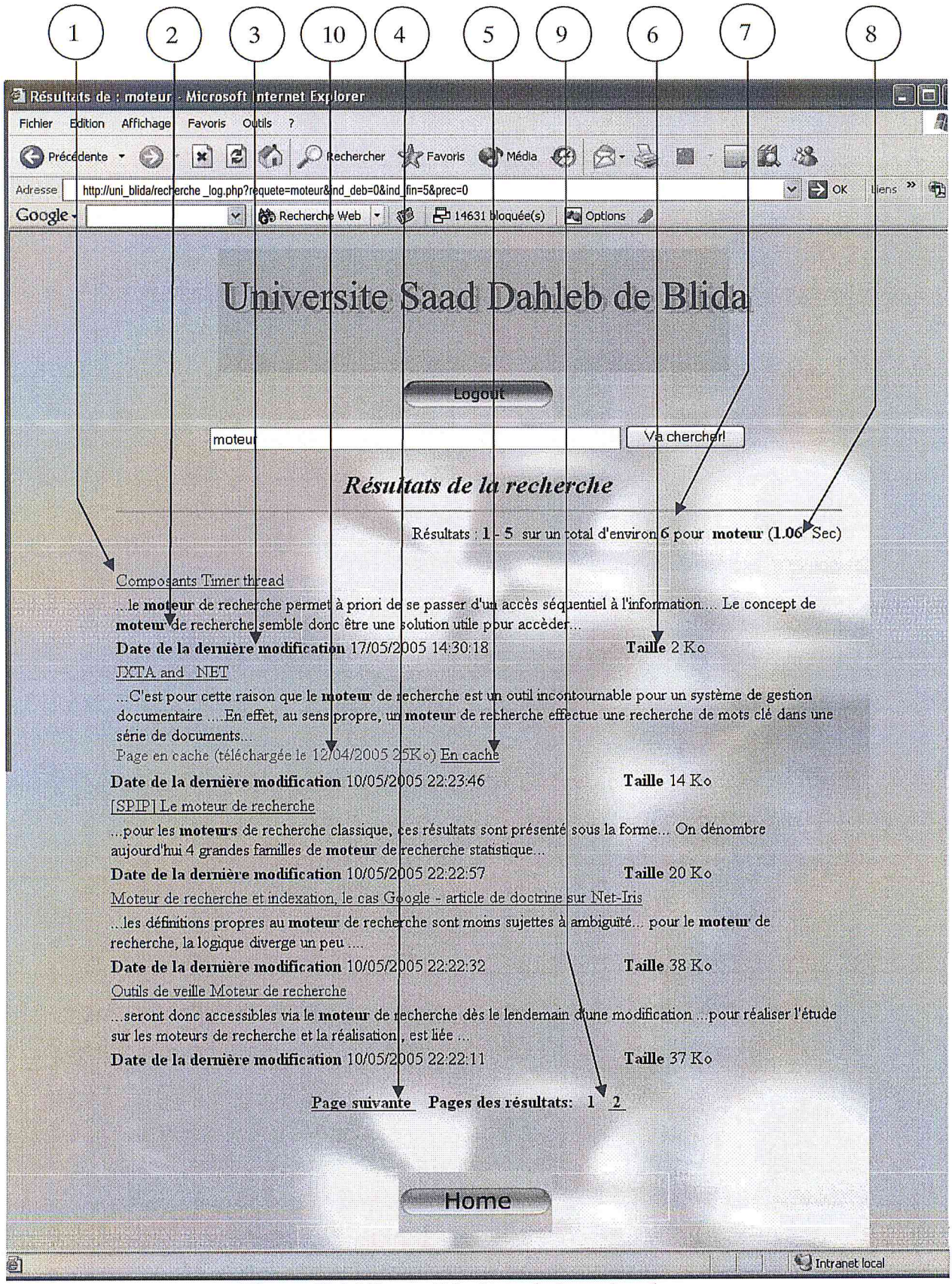


Figure V.14 Page «Résultats de la recherche»

1) Lien hypertexte vers le document résultat.
2) Extrait du document en mettant en évidence les mots clés trouvés.
3) Date de la dernière modification du document.
4) Lien vers la page suivante des résultats.
5) Lien vers le même document stocké au niveau du serveur cache.
6) La taille (en Ko) du document.
7) Le nombre total de documents trouvés.
8) La durée de la recherche.
9) Lien vers une page particulière parmi les pages des résultats.
10) Date (de stockage) et taille du document stocké dans le serveur cache.

❖ **Page SERVICE DE PUBLICATION**

Elle permet d'accéder aux différents services proposés : publication documentaire automatique et manuelle.

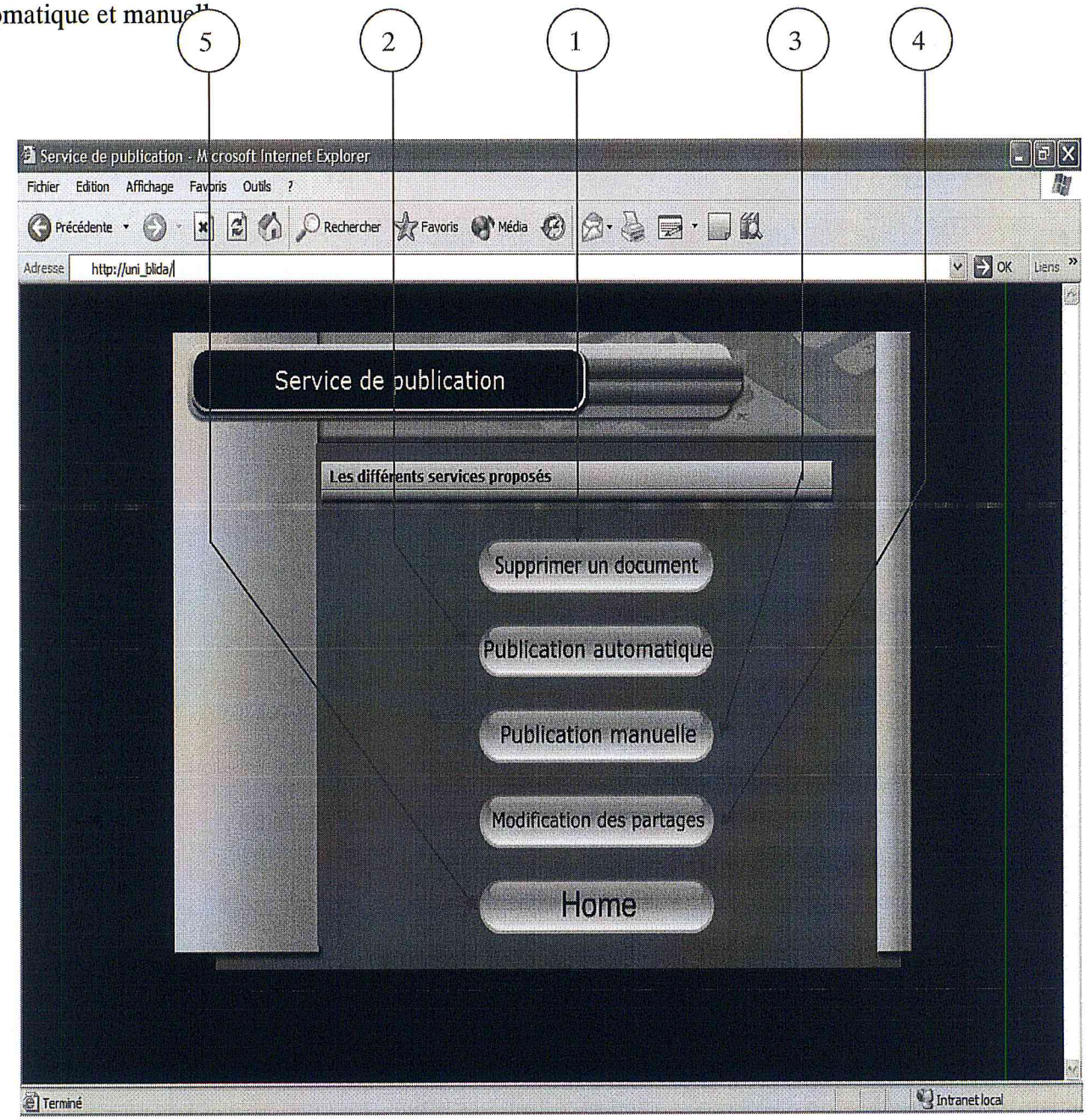


Figure V.15 Page «Service de publication»

1) Lien d'accès à la page de suppression des documents.
2) Lien d'accès à la page de publication automatique des documents.
3) Lien d'accès à la page de publication manuelle des documents.
4) Lien d'accès à la page de modification des partages.
5) Lien d'accès à la page de recherche.

❖ **Page PUBLICATION AUTOMATIQUE**

Cette page permet à l'utilisateur de publier des documents de manière automatique.

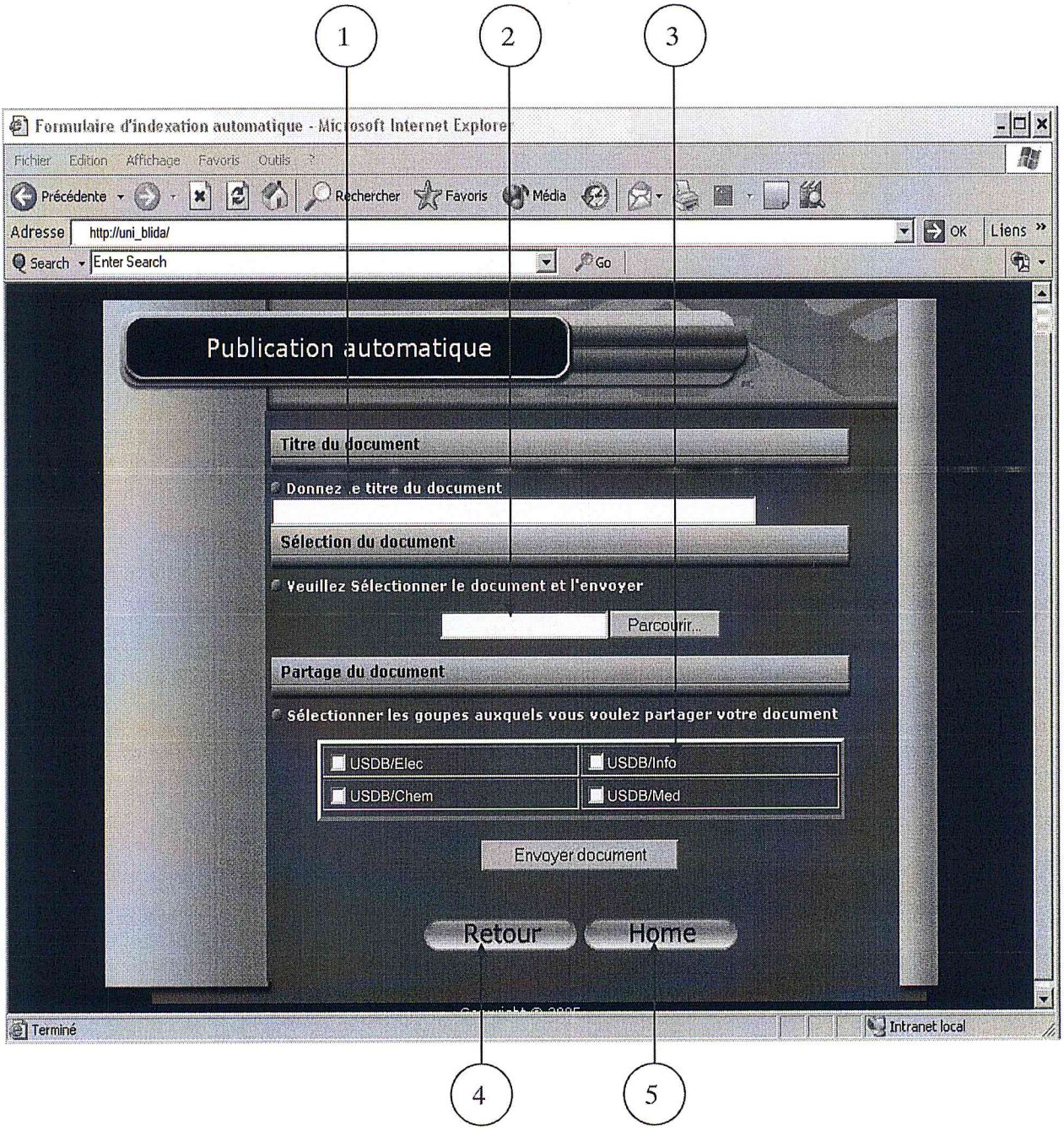


Figure V.16 Page «Publication automatique»

1) Titre du document à publier.
2) Champs qui permet la sélection du document à publier.
3) Cases à cocher qui permettent la sélection des groupes pour lesquels l'utilisateur veut partager le document.
4) Lien d'accès à la page du service de publication.
5) Lien d'accès à la page de recherche.

❖ Page PUBLICATION MANUELLE

Elle permet de publier des documents manuellement (l'utilisateur doit fournir toutes les informations qui concernent le document à publier)

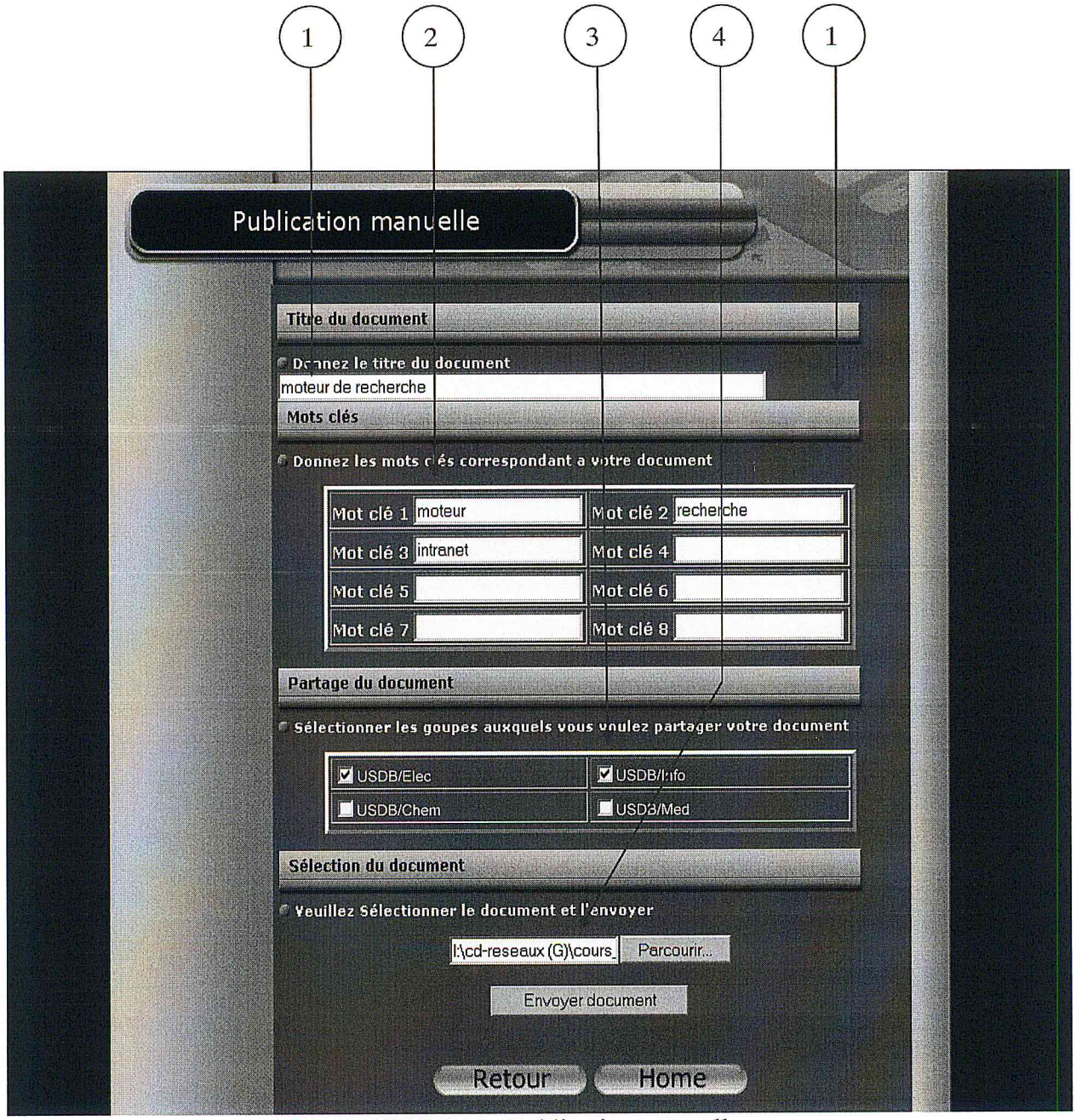


Figure V.17 Page «Publication manuelle»

1) Titre du document à publier.
2) Mots clés correspondants au document à publier.
3) Case à cocher qui permettent la sélection des partages.
4) Champs qui permet la sélection du document.

Si le document existe déjà (publié auparavant), la page suivante sera affichée :

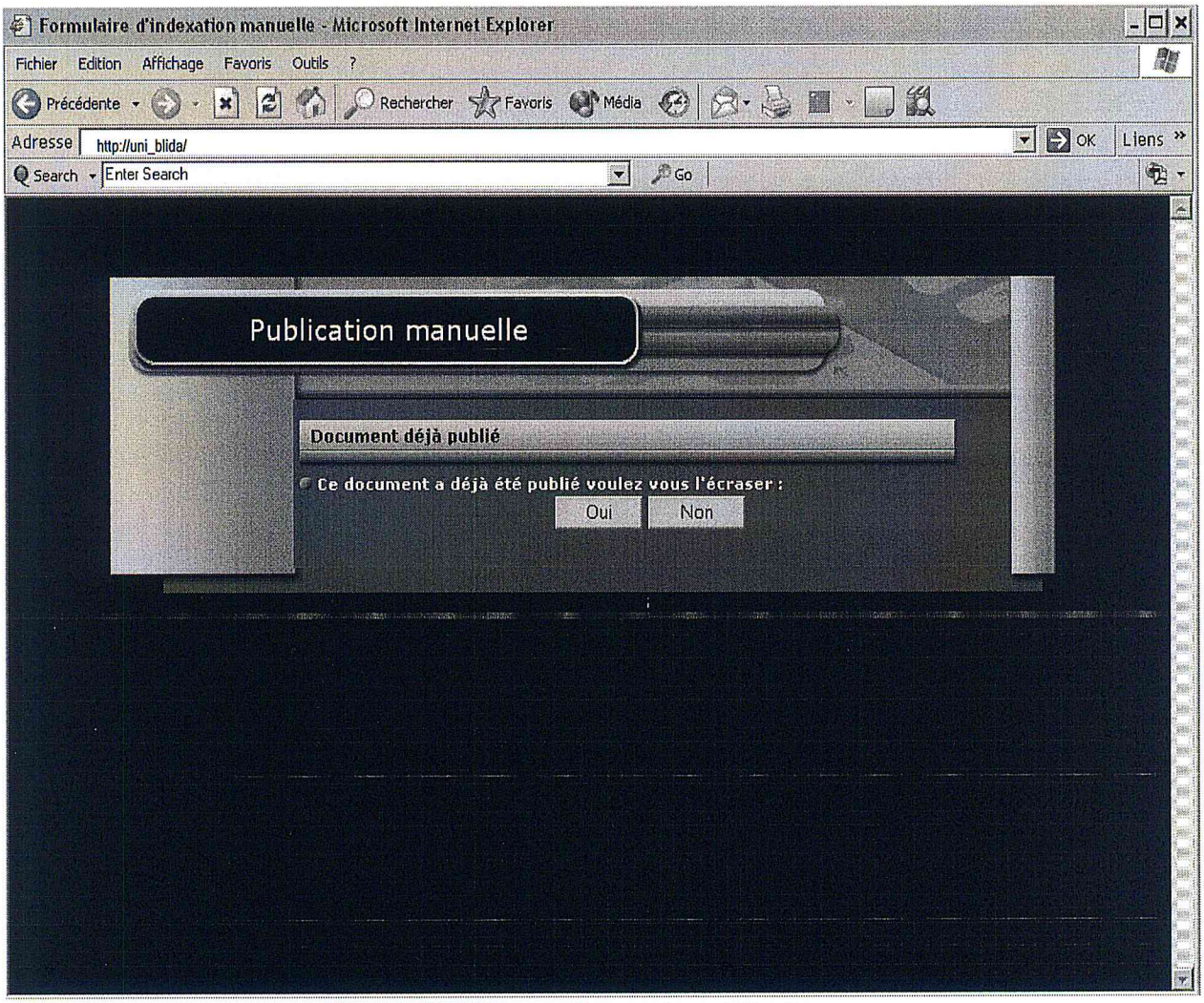


Figure V.18 Page «Erreur de publication»

Si la publication se termine avec succès, la page suivante sera affichée :

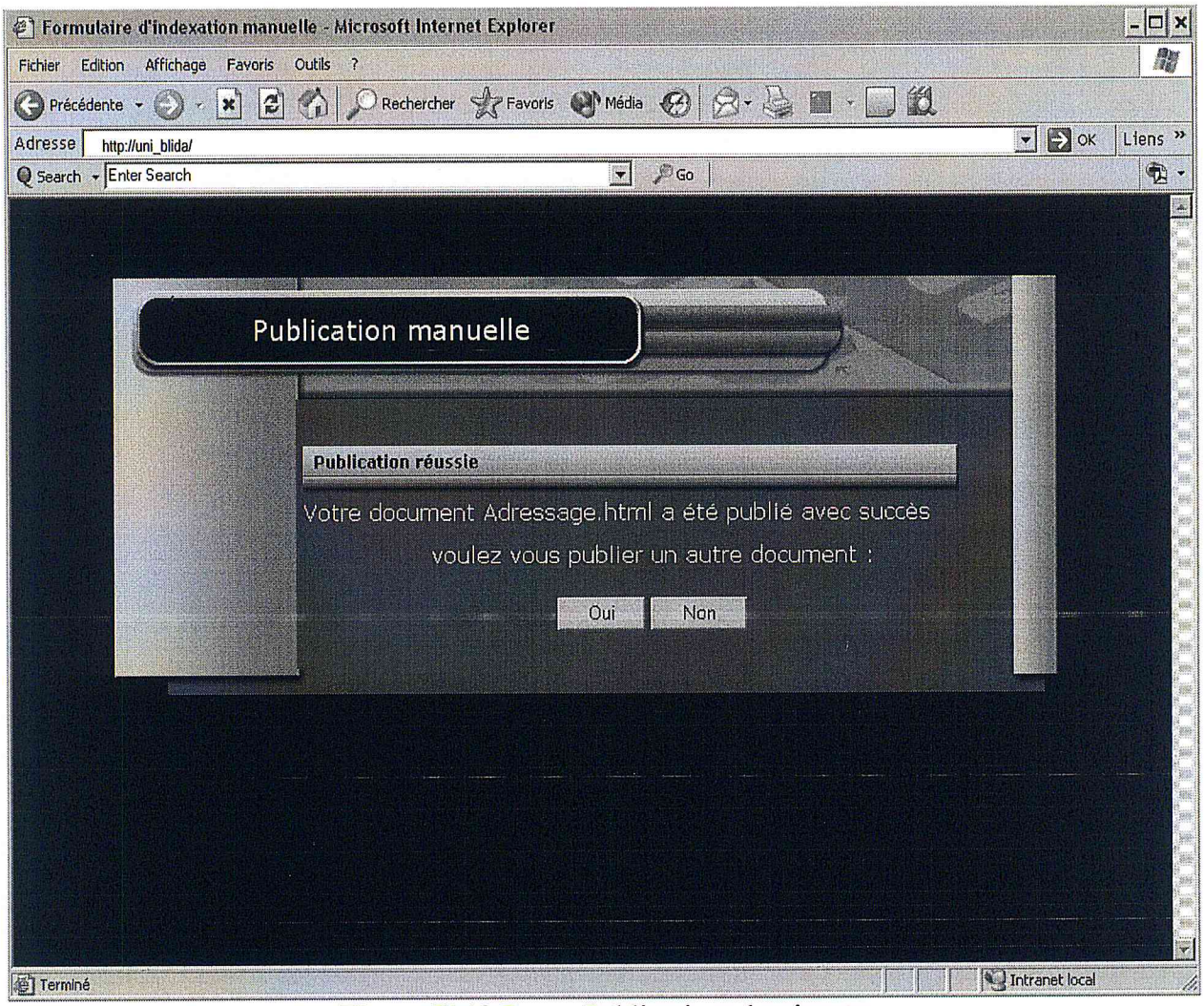


Figure V.19 Page «Publication réussie»

❖ Page SUPPRESSION DES DOCUMENTS PUBLIES

Elle permet à l'utilisateur la suppression de ses documents publiés.

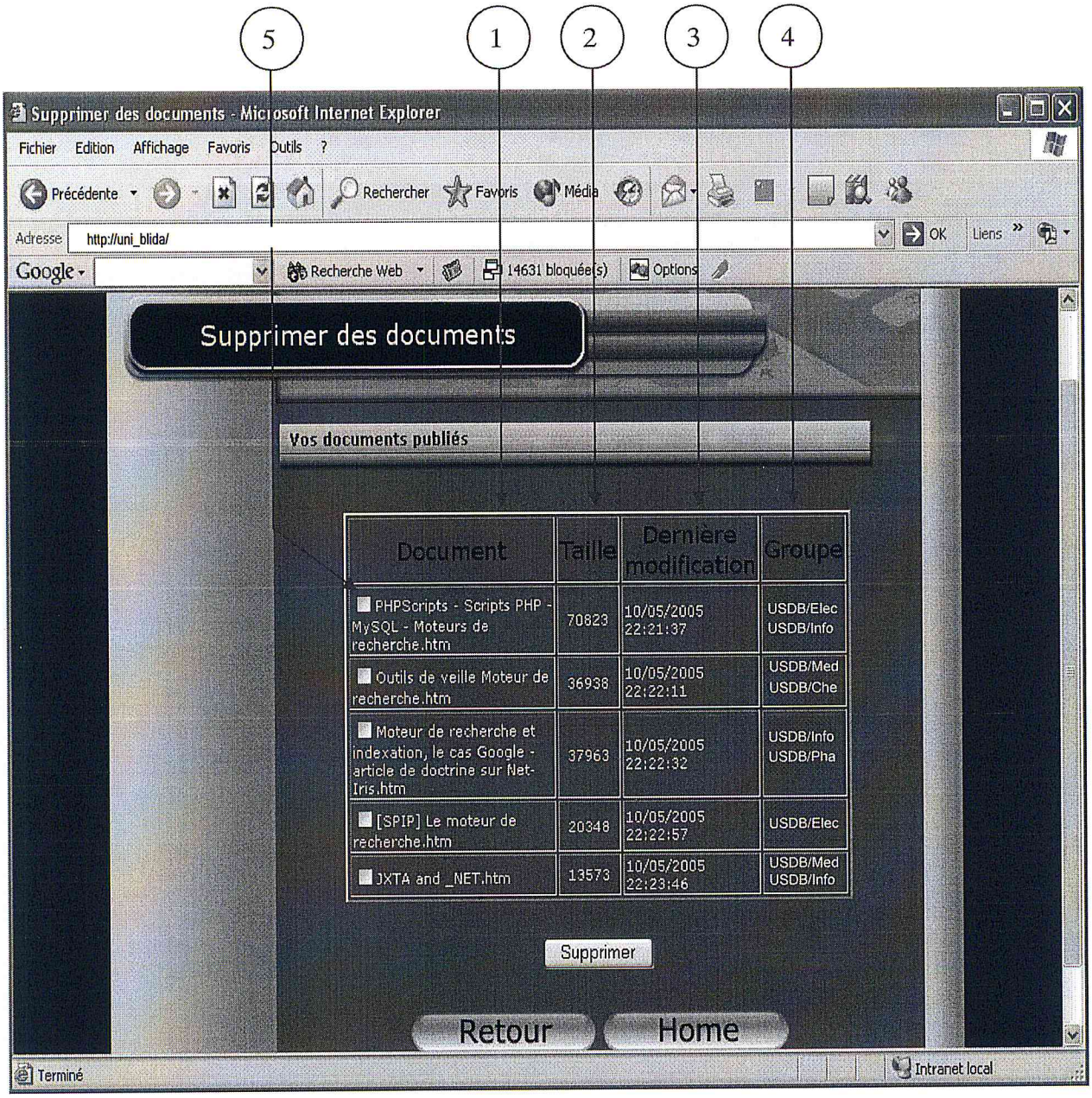


Figure V.20 Page «Suppression de documents»

1) Nom du document.
2) Taille en Octets du document.
3) La date de la dernière modification du document.
4) Groupes pour lesquels le document est partagé.
5) Case à cocher qui permet la sélection des documents à supprimer.

❖ Page MODIFICATION DES PARTAGES

Elle permet à l'utilisateur d'ajouter/supprimer des partages de ses documents publiés.

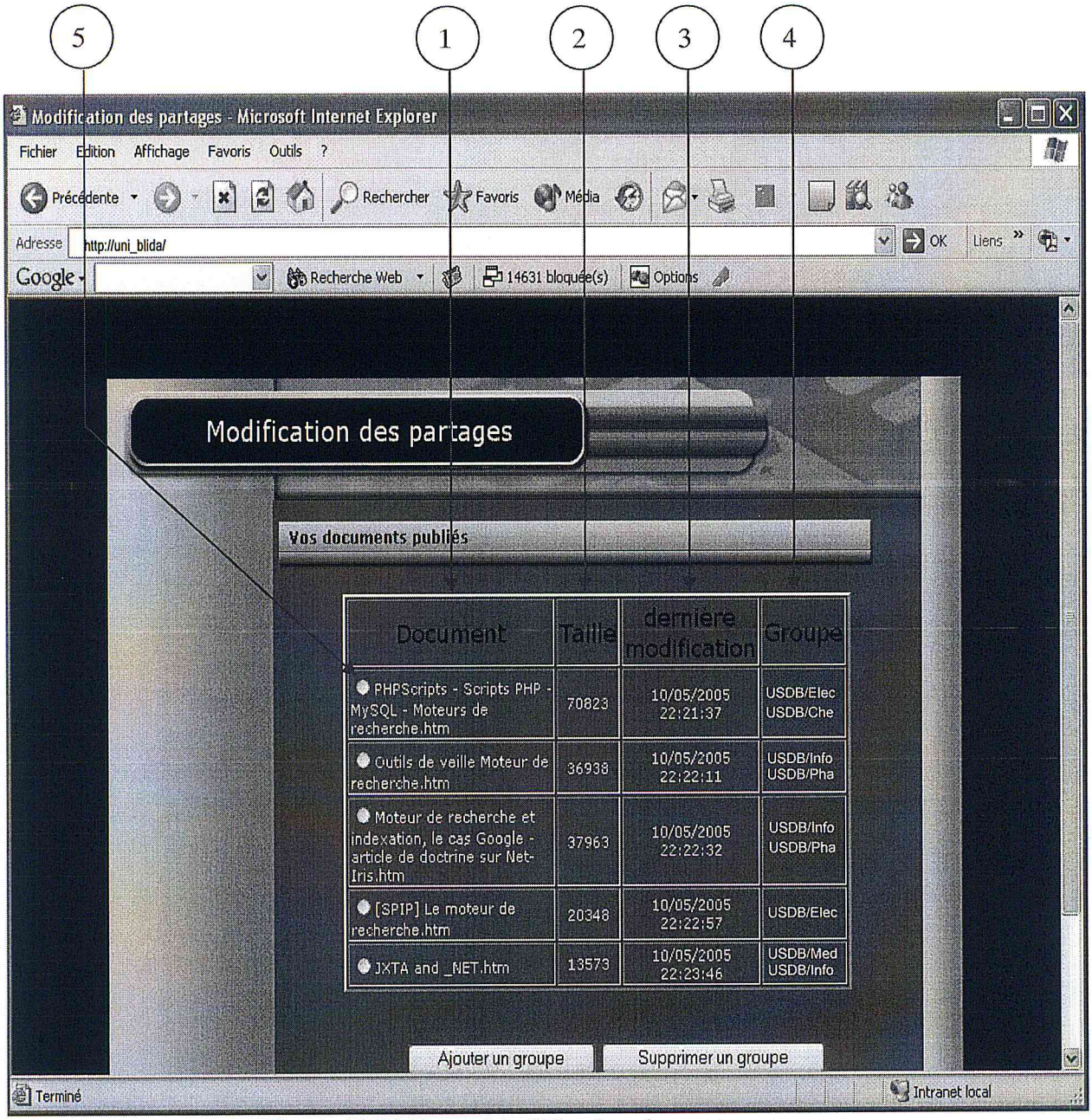


Figure V.21 Page «Modification des partages »

1) Nom du document.
2) Taille en Octets du document.
3) La date de la dernière modification du document.
4) Groupes pour lesquels le document est partagé.
5) Bouton radio qui permet la sélection du document à modifier.

L'ajout de partages se fait via la page suivante :

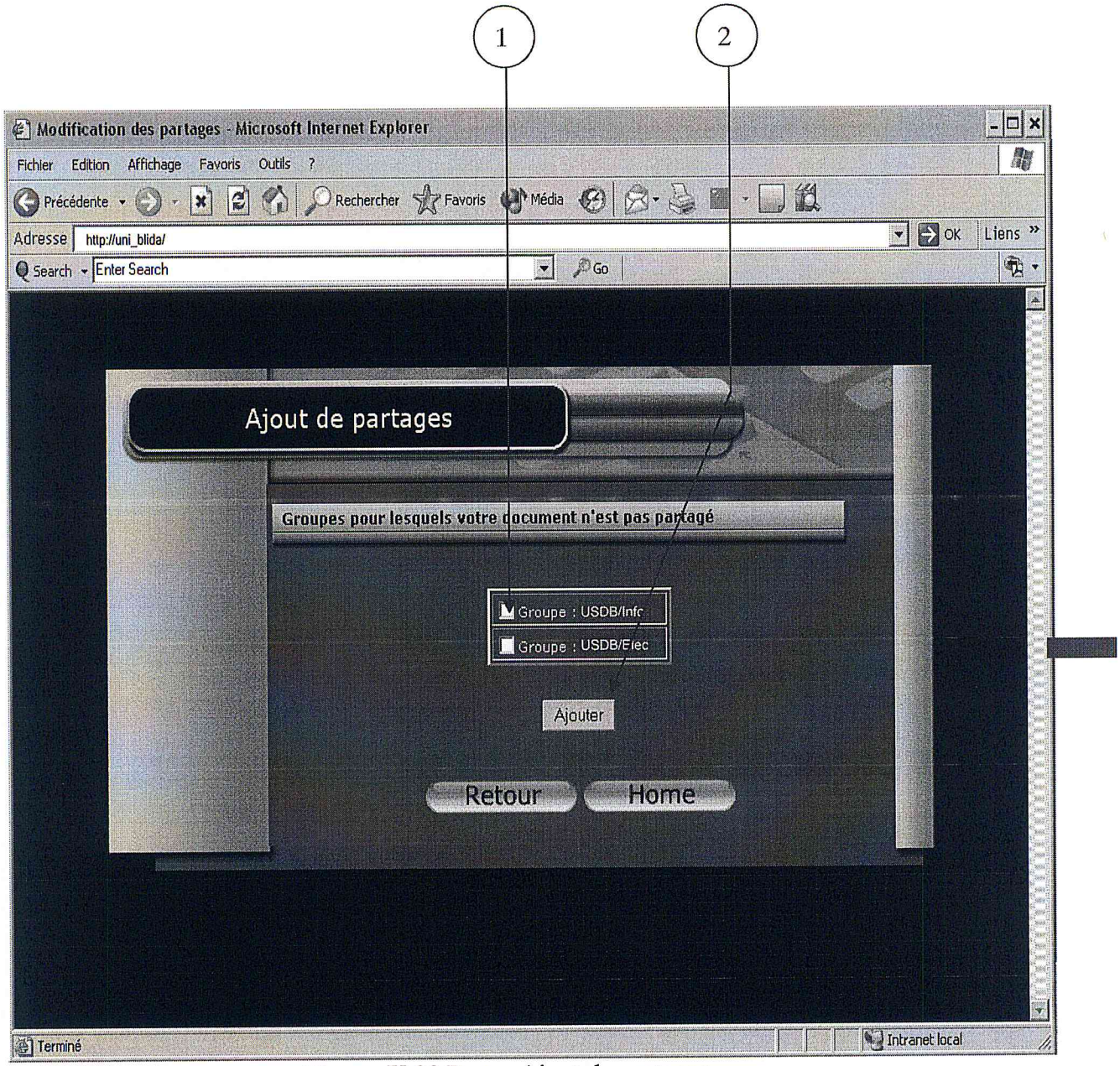


Figure V.22 Page «Ajout de partages »

1) Cases à cocher permettant la sélection des groupes à ajouter.
2) Ajouter groupe(s).

La suppression de partages se fait via la page suivante :

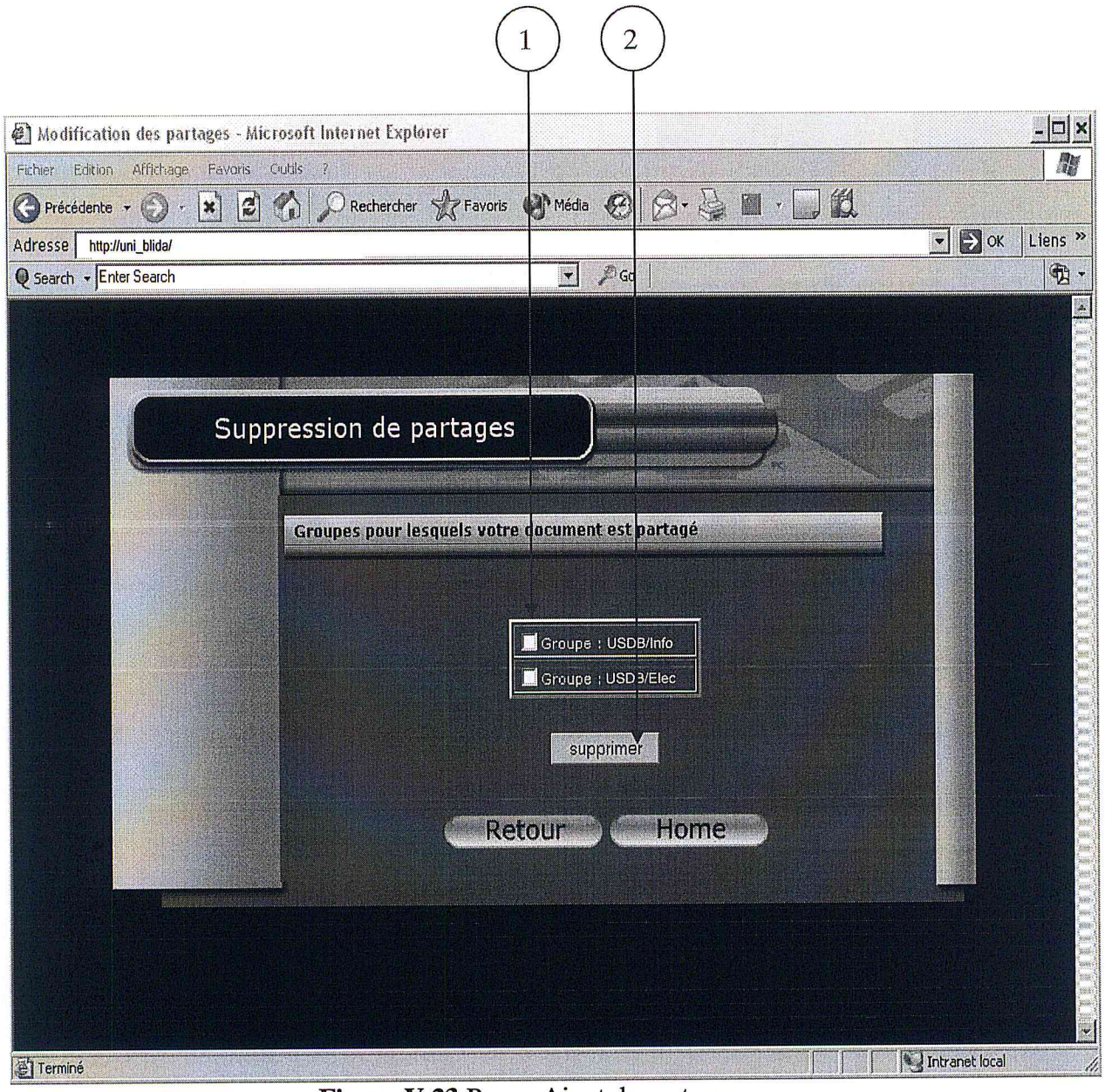


Figure V.23 Page «Ajout de partages »

1) Cases à cocher permettant la sélection des groupes à supprimer.
2) Supprimer groupe(s).

❖ **Page MODIFICATION DU COMPTE**

Elle permet à l'utilisateur de modifier son Login ou son mot de passe.

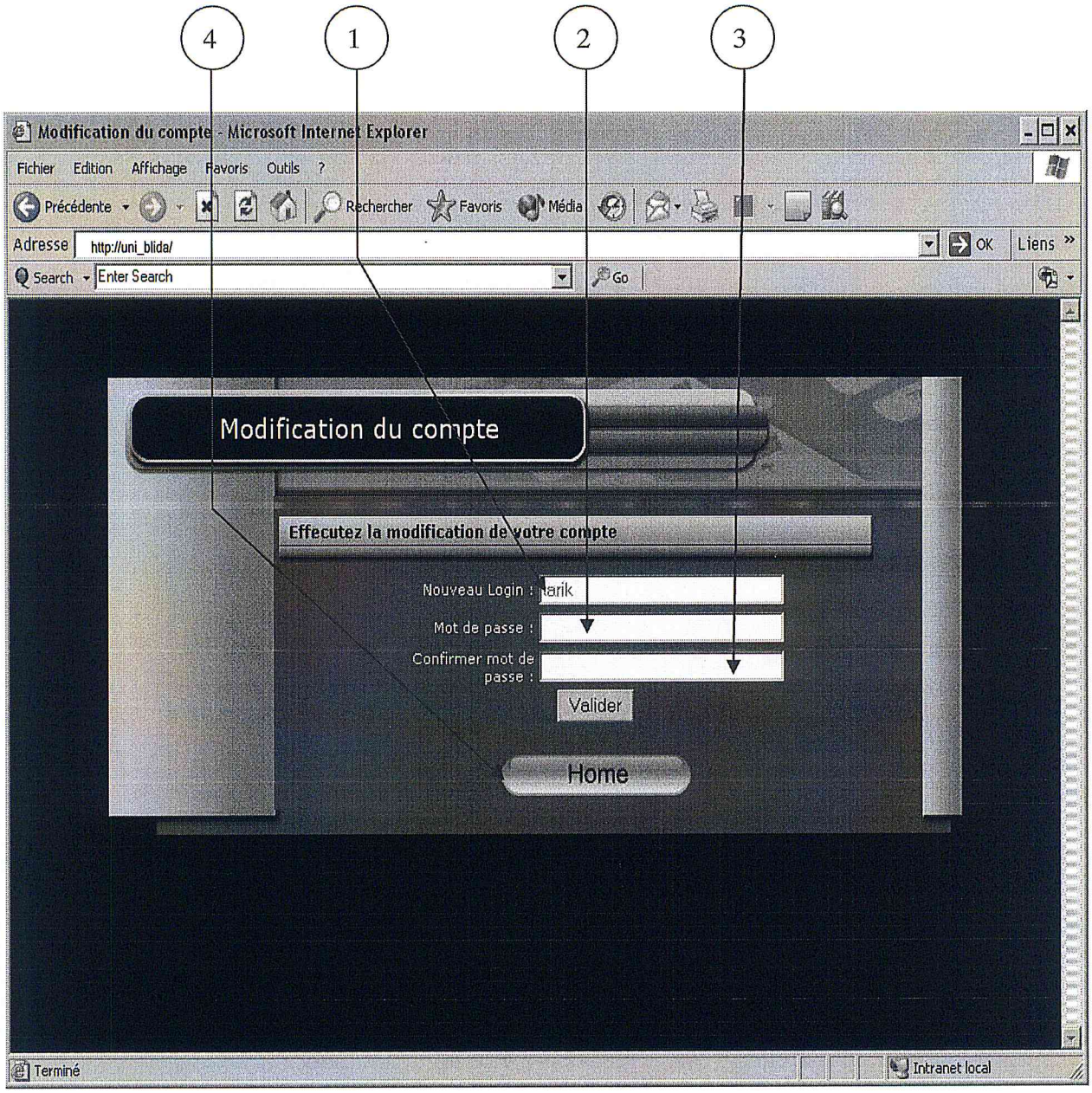


Figure V.24 Page «Modification du compte»

1) Zone d'introduction du nouveau Login.
2) Zone d'introduction du nouveau mot de passe.
3) Confirmation du nouveau mot de passe.
4) Lien d'accès à la page de recherche.

CONCLUSION ET PERSPECTIVES

Au cours de notre travail, nous nous sommes rendu compte que le domaine de la recherche documentaire reste un vaste univers d'exploration, par ses diverses théories et méthodes d'extraction de l'information des textes, de l'analyse des corpus pour aboutir à une meilleur indexation et dans le traitement des requêtes et la recherche des réponses adéquates.

Notre projet s'inscrit dans le cadre du développement d'un moteur de recherche pour l'indexation de documents : HTML, TXT, PDF et DOC.

Le moteur de recherche concernera :

- **Un spider** chargé de scruter le Web continuellement afin de référencer automatiquement des documents.
- **Une base de données** qui constitue un index incluant les informations concernant les documents à accès publique ainsi que celles des publications des utilisateurs.
- **Une interface d'interrogation** permettant d'exploiter les résultats retrouvés par le spider et incluant un service de publication de documents pour les utilisateurs.

Un des objectifs majeurs de notre travail est l'élaboration d'un mécanisme d'indexation fiable qui assure une meilleure performance de tout le système, car il constitue la matière grise sur laquelle se base la recherche des utilisateurs. Nous avons étudié les différentes méthodes d'indexation documentaire, et nous avons jugé que la méthode statistique était la plus adaptée à un environnement caractérisé par une circulation de données accrue. C'est une méthode simple, pas très gourmande en ressources mais qui assure de bons résultats.

Notre système Uni_Blida SEARCH est dédié à une utilisation restreinte et sécurisée dans un Intranet. Il est aussi administré pour pouvoir gérer les utilisateurs, l'index et les documents pour une meilleure utilisation et une adaptation à chaque champ d'application. Il tient compte de la normalisation des mots (lemmatisation), des traitements des requêtes complexes et de la sécurisation des documents privés.

Il est évident que le moteur de recherche développé ne pourra assurer une couverture intégrale du réseau comme c'est le cas pour les leaders dans le domaine de la recherche d'information sur le Web. Néanmoins, au cours de notre étude, nous avons essayé de résoudre

les problèmes liés à la recherche documentaire (du texte). Nous souhaitons élargir cette étude à la recherche multimédia touchant les images, la vidéo et le son.

Les perspectives d'évolutions du projet s'annoncent comme suit :

- La reconnaissance de plusieurs langues (réalisable par la méthode statistique).
- La sécurisation du transfert des fichiers personnels (établissement de politiques de sécurité).
- L'indexation des images, de la vidéo et du son.
- Une lemmatisation plus approfondie avec des traitements sémantiques poussés.
- L'indexation de plusieurs formats de documents (.PPT, .XLS, ...)
- L'introduction des opérateurs d'adjacence et de proximité ainsi que la recherche des phrases.
- Orienter le moteur de recherche vers la recherche sémantique comme phase initiale dans le processus d'indexation.

ANNEXES

Annexe A

Le Framework .NET

Qu'est ce que le Framework .NET ?

Le Framework .NET est une nouvelle plate-forme informatique qui simplifie le développement d'applications dans l'environnement fortement distribué d'Internet. Le Framework .NET est conçu pour remplir les objectifs suivants :

- Fournir un environnement cohérent de programmation orientée objet que le code objet soit stocké et exécuté localement, exécuté localement mais distribué sur Internet ou exécuté à distance.
- Fournir un environnement d'exécution de code qui minimise le déploiement de logiciels et de conflits de versionning.
- Fournir un environnement d'exécution de code qui garantit l'exécution sécurisée de code y compris le code créé par un tiers d'un niveau de confiance moyen ou un tiers inconnu.
- Fournir un environnement d'exécution de code qui élimine les problèmes de performance des environnements interprétés ou écrits en scripts.
- Fournir au développeur un environnement cohérent entre une grande variété de types d'applications comme les applications Windows et les applications Web.
- Générer toutes les communications à partir des normes d'industries pour s'assurer que le code basé sur le Framework .NET peut s'intégrer à n'importe quel autre code.

Le Framework .NET contient deux principaux composants : la Common Language Runtime et la bibliothèque de classes du Framework .NET.

Le Common Language Runtime est la base du Framework .NET. Le runtime peut être considéré comme un agent qui manage le code au moment de l'exécution, fournit des services essentiels comme la gestion de la mémoire, la gestion des threads, et l'accès distant. Il applique également une stricte sécurité des types et d'autres formes d'exactitude du code qui garantissent un code sécurisé et robuste. En fait, le concept de gestion de code est un principe fondamental du runtime. Le code qui cible le runtime porte le nom de code managé (par opposition au code non managé).

La bibliothèque de classes, l'autre composant principal du Framework .NET, est une collection complète orientée objet, de types réutilisables que vous pouvez utiliser pour développer des applications allant des traditionnelles applications à ligne de commande ou à interface graphique utilisateur (GUI, Graphical User Interface) jusqu'à des applications qui

exploitent les dernières innovations fournies par ASP.NET, comme les services Web XML et Web Forms.

Qu'est ce que la Common Language Runtime (CLR) ?

La Common Language Runtime (CLR) est un environnement d'exécution sécurisé et robuste qui supporte du code écrit dans plusieurs langages différents (C++, VB, C#, Pascal, Cobol ...) et simplifie le développement, la gestion et le déploiement d'applications. On peut la comparer à la Java Virtual Machine (JVM) ou au Runtime Visual Basic 6 (msvbvm60.dll).

La CLR est constituée d'un ensemble de services standards (Modèle de programmation orientée objet, sécurité, ramasse miettes) dont chaque programme .NET peut tirer profit

Qu'est-ce que le Common Type System (CTS) ?

Afin que des classes définies dans plusieurs langages puissent communiquer entres elles, elles ont besoin d'un ensemble de types de données communs. C'est l'objet de la CTS, elle définit les types de données que le Runtime .NET comprend et que les applications .NET peuvent utiliser. A noter que la CTS est un sur ensemble de la CLS.

Qu'est-ce que la Common Language System (CLS) ?

C'est un sous-ensemble de la CTS que chaque langage .NET est supposé supporter. Un programme qui utilise des types de la CLS peut interagir avec un autre programme .NET écrit dans un autre langage. Il est donc ainsi possible par exemple qu'une classe C# hérite d'une classe VB .NET.

Qu'est-ce que l'Intermediate Language (IL) ?

Tous les programmes .NET avant d'être déployés sont compilés dans un langage de bas niveau appelé Intermediate Language ou Microsoft Intermediate Language (MSIL) : ce code IL est ensuite compilé dans du code natif au moment de l'exécution. Ce qui signifie que quelque soit le langage utilisé dans votre programme, vos exécutables et DLL seront toujours déployés sous la forme de code IL ; il n'y a donc aucune différence entre un composant écrit en C# et en VB .NET.

Qu'est-ce qu'un Assembly ?

Les assemblies sont un élément fondamental de la programmation avec le Framework .NET. Un assembly exécute les fonctions suivantes :

- Il contient le code que le Common Language Runtime exécute. Le code MSIL (Microsoft Intermediate Language) figurant dans un fichier exécutable portable ne sera pas exécuté s'il ne possède pas de manifeste de l'assembly associé. Notez que chaque assembly ne peut avoir qu'un seul point d'entrée (DllMain, WinMain ou Main).

- Il forme une limite de sécurité. Un assembly correspond à l'unité au niveau de laquelle les autorisations sont demandées et accordées.

- Il forme une limite de type. L'identité de chaque type inclut le nom de l'assembly dans lequel il réside. Un type nommé MyType chargé dans la portée d'un assembly est différent d'un type nommé MyType chargé dans la portée d'un autre assembly.

- Il forme une limite de portée de référence. Le manifeste de l'assembly contient les métadonnées de l'assembly qui permettent de résoudre les types et de satisfaire aux demandes des ressources. Il spécifie les types et les ressources qui sont exposés en dehors de l'assembly. Le manifeste énumère également les autres assemblies dont il dépend.

- Il forme une limite de version. L'assembly correspond à la plus petite unité versionable du Common Language Runtime ; tous les types et les ressources figurant dans le même assembly sont versionés sous la forme d'une unité. Le manifeste de l'assembly décrit les dépendances de versions que vous spécifiez pour les assemblies dépendants.

- Il forme une unité de déploiement. Lorsqu'une application démarre, seuls les assemblies que l'application appelle initialement doivent être présents. Les autres assemblies, tels que les ressources de localisation ou les assemblies contenant des classes d'utilitaires, peuvent être extraits sur demande. Cela permet aux applications de rester simples et basiques lors de leur premier téléchargement.

- Il s'agit de l'unité au niveau de laquelle l'exécution côte à côte est prise en charge.

Qu'est-ce qu'un service web XML ?

Un service Web est un composant logiciel encapsulant des fonctionnalités métier de l'entreprise et accessibles, grâce à des protocoles Internet standards, depuis n'importe quelle plate-forme ou langage de programmation. Ils sont décrits dans des documents WSDL (Web Service Description Language), qui précisent les méthodes pouvant être invoquées, leurs signatures et les points d'accès du service (URL, port .). Les services Web sont accessibles via SOAP, la requête et les réponses sont des messages XML transportés sur HTTP.

Qu'est ce que le code Managed/Unmanaged ?

Managed :

Le framework .NET propose un ensemble de services aux programmes qui l'utilisent comme la gestion des exceptions et la sécurité. Pour que ces services fonctionnent, le code doit implémenter un minimum d'informations ; un tel code est appelé Managed Code. Toutes les sources C# et VB .NET sont managées par défaut. Les sources C++ ne le sont pas mais il est possible de spécifier au compilateur de produire du code managé en ligne de commande (/com+).

Unmanaged :

C'est du code natif accédant directement aux services des librairies d'exécution du C/C++ et du système d'exploitation dont la compilation génère de l'assembleur type 80x86.

Qu'est ce que SOAP ?

SOAP (Simple Object Access Protocol) est un protocole d'échange de données entre des systèmes distribués, décentralisés. SOAP a vocation d'universalité : les données sont échangées au format XML et le protocole de transport " par défaut " est http (d'autres protocoles de transport sont envisageables). L'unité de base SOAP est le message. Ce message est vu comme une enveloppe qui est composée d'une entête ("header") contenant quelques attributs généraux (destinataires et un corps de message (" body ")) qui contient les données structurées et typées. Le champ d'action de SOAP est immense : il n'est pas lié à un langage (Java, C#, C, C++, Cobol ou à une plate-forme particulière : mobile, pda, pc, mac, appareil photo, appareil électroménager. On peut tout imaginer via SOAP : envoyer des photos, effectuer des virements, obtenir les derniers cours de bourse, interroger un annuaire. Les services web sont techniquement basés sur SOAP : découvrir et accéder à des services en ligne.

Qu'est ce que Universal Description, Discovery and Integration (UDDI) ?

Les spécifications UDDI (Universal Description, Discovery and Integration) définissent une méthode standard de publication et de découverte d'informations sur des services Web XML. Les schémas XML associés à UDDI déterminent quatre types d'informations qui permettent à un développeur d'utiliser un service Web XML publié : les informations sur l'entreprise, les informations de service, les informations de liaison et les informations sur les spécifications des services.

Annexe B

HTML

HTML

HTML signifie *HyperText Mark-up Language*. Comme son nom l'indique, c'est un langage de description qui permet de définir l'habillage d'un document, c'est à dire la façon dont il doit s'afficher à l'écran d'un navigateur.

Cette notion d'habillage est importante : elle signifie qu'une page écrite en HTML comportera du texte, bien sûr, mais aussi des codes ou balises permettant de modifier l'affichage de ce texte, à savoir sa forme, sa taille ou sa couleur. Le HTML permet également d'inclure des images, du son ou des animations dans une page Web et d'établir des relations cohérentes entre ces informations grâce aux liens hypertextes.

Le HTML est très simple. Ce n'est pas un langage de programmation dans le sens où il n'existe pas de variables, boucles, expressions conditionnelles. En fait, c'est plus un ensemble de codes qu'un langage, comme on le conçoit en informatique.

Il faut également signaler qu'un document HTML n'est autre qu'un fichier texte auquel on a ajouté des balises HTML **[6]**.

1.2/ Les balises HTML

Les pages HTML sont construites selon le modèle suivant:

<HTML> balise de début

<HEAD>	En-tête
<TITLE>	Information non obligatoire: titre de la page,
	métadonnées.
</TITLE>	Description du contenu de la page.
<META>	Ex: Afficher la source de la page d'acceuil du site, le
	titre dans la balise TITLE est repris au sommet de la
	page, dans la barre de titre.
</META>	meta name=Keyword... ...., description, creator...,
</HEAD>	
<BODY>	Corps du document.
</BODY>	C'est le contenu visible du document

</HTML> balise de fin.

Lors de l'indexation, les balises les plus importantes sont :

- <TITLE> : qui contient le titre de la page.
- <META> : qui contient éventuellement les mots clés.
- <A href> : qui contient les liens hypertextes présents dans la page. On trouve aussi la balise <FRAME> qui peut contenir des liens hypertextes vers d'autres documents.

Exemple de page HTML :

<HTML> <HEAD>

<TITLE> Moteur d'indexation et de recherche **</TITLE> </HEAD>**

<BODY>

<H1> TITRE 1 **</H1>**

Cliquer **<A HREF="**http://Uni_Blida/index.html**">ICI</A>**

</BODY> </HTML>

Résultat :

Figure Résultat du code source de l'exemple

Annexe C

Intranet

1 INTRANET

1.1 Généralités

L'usage des **technologies de l'information et de la communication** constitue un enjeu majeur pour les entreprises. L'intranet, outil multi ressources par excellence, ouvre des opportunités considérables en matière de gestion des ressources humaines et de développement des compétences grâce à l'interactivité et la convivialité du multimédia, associées à la puissance des réseaux.

Les entreprises privilégient la communication interne sur leur Intranet. Une récente étude du Cutter Consortium auprès de 154 grandes entreprises a montré que 77,6% d'entre elles prévoient d'utiliser (ou utilisent déjà) l'intranet pour gérer leur communication interne. Le principal usage de l'intranet pour 60% des répondants est la transmission des dernières informations concernant l'entreprise et à 48% le travail en groupe **[10]**.

1.2 Qu'est qu'un réseau Intranet ?

Pour faire simple, c'est un ensemble d'ordinateurs et autres équipements informatiques (imprimante, scanner, modem, ...) interconnectés via des moyens de télécommunication (cartes réseau, hubs, paire torsadée, ...) pour permettre à ces équipements de communiquer. L'architecture Intranet comporte plusieurs services : Messagerie, forums de discussions, partage de fichiers, accès aux applications, bases de données de l'entreprise et le serveur Web qui joue le rôle de relais entre ces différents services **[11]**.

1.3 Les Objectifs d'un Intranet

Tout Intranet doit pouvoir réaliser les services suivants **[10]**:

- Publier des informations de l'entreprise auprès des acteurs concernés, indépendamment de leurs situations géographiques et du type de matériels qu'ils utilisent.
- Partager les informations et les services nécessaires à la gestion de l'entreprise.
- Utiliser des applications décisionnelles ou transactionnelles, en autorisant la saisie et la mise à jour d'un certain nombre de données et documents.
- Intégrer des applications existantes sans aucune peine.

1.4 Avantages des Intranets

Au moment où les sites Web d'Internet sont en cours d'évolution vers des applications Web sophistiquées, les sociétés commencent à utiliser la technologie Web au sein de leurs réseaux privés comme un moyen efficace pour créer et étendre des applications à usage interne. Les Intranets présentent les avantages suivants :

- **Coûts d'extension moins élevés**

Dans la mesure où les applications Intranet résident sur les serveurs, les utilisateurs peuvent passer directement au site Intranet et accéder à l'application sans avoir à installer de nouveaux composants. Si les fonctionnalités de l'application doivent être modifiées, l'administrateur peut procéder à des mises à jour sur le serveur et ainsi permettre aux utilisateurs de bénéficier instantanément des nouvelles possibilités de l'application **[10]**.

- **Applications multi plates-formes**

Les applications Intranet se présentent sous forme de pages HTML et sont par nature compatibles avec plusieurs plates-formes. Ainsi, les organisations équipées de plates-formes bureautiques hétérogènes peuvent assurer que les utilisateurs pourront accéder à l'application sans qu'il soit nécessaire de prévoir une version spéciale pour chaque plate-forme.

- **Des applications à faible débit**

Vu que la plupart des traitements sont effectués sur le serveur et seules les pages HTML sont remises au client, les applications Web sont automatiquement adaptées aux connexions à faible débit.

1.5 Les serveurs d'un Intranet

Les serveurs constituent l'une des richesses fondamentales d'Internet. Des logiciels spécifiques (*navigateur ou browser*) permettent l'accès à l'information en utilisant les fonctionnalités de ces serveurs, dont nous citons les suivants **[11]**:

- **Serveur Web**

La convivialité des systèmes clients (*navigateur*) et la richesse fonctionnelle apportée par le concept d'hyperlien (*lien entre documents*) ont contribué au succès rapide de serveurs Web qui sont devenus le standard de développement de bases d'infor mations sur le Net.

➢ **Serveur de messagerie électronique**

La messagerie concerne l'échange de courrier entre deux personnes, entre une personne et un groupe de personnes ou encore entre une personne et un système automatique. Les protocoles standards sont :

- SMTP : Simple Mail Transfert Protocol.
- POP3 : Post Office Protocol.
- IMAP4 : Interactive Mail Access Protocol.
- MIME : Multipurpose Internet Mail Extension.

➢ **Serveur de fichier (serveur FTP)**

FTP ***(File Transfert Protocol)*** permet de stocker des fichiers et les fournir aux utilisateurs ayant un compte sur des serveurs. Il permet aussi d'exporter des fichiers vers le serveur et d'importer des fichiers sur l'hôte local. Selon les droits d'accès, il est possible de créer, de supprimer et consulter des fichiers et des répertoires du serveur.

➢ **Pare-feu (firewall) et le Serveur Proxy**

L'ouverture des réseaux d'entreprises au monde extérieur, les accès à Internet par exemple, la décentralisation des traitements et des données ainsi que la manipulation des postes de travail accroissent les risques et fragilisent les réseaux à toute attaque de l'extérieur. Les menaces peuvent être regroupées en deux catégories : **les menaces contre les systèmes, les menaces contre les données.** Une des solutions les plus utilisées dans les Intranets, est le ***Pare-feu*** (ou firewall). Le pare-feu est une machine qui réalise des fonctions de filtrage avancées.

➢ **Le serveur de nom de domaine DNS**

Les machines appelées **serveurs de nom de domaine** permettent d'établir la correspondance entre le nom de domaine et l'adresse IP sur les machines d'un réseau. Chaque domaine possède ainsi, un serveur de noms de domaines, relié à un serveur de nom de domaine de plus haut niveau. Ainsi, le système de nom est une architecture distribuée, c'est-à-dire qu'il n'existe pas d'organisme ayant à charge l'ensemble des noms de domaines. Par contre, il existe un organisme (l'InterNIC pour les noms de domaine en .com,.net,.org et .edu par exemple). Le système de noms de domaine est transparent pour l'utilisateur.

2 LES PROTOCOLES

Les protocoles de transfert de données sont des règles et des standards adoptés pour assurer une communication entre différents ordinateurs ayant des systèmes exploitations hétérogènes **[11]**.

2.1 HTTP (HyperText Transfert Protocol)

Le but de ce protocole est de répondre au client en temps utile. C'est pour cela que la connexion est fermée à chaque fois qu'une requête a été satisfaite. On ne maintient pas de connexion ouverte quand on n'en a plus besoin. Ainsi le serveur ne peut pas maintenir l'état de la connexion : nous parlons alors de protocole sans état contrairement à FTP.

C'est le premier protocole de distribution de l'information sur le Web, il assure le transfert de fichiers hypertextes entre un serveur Web et client Web (navigateur). La procédure suivie par ce protocole est très simple :

- Etablissement de la connexion,
- Envoi de la requête précisant le document à consulter,
- Réponse du serveur avec un code d'état pour la requête,
- Déconnexion par l'une ou l'autre des machines.

Dans ce protocole, nous trouvons la notion d'URL (Uniform Resource Locator) : il s'agit là d'un schéma d'adressage standardisé permettant de localiser et de retrouver tous document (programme, ficher et répertoire) sur Internet/Intranet.

2.2 Concepts des réseaux TCP/IP

À mesure qu'Intranet et Internet s'étendent, l'emploi de TCP/IP (Transmission control Protocol/Internet Protocol) sur les réseaux internes se généralise. TCP/IP offre un éventail considérable de protocoles ouverts, tout à fait adaptés aux réseaux étendus (WAN pour Wide Area Network).

2.2.1 Internet Protocol (IP)

IP est un protocole sans connexion, car il permet l'envoi des paquets vers l'hôte de destination sans établir une connexion. L'entête de celui-ci contient suffisamment d'informations pour qu'il soit transporté jusqu'à destination. Il est non fiable, car il n'effectue

aucun contrôle d'erreur (ce dernier sera à la charge d'un protocole de couche supérieure s'il est nécessaire).

2.2.2 Transmission Control Protocol (TCP)

TCP est un protocole orienté connexion qui permet à une application d'être sûre de l'envoi et de la réception des paquets IP sur le réseau, en assurant qu'ils parviennent de la machine source à la machine cible.

3 LES TYPOLOGIES DES ARCHITECTURES INTRANET

La technologie informatique évolue et les applications d'aujourd'hui sont de plus en plus complexes. Le développement d'applications fonctionnant en mode Client-Serveur peut être vu comme l'une d'entre elles. elle constitue un chaînon logique dans l'évolution des architectures informationnelles. Après l'informatique centralisée des années 70, s'est développée une informatique individuelle autour du phénomène «micro-ordinateur» dans les années 80. La rencontre entre ces deux mondes dans les années 90 s'est concrétisé par le modèle Client-Serveur et l'informatique de groupe. On ne choisit pas une architecture Client-Serveur en suivant un phénomène de mode, elle est dictée par des raisons économiques et stratégiques **[11]**.

Le développement rapide et l'adoption par tous les acteurs majeurs du marché des technologies Internet/Intranet font émerger un nouveau modèle d'architecture Client/Serveur (dit «multi-tiers»), dont la caractéristique principale est de renforcer le rôle du réseau et des serveurs d'applications au détriment de celui des postes de travail.

L'intérêt majeur de ce modèle est que, par construction, il combine les avantages des systèmes centralisés (cohérence globale, simplicité de développement, d'exploitation et de maintenance, maîtrise des coûts,...) et ceux du Client/Serveur (meilleure ergonomie du poste de travail, bonne intégration avec les outils bureautiques ou les autres applications, gains de productivité,...).

Les applications constituant le système d'information d'une entreprise comportent généralement trois grands types de composants :

- Les fonctions de présentation (ou Interface Homme/Machine).
- Les traitements.
- L'accès aux données réparties ou distribuées sur plusieurs serveurs hétérogènes.

Les architectures Intranet se distinguent en deux grandes familles :

- Les architectures deux niveaux, où un client communique directement avec un serveur de traitement de données ou de base de données.
- Les architectures trois niveaux ou multi niveaux où au moins un serveur intermédiaire est intercalé entre les postes clients et les serveurs de données.

3.1 Architectures à deux niveaux

Dans ce type d'Intranet, il y a deux pôles, le premier est le poste client avec une interface qui permet d'exprimer des requêtes directement vers le deuxième pôle, qui est le serveur de base de données. Autrement dit, pas d'intermédiaire entre interface utilisateur, traitement et données **[4]**.

Comme le schématise la figure ci-dessous, il existe deux types d'architectures à deux niveaux :

- Le premier ne différencie pas réellement l'interface utilisateur des traitements : les objets de l'interface utilisateur lancent eux-mêmes les requêtes (SQL, ...) vers le moteur de base de données et récupèrent les résultats directement.
- Quant au second, il ne différencie pas les traitements des données : c'est le cas des procédures stockées pour un SGBD relationnel par exemple.
- Les architectures à deux niveaux sont faciles à mettre en oeuvre mais disposent de certaines limitations qui peuvent être résolues par les architectures à trois ou multi niveaux.

Cependant les architectures à deux niveaux peuvent être mis en place assez aisément dans le cadre de petites applications et un certain nombre d'outils de développement s'appuie sur ce modèle.

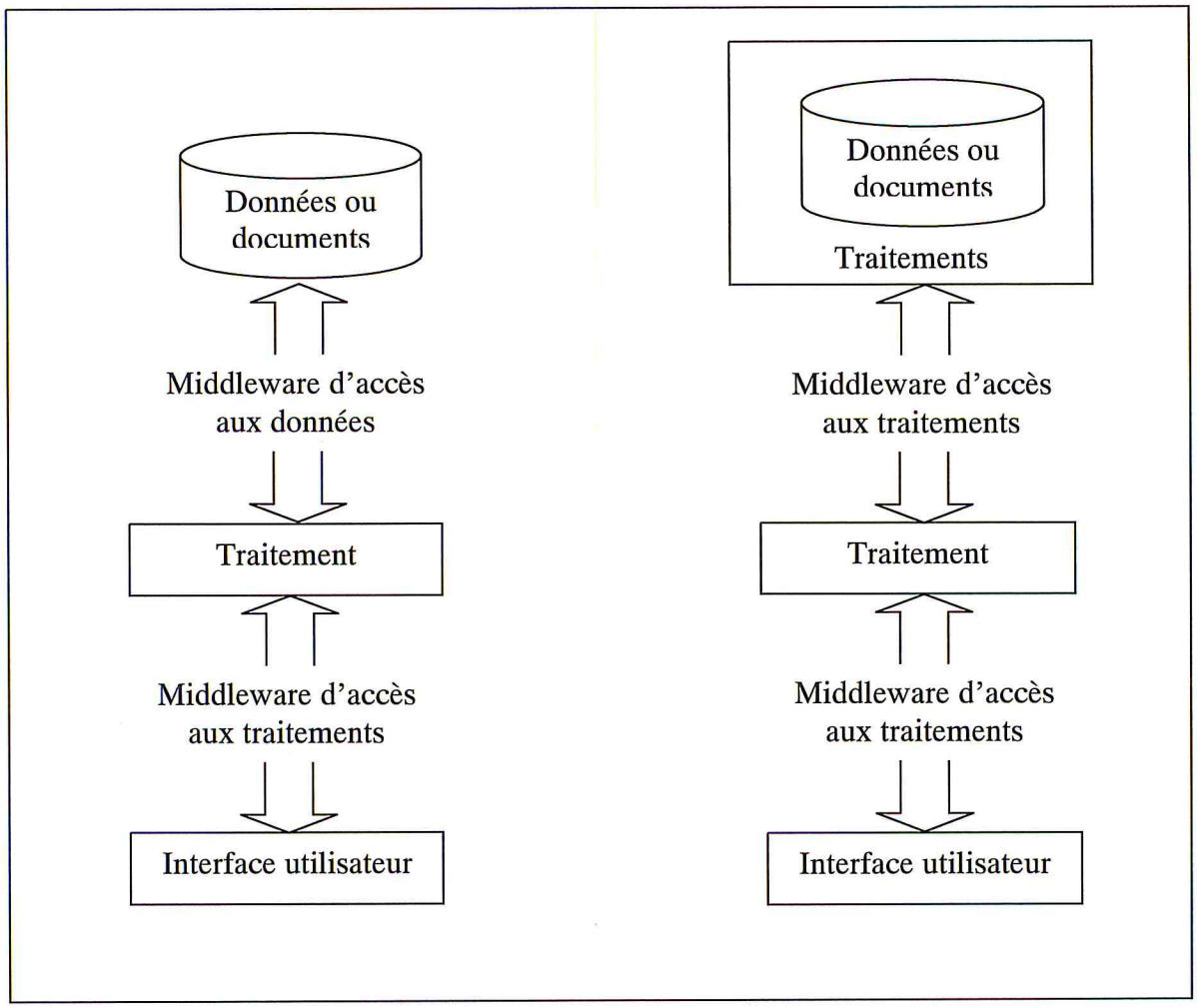


Figure 2 Les types d'architectures trois niveaux

3.3 Communication entre Serveur (WEB) et Client (Browser)

3.3.1 Architecture traditionnelle d'une application Web

L'architecture d'une application WEB présente traditionnellement trois niveaux (on parle d'architecture à trois tiers):

- Le navigateur représente le premier pôle : comme interface homme/machine.
- Le serveur WEB constitue le second pôle : Il est au coeur de l'architecture. C'est le point de passage obligé entre les demandes des clients et les traitements.
- Et comme troisième pôle les bases de donnés et les ressources détenant les services.

3.3.2 Connexion à un serveur Web

Le browser Internet (ou navigateur) se comporte comme un client puisqu'il demande un service au serveur Web quand l'utilisateur formule l'url d'une page HTML ou en cliquant

sur un *ancre* (lien Hypertexte) dans la page courante. Le browser (client) commence toujours par établir une connexion avec le serveur en utilisant l'adresse URL.

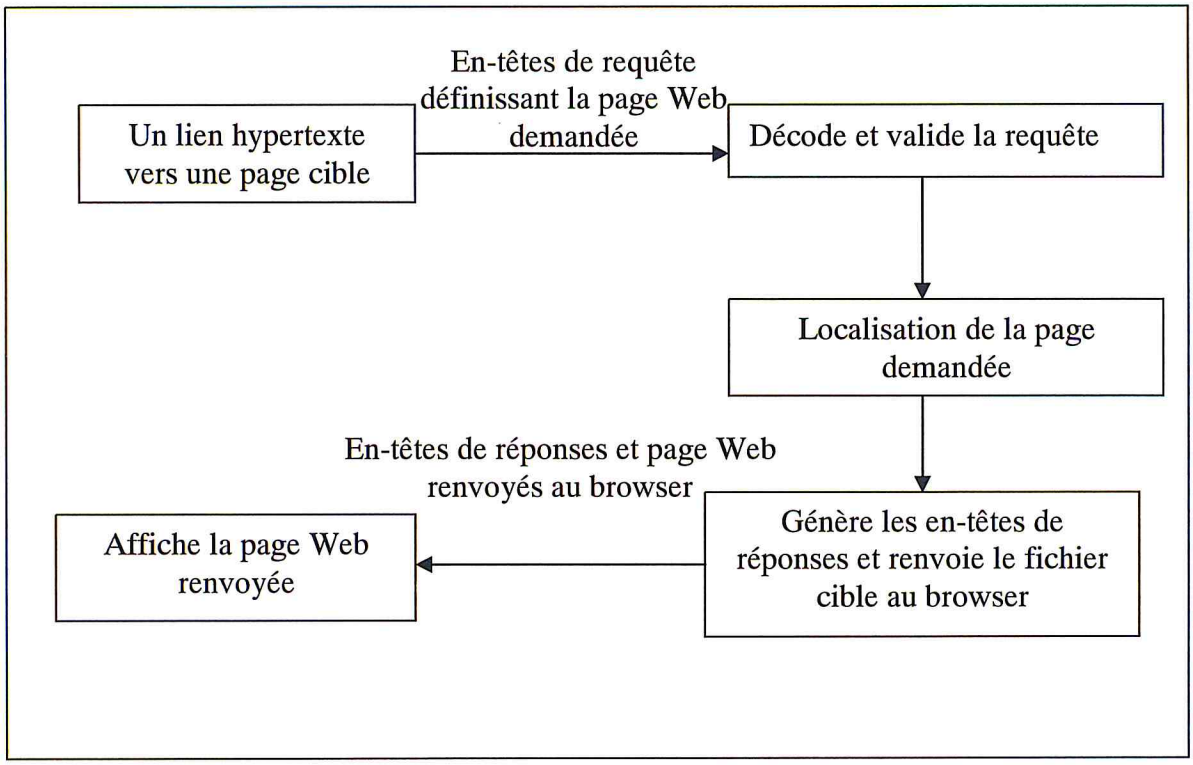


Figure 3 Les étapes d'une connexion entre un client et un serveur Web

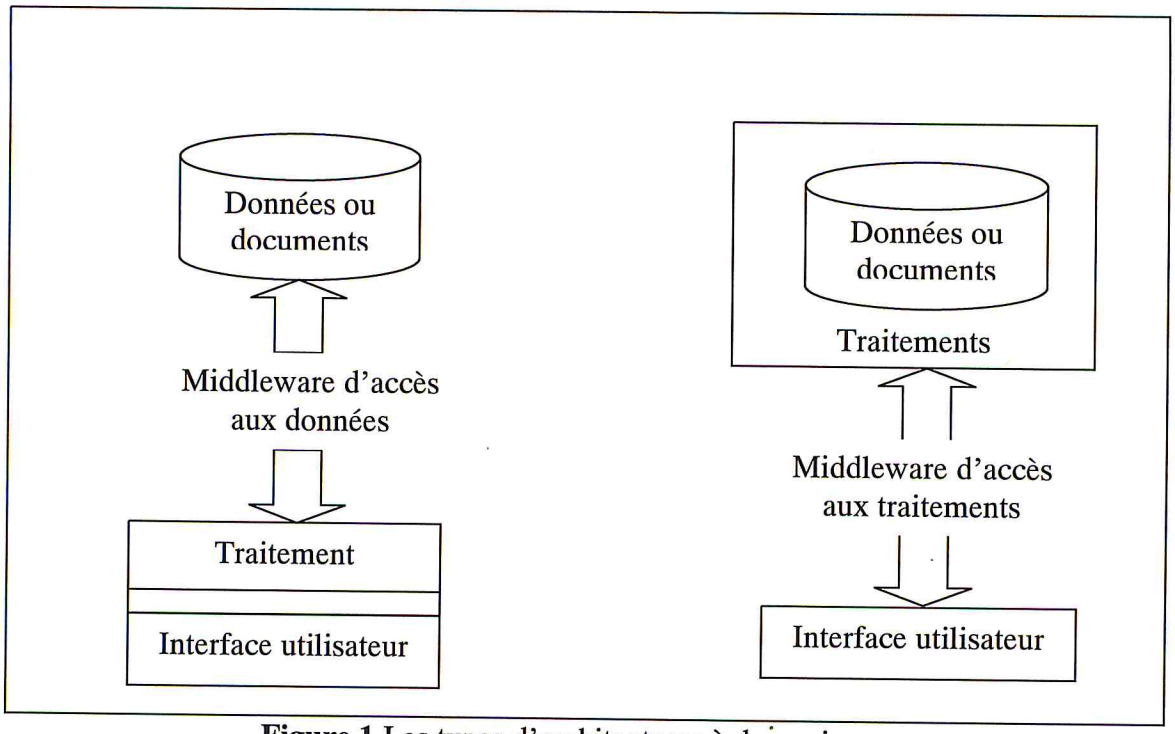


Figure 1 Les types d'architectures à deux niveaux

3.2 Architectures à trois niveaux

Une architecture logicielle à trois niveaux ou plus effectue nécessairement une séparation dans le codage et les outils mis en oeuvre entre interface utilisateur, traitements et données. Comme l'architecture à deux niveaux, l'architecture à trois niveaux comporte deux types **[4]** :

- Le premier n'affecte qu'une partie du traitement aux bases de données : les objets de l'interface utilisateur lancent les requêtes (SQL, ...) vers le serveur de traitements qui les transmet au moteur de base de données. Ce dernier effectue quelques traitements sur ces données et remet les résultats au serveur de traitements qui les remet à l'utilisateur.
- Quant au second, il affecte tous les traitements possibles aux serveurs de traitements, qui ont le rôle de relais entre l'interface utilisateur et les bases de données.

Cette architecture à trois niveaux est particulièrement bien adaptée pour les architectures Intranet, car elle confère les avantages suivants :

- La gestion des données reste centralisée, ce qui d'une part simplifie l'exploitation et d'autre part garantit la cohérence globale du système d'information.
- L'usage d'un Browser comme client universel, adapté aux besoins des applications, assure l'indépendance vis à vis du type de poste de travail.

GLOSSAIRE

GLOSSAIRE

ASCII

Acronyme de : American Standard Code for Information Interchange. L'ASCII standard est codé sur sept bits, et autorise donc le codage de 127 caractères au maximum, ce qui est insuffisant pour les langues européennes. Il existe des ASCII « étendus » codés sur huit bits, plus ou moins officiels et correspondant aux jeux de caractères nationaux. Par extension, on a coutume de parler de « format ASCII » pour désigner un texte ne comportant pas de « code de contrôle » (généralement, les 31 premiers codes), en particulier pour la mise en forme, même s'il comporte des caractères huit bits (comme les lettres accentuées de l'alphabet occidental), voire aussi s'il contient en outre des « balises ».

DBA

En anglais : Data Base Administrator. Administrateur de base de données.

DLL

En anglais : Dynamic Link Library. Librairie de routines spécialisée chargées en mémoire par un programme selon ses besoins et exploitable également par d'autres programmes chargés en même temps. Présente l'avantage d'un code commun sauve dans un fichier séparé d'où gain de place sur disque et en mémoire.

DNS

En anglais : Domain Naming System / Domain Name Server. Base de données de transcodage d'une adresse réseau littérale en une adresse IP.

FTP

En anglais : File Transfert Protocol. Protocole de transfert de fichiers entre stations de travail.

HTML

En anglais : Hyper Text Markup Language. Langage de programmation de pages consultables dites HTML. Ce langage est constitué de balises ou tags définissant le document (entête, corps, titre, texte, ...)

HTTP

En anglais : Hyper Text Transfert Protocol. Protocole de communication utilisé pour le transfert entre un serveur et un poste client des pages au format HTML des sites web.

IP

En anglais : Internet Protocol. Protocole d'interconnexion de deux sous réseaux. Développé à l'origine pour la défense US. Ce protocole est chargé principalement de la gestion des adresses de destination affectées à chaque paquet de données transitant entre les sous réseaux interconnectés.

MD5

En anglais : Message-Digest 5. L'algorithme MD5 a de multiples applications en cryptographie et sur l'Internet en général, pour vérifier l'intégrité d'un message ou d'un téléchargement.

PHP

Officiellement : Hypertext Pre-Processor. C'est un langage de script HTML, qui fonctionne côté serveur.

POP3

En anglais : Post Office Protocol, Version 3. Protocole de la couche applicative. Utilisé pour télécharger ses e-mails avec les logiciels de messagerie électronique (KMail, Messenger, etc.)

SMTP

En anglais : Simple Mail Transfer Protocol. Protocole de la couche applicative. Utilisé pour envoyer des e-mails par les logiciels de messagerie électronique (KMail, Messenger, etc.)

SSL

En anglais : Secure Socket Layer. Le protocole SSL a pour but de sécuriser la communication entre deux points d'un réseau.

TCP

En anglais : Transmission Control Protocol. Protocole de la couche transport, fiable, full-duplex et en mode connecté. Utilisé par la majorité des applications et des protocoles de la couche applicative d'Internet, notamment avec HTTP (Netscape), Telnet (Connexion sur un

ordinateur distant), SMTP (envoi d'e-mail), POP3 (réception d'e-mail) et FTP (transfert de fichier).

TCP/IP

TCP/IP est un ensemble de protocoles permettant aux ordinateurs de partager leurs ressources à travers un réseau. Il a été développé par une communauté de chercheurs dans le cadre de ARPAnet. ARPAnet est certainement le plus connu des réseaux TCP/IP. En juin 87 au moins 130 fournisseurs offraient des produits supportant TCP/IP, et des milliers de réseaux les implémentaient.

TELNET

Protocole de la couche applicative. Utilisé pour se connecter à un serveur distant, Terminal Virtuel Internet.

U.M.L

U.M.L est un langage destinée aux phases amont de la réalisation d un logiciel. U.M.L est une technique de modélisation unifiée issue de méthodes plus anciennes comme O.M.T, de O.O.S.E et de O.O.D.

URL

En anglais : Uniform Ressource Locator. Méthode d'identification d'une ressource Internet (serveur, fichier, groupe de discussion,...) et le moyen d'y accéder. Ex : http://www.usad/index.php indique d'aller chercher la page index.html du serveur usad à l'aide du protocole de transfert http.

BIBLIOGRAPHIE

Bibliographie :

[OSMA 99] Osmar R. Zaiane.
From resource discovery to knowledge discovery
on the Internet, 1999

[FORE 03] Alain Forestier
Qu'est ce qu'un algorithme de moteur de recherche?, 2003

[LAWR 98] Steve Lawrence & C.Lee Giles. « How big in the Web», 1998.

[BENA 00] BENAOUDA Adel & MOKHTARI Amdjed.
Conception et réalisation d'un moteur d'indexation et de recherche
dans un intranet . INI 2000 (PFE).

[KOST 95] Martijn Koster.
Robots in the Web: threat or treat, 1995

[MICH 03] Michael Chau & Hsinchun Chen.
Personalized and Focused Web Spiders», 2003

[LELO 98] Catherine LELOUP
MOTEURS d'INDEXATION et de RECHERCHE
Environnements Client-Serveur, Internet et Intranet
Edition : Eyrolles, 1998

[LATO 98] Djamel LATOUI & Zahir MAAFA
Réalisation d'un système pour la recherche documentaire
dans une base des Hyperdocuments, INI 1998

[DROU 03] Patrick Drouin
Acquisition automatique des termes
l'utilisation des pivots lexicaux Spécialisés, 2003

[MONT 98] Jean-Luc MONTAGNIER
Pratique des réseaux d'entreprise
Edition : Eyrolles, 1998

[GARD 99] Georges et Olivier GARDARIN
Le Client Serveur
Edition : s&sM, 1999

[DELA 99] Quentin DELACROIX (Doctorat)
Un système pour la recherche plein texte et la consultation
hypertexte de documents techniques » Université Blaise Pascal
Clermont-Ferrand II, Juillet 1999

[ELHA 97] Mabrouka EL HACHANI (DEA)
Indexation automatique
Ecole Nationale Supérieure des Sciences de l'Information
et des Bibliothèques ENSSIB) 1997.

[MULL 00] Pierre-alain Muller « modélisation objet avec UML, 2000

[URL01] www.dsi-info.ca/moteur-de-recherche/

[URL02] www.developpez.com/faq

[URL03] PHP.NET - Le site officiel de PHP.
http://www.php.net

RESUME

Le travail réalisé dans ce projet de fin d'études fait partie du domaine de la recherche documentaire. Il représente un outil de recherche d'informations renfermées dans le corpus du système d'information de l'Universite de Blida. Il vient d'amélioré le travail effectué au niveau de ce dernier sur le plan d'échange d'informations entre utilisateurs en les dotant d'un outil qui facilite l'accès aux informations désirées dans des délais raisonnables. Il leur permet aussi la publication des documents en les partageant avec un nombre restreint d'utilisateurs. Les méthodes adoptées pour la réalisation d'un tel système restent ouvertes à d'éventuelles améliorations, particulièrement dans le processus d'indexation des documents.

Mots clés :

Recherche documentaire, indexation, recherche, présentation, Robot, Spider, intranet,

ABSTRACT

The tool developed in this project takes a part in the field of documentary search. It contributes in the improvement and enrichment of Blida Universite Intranet network, making easier for its users researches for information contained within the corpus of the network. Also, it allows to them to publish their files and share them with other users. The different methods used to elaborate this tool can be enhanced by introducing of several improvements specially in the process of indexation.

Keywords :

Search engine, indexation, presentation, research, spider, Intranet