Institut Supérieur des Etudes Technologiques en Communications de Tunis

Projet de fin d'études

CONCEPTION ET IMPLEMENTATION D'UN MOTEUR DE RECHERCHE

Réalisé par

Sami Turki

Sabri Ben Amara

TS5- Télécommunications

Encadré par

Sami Ben Khélifa

2001- 2002

REMERCIEMENTS

C’est avec un grand plaisir que nous réservons cette page en signe de gratitude et profonde reconnaissance à tous ceux qui nous ont aidé à la réalisation de ce projet de fin d’étude.

Nous tenons à exprimer nos vifs respects à tous le corps enseignant et administratif de l’Iset’Com pour leurs efforts.

Nous remercions également M Sami **Ben Khelifa** notre encadreur qui nous a accueilli favorablement malgré la densité dont il a à sa charge.

Chacun de :

- M[r] Facker **Ben Fradj** chef service web .
- M[r] **Miled Raouf** Infographiste;
- Tous ceux qu’on a côtoyés au sein de 3S GlobalNet.

Comme nous remercions nos professeurs, membres ou Jury pour l’honneur qu’ils ont fait en acceptant de faire partie de notre Jury.

SOMMAIRE

ANNEXE

Bien Lancer Son Site

C'est Réussir Son Insertion Au Web

Le lancement d'un site est une occasion pour rassembler pleins **d'opinions** de **partenaires** et de **visiteurs**. Sachez que ce qui est nouveau attire l'attention, il suffit de voir les publicités à la TV, d'où l'opportunité de rassembler les 3 éléments ci-dessus.

- **L'opinion :** c'est en fait l'avis des Internautes souvent impartial qui permet de prévenir les différents bugs et de déceler les défauts que vous n'avez pas pu voir avec votre oeil de webmaster (on est souvent aveuglé par ce que l'on fait et on refuse donc de voir par nous même certaines erreurs, ou améliorations potentielles). Cela peut paraître élémentaire, mais plusieurs internautes font de flagrantes erreurs (présentation inadaptée, chargement trop long avec certains modems...). Et ce parce qu'ils se sont fiés à des amis, ou à leur propre instinct. Lorsque votre site est fini, c'est la première chose à faire avant le "grand" référencement.
- **Les partenaires :** le système de partenariat est très exploité sur Internet.. C'est au lancement de site qu'il faut tisser un réseau de partenaire efficace. Alors comment faire ? Par la même occasion que lors de la recherche de partenaires, surfez un peu chez vos "concurrents" et repérez les sites qui mettent bien en valeur les partenaires. Et affichez vos partenaires comme un militaire qui exhiberait fièrement ses récompenses d'ancien combattant. Par ces 2 choses simples, vous pourrez gagner un capital de visiteurs de 25% du total. J'insiste sur le fait que certains partenaires doivent être en place dès le lancement, c'est comme ça que vous déclencherez un maillon du fameux effet "boule de neige
- **Les visiteurs :** c'est évidemment l'élément le plus important du lancement. Avoir un grand nombre de visiteurs dès le départ, c'est avoir un **trafic important pour toujours** (ceci **si votre site est souvent mis à jour)**. Sachez qu'il y a 2 types de visiteurs : ceux qui viennent des moteurs, annuaires et autres top listes, et les autres qui ne dépendent pas d'un autre site Internet. Je parle de l'action sur les forums et les chats, mais pas du Spamindexing. En outre, si en plus d'être "bien" votre site est aussi souvent mis à jour et dispose d'un nom de domaine, les personnes qui le visiteront y reviendront et là, le bénéfice de trafic sera conséquent.

I- Erreurs à éviter lors du lancement :

- **Ne pas mettre de pages en construction** : il n'y a rien de plus agaçant qu'un site avec plein de pages en construction. Il est conseiller d'annoncer des mises à jour plutôt que de mettre de pages vides.

- **Le spam** : si vous ne savez pas ce que c'est tant mieux pour vous. Sinon sachez qu'il apporte plus de problèmes que de visiteurs en vous considérant ficher dans les listes rouges des moteurs de recherche.
- **Hors sujet** : si vous voulez que votre site ne se limite pas à une petite page perso (parmi tant d'autres) évitez les textures et motifs de fond trop voyants, les gifs animés à chaque coin de pages, l'emploi des frames et vérifiez le temps de chargement de vos pages.

II- Conseils et astuces :

- **Faites preuve d'empathie** : faire preuve d'empathie c'est savoir se mettre à la place de quelqu'un pour mieux le comprendre. Si vous même étiez un simple, quel genre de dialogue pourrait vous convaincre de visiter tel ou tel site ?
- **Les balises méta-tags sont indispensables dès le lancement** : car certains moteurs revisitent très rarement les sites (voire qu'une seule fois, mais ce sont plutôt des annuaires avec moteur interne).

Conclusion :

Finalement, le lancement du site devient une vraie science qui devra mobiliser tout votre temps lorsqu'il aura lieu. Remarquez que plusieurs conseils ici présents s'appliquent durablement, et ne sont pas forcement indispensable dès le lancement.

Description Des Critères D'évaluation
Des Moteurs De Recherche

I- Description des moteurs :

1) Editeur :

La liste des organisations/sociétés qui sont intervenues dans la mise en place du moteur de recherche. Cette liste contient le nom de l'éditeur, un lien vers cet éditeur (s'il possède un site Web), ainsi que son URL et le pays de son site.

2) Type de Service :

Le type de service fourni par le moteur : Index Automatique, Index Manuel, Annuaire, Index Automatique et Annuaire, Index Manuel et Annuaire.

3) Type d'accès :

L'accès à ce service est-il gratuit (Public) ou payant (Commercial), ou bien les deux types d'accès sont-ils proposés (Public et Commercial).

4) Fréquence de mise à jour :

Evaluation moyenne de la fréquence de mise à jour de l'ensemble de l'index de la base.

5) Fréquentation moyenne :

Evaluation moyenne du nombre de requêtes adressées au service pour une période donnée.

6) Sites Miroirs :

Liste des sites miroirs du moteur de recherche. Les sites miroirs sont des répliques du site original à des localisations différentes, afin de répartir la charge des machines et de réduire les temps de communication. Cette liste comprend, s'il y a lieu, le nom du site miroir, un lien vers ce site, son URL ainsi que sa localisation.

II- Collecte des documents :

1) Méthode de collecte :

Décrit la manière dont les documents qui seront plus tard indexés sont déjà dans un premier temps collectés. Quatre cas sont possibles :

- Manuelle : des *net-surfer* passent leurs journées à parcourir le Web et à noter les adresses des sites intéressants.

- Automatique : un robot (petit programme) se promène sur le WWW et rapatrie les documents qu'il trouve en se déplaçant de lien en lien.

- Soumission d'URL : dans ce cas, ce sont les auteurs de pages et/ou sites Web qui envoient l'adresse de leurs créations aux moteurs afin que ces derniers indexent leurs pages.
- Suppression d'URL : Ce n'est pas un moyen de collecte de document, mais au contraire un moyen de suppression de document qui devrait logiquement être présent dans tout système permettant de soumettre des URLs (ce qui n'est d'ailleurs bien souvent pas le cas!). Dans le cas d'un moteur acceptant cette fonctionnalité, l'auteur ayant fourni l'URL de ses pages pour indexation a la possibilité de retirer ses dernières de l'index du moteur. Cette décision peut avoir plusieurs causes : Les pages ont été déplacées, ou bien elles n'existent plus, ...

2) Robot de collecte :

Le nom (s'il en possède un) du logiciel effectuant la collecte automatique des documents.

3) Méthode de parcours :

Méthode utilisée par le robot de collecte pour parcourir le Web. Deux possibilités: Largeur d'abord pour les programmes qui, à partir d'une page, parcourent d'un seul niveau tous les liens présents sur celle-ci, profondeur d'abord pour ceux qui à partir d'une page, explorent le premier lien, puis sur la page résultante parcourent à nouveau le premier lien, etc...

4) Protocole standard d'exclusion :

Deux réponses possibles, oui ou non. Dans le cas de l'affirmative, cela signifie que le robot de collecte respecte le protocole standard d'exclusion permettant à tout WebMaster de spécifier des pages Web ne devant pas être collectées par le robot.

5) Serveurs collectés :

Ce critère décrit l'ensemble des types de serveurs collectés par le moteur de recherche. Nous avons restreint leur nombre aux plus essentiels : WWW, Usenet, F.T.P., Gopher, et une rubrique 'Autre', pour les outils collectant des documents à partir d'autres sources.

6) Couverture géographique :

Décrit la couverture géographique du robot de collecte des documents. En effet, de plus en plus d'outils sont spécifiques à un domaine géographique particulier (Europe, France, Pays Francophones, Suisse,...).

7) Type de contenu :

On trouve ici le sujet des documents collectés par le système. Si la plupart s'intéressent à tous les documents (dans ce cas, le type de contenu est étiqueté général), certains vont restreindre leur processus de collecte à certains sujets bien précis (médecine, brevets, informatique, ...)

8) Fréquence de visite des documents :

Ce critère donne une évaluation de la fréquence moyenne de visite des documents par le robot de collecte. En effet, ce dernier doit parcourir le plus fréquemment possible les pages

qu'il a déjà récupérées afin de tenir compte de toute modification du document. Ainsi, plus cette fréquence est élevée, et plus les résultats d'une recherche seront à jour par rapport à la réalité (si l'indexation est aussi fréquente).

III- Indexation des documents :

1) Méthode d'indexation :

Nous distinguons ici deux méthodes d'indexation, l'indexation automatique et l'indexation manuelle.

2) Nom du moteur :

Le nom du moteur d'indexation dans le cas d'une indexation automatique.

3) Données indexées :

Les critères d'indexation peuvent être multiples et variés. Nous avons retenu :

- Le titre du document .
- Ses différents sous-titres (balises <H1>...<Hn>) .
- Son en-tête (le <META> tag) .
- Sa date de création et/ou modification .
- Sa taille .
- Les URLs qu'il cite .
- Le texte des URLs qu'il cite .
- D'autres balises éventuelles .
- Un résumé du document .
- Un extrait du document .
- Enfin le texte intégral du document.

4) Traitements manuels :

Nous regroupons ici quelques traitements manuels complémentaires à l'indexation et pouvant apporter une valeur ajoutée pour la recherche. Ainsi, nous prenons en compte le catalogage des documents qui consiste à regrouper ces derniers par thèmes ou sujets, et la création de résumés manuels, permettant de mieux se rendre compte de la pertinence d'un document lors de l'affichage des résultats qu'un simple extrait.

IV- Recherche des documents :

1) Type de question :

Les systèmes de recherche peuvent proposer différents modes d'interrogation. Cela va de la requête en langage booléen (mots séparés par des opérateurs ET, OU et NON) à la question en langage naturel (formulation d'une question en langage libre) en passant par la requête comportant une liste de mots (qui revient souvent à une question booléenne dont tous les mots sont séparés par un opérateur implicite - généralement le OU) ou l'interrogation par

expression régulière (expression définissant des ensembles de chaînes de caractères à rechercher).

Remarque : Nous ne distinguons pas ici l'usage des opérateurs ET, OU NON, et des opérateurs "+" et "-" souvent utilisés par les moteurs de recherche. En effet, les deux sont plus ou moins interchangeables comme nous le soulignons dans la partie consacrée aux

2) La Troncature automatique :

Nous évaluons ici les possibilités de troncature automatique du système. La troncature automatique consiste à ne fournir qu'une partie du mot et à ce que le système recherche un ensemble de chaînes de caractères dérivées de cette sous chaîne. Ainsi, nous testons si le système supporte

- La troncature automatique gauche : on spécifie une chaîne de caractères, et le système recherche les documents comportant le mot spécifié ou les mots terminant par cette chaîne. Par exemple, pour la chaîne matique, le moteur de recherche va récupérer les documents contenant les termes informatique, télématique, mathématique, ...
- La troncature automatique droite : dans ce cas, on fournit au système la partie gauche d'un mot, et il recherche toutes le chaînes de caractères commençant par cette partie gauche. Par exemple, pour la chaîne inform, le moteur de recherche doit retourner les documents contenant les mots informatique, informaticien, informaticiens, informaticienne, informaticiennes, information, informations, informationnel, informationnels, informationnelle, informationnelles, ... informe, informes, informel, informels, informelle, informelles, informulé, informulés, informulée, informulées. Comme nous le voyons, c'est donc un bon moyen d'étendre une recherche au pluriel, féminin, ou mots de la même famille qu'un terme. Mais comme nous le constatons également, le bruit de la recherche croit énormément, et il est facile d'être rapidement submergé par des mots n'ayant aucun rapport avec la signification recherchée.
- La lemmatisation, qui consiste à rechercher pour un mot toutes ses formes possibles dans la langue (tous les genres, nombres, conjugaisons). Etant donné l'état de l'art dans ce domaine au niveau des moteurs de recherche sur Internet, nous avons étendu le terme de lemmatisation à tout traitement un peu plus évolué qu'une simple troncature droite .
- La possibilité pour l'utilisateur de désactiver les traitements de troncature automatique effectués par le système.

3) La Troncature manuelle :

Tout comme pour la troncature automatique, nous distinguons ici la troncature manuelle gauche et droite. Elles ont les mêmes effets que la troncature automatique, et généralement elle est mentionnée en plaçant le caractère '*' à l'endroit où doit se faire la troncature. Pour reprendre les exemples précédents de la troncature automatique, il faudra donc saisir *matique

et inform*. Nous trouvons également la troncature interne, qui permet de spécifier le début et la fin d'un mot, en laissant une partie *libre*. Par exemple, poi*on donnera poison, poisson, poivron, ...

4) Champs de recherche :

Cette rubrique énumère les divers champs dans lesquels le système effectue la recherche. Nous avons dégagé cinq champs : l'URL du document, son titre, son résumé (généralement dans le cas d'une indexation manuelle, ce critère regroupe aussi les moteurs recherchant dans une liste de mots-clés créée manuellement), son texte intégral (l'ensemble du document moins son titre puisque ce critère est déjà pris en compte).

5) Champs spécifiques :

Ce critère énumère la possibilité offerte à l'utilisateur de spécifier dans sa requête dans quel(s) champs la recherche doit s'effectuer. Aux vues de ce que proposent certains moteurs de recherche, nous avons été très exhaustifs pour ce critère, et les champs spécifiables que nous avons retenus sont les suivants :

- L'URL du document .
- Son titre .
- Ses mots clés (META tag) .
- Son résumé .
- Le texte intégral (pas l'url ni les mots clés ou le résumé,...
- Les URL citées dans le document.

6) L'Elimination des mots vides :

L'élimination des mots vides permet de ne pas prendre en compte certains mots (les mots vides!) trop communs et surtout n'apportant que peu de sens dans un texte tels que les articles. Nous distinguons deux moyens utilisés par les moteurs de recherche pour éliminer les mots vides. La méthode consistant à utiliser une liste de mots vides (Liste), et une autre à éliminer les mots dépassant un certain nombre d'occurrences dans la base. Bien sur, chaque moteur de recherche effectue des variations sur chacune de ces deux méthodes, et c'est pourquoi pour certains d'entre eux un commentaire complémentaire est disponible.

7) Prise en compte de la casse :

Exprime comment le moteur de recherche réagit aux caractères majuscules et minuscules. Dans chacun des deux cas, nous envisageons que le système peut ne rechercher que les chaînes majuscules, ou majuscules, ou bien les deux à la fois.

8) Prise en compte de l'accentuation :

Exprime comment le moteur de recherche réagit aux caractères accentués et non-accentués. Le tableau présente les chaînes recherchées pour une question accentuée et une

question non-accentuée. Dans chacun des deux cas, nous envisageons que le système peut ne rechercher que les chaînes accentuées, ou non-accentuées, ou bien les deux à la fois.

V- Présentation des résultats :

1) Informations générales :

Ce sont des informations de nature diverses relatives à la recherche : le nombre de documents-réponses trouvés par le moteur de recherche, la liste des termes de la question reconnus, de ceux non reconnus, et enfin la liste des mots vides de la question.

2) Organisation de documents-réponses :

Elle doit permettre d'avoir une vue synthétique et efficace des documents-réponses. Pour cela, plusieurs critères essentiels :

- La méthode de tri des documents-réponses qui est propre à chaque système est décrite en quelques mots .
- La caractérisation des documents qui correspond à un regroupement des documents en sous-ensembles permettant d'avoir une bonne vue synthétique des résultats .
- L'élimination des liens dupliqués qui évite à l'utilisateur de voir apparaître plusieurs fois le même document dans la liste des documents-réponses.

3) Informations concernant les documents :

La liste des informations affichées par le système concernant chaque document-réponse. Nous avons retenu :

- L'URL du document .
- Le lien vers le document permettant d'aller le consulter directement (normalement toujours présent) .
- Le titre du document .
- La liste des mots-clés du document .
- Un résumé du document .
- Un extrait du document (différent du résumé dans le sens que l'extrait correspond souvent aux premières lignes du document) .
- Les URL citées dans le document .
- Les URL citant le document .
- La taille du document .
- Sa date de dernière mise à jour connue (par l'auteur) .
- Sa date de dernière visite (par le système) .
- Une mesure de pertinence (score) .
- Enfin la mise en évidence des mots de la question présents dans le document.

4) Informations concernant les documents:

Nous retrouvons ici les mêmes critères que dans la rubrique précédente. La différence est que nous mettons ici en évidence les informations qui peuvent être affichées par le biais d'une option proposée par le système de recherche.

SITES HEBERGES CHEZ

3S GlobalNet

A ce jour, 3S GlobalNet héberge 108 serveurs opérationnels dont les URLs sont les suivantes :

3S GlobalNet
http://www.gnet.tn
L'acropolium de Carthage
http://www.acropolium.com.tn
Africantours
http://www.africantours.com.tn
Agostino améliorants
http://www.agostino.ameliorants.com.tn
Agriforum
http://www.agriforum.com.tn
AON Socargest Tunisie s.a
http://www.aon.com.tn
AMI
http://www.ami.com.tn
AMEN BANK
http://www.amenbank.com.tn
Art Moden
http://www.art-moden.com.tn
Art Moden
http://www.artmoden.com.tn
Apolonia Edition
http://www.apollonia.com.tn
Arabesk
http://www.arabesk.com.tn
Africa Transit Magasin Cale
http://www.atmc.com.tn
Bacosport
http://www.bacosport.com.tn
Ben Rejeb Trading
http://www.benrejeb-trading.com.tn
Beta
http://www.beta.com.tn
Btp-O-Net
http://www.btponet.com.tn
Carthage Tours
http://www.carthagetours.com.tn
Cawtar
http:// www.cawtar.org.tn
CCAT
http://www.ccat.org.tn
CECAI
http://www.cecai.com.tn
Cofitec
http://www.cofitec.com.tn
CJD
http://www.cjd.org.tn
Corail Royal
http://www.corailroyal.tourism.tn
Cofivet
http://www.cofivet.com.tn
Inesfood
http://www.inesfood.com.tn
Interieurs
http://www.interieurs.com.tn
Khrystaleng
http://www.khrystaleng.com.tn
KG Conseil
http://www.kgconseil.com.tn
Kingflex
http://www.kingflex.com.tn
Kanoun
http://www.gnet.tn/kanoun/
Olle-General Dairies
http://www.olle.com.tn
Société Promedia
http://www.leconomiste.com.tn
Hôtel Le palace
http://www.lepalace.com.tn
Pyra-Softwares
http://www.pyra.com.tn
LG electronics
http://www.lg.com.tn
LG electronics
http://www.lge.com.tn
LG electronics
http://www.lgetn.com.tn
Lufthansa
http://www.lufthansa.com.tn
Made
http://www.made.com.tn
Masmoudi
http://www.masmoudi.com.tn
Mahrajanet
http://www.mahrajanet.com.tn
Mahdia Beach
http://www.mahdiabeach.com.tn
Mafamec
http://www.mafamec.com.tn
Ma Maison
http://www.mamaison.com.tn
Medianet
http://www.medianet.com.tn
Mosaica
http://www.mosaica.com.tn
MP Consulting
http://www.mpconsulting.com.tn
MSMab
http://www.msmab.com.tn
Nadhour Tunisie
http://www.nadhour.org.tn

Comecab
http://www.comecab.com.tn

Comete Engineering
http://www.comete.com.tn

Confection Ras Jebal
http://www.crj.com.tn

Dahr
http://www.dahr.com.tn

Dar el Jeld
http://www.dareljeld.tourism.tn

Serveur Immobilier Diar
http://www.diar.com.tn

Festival de Douz
http://www.douz.festival.tourism.tn

Espace Pictural
http://www.espace-pictural.info.tn

Electro Diesel Tunisie
http://www.edt.com.tn

Etiquettes & Accessoires
http://www.e-a.com.tn

Institut Elamouri
http://www.elamouri.com.tn

Société Eltaief
http://www.eltaief.com.tn

Founoun
http://www.founoun-online.com.tn

FMCI
http://www.fmci.ens.tn

GMD Marine
http://www.gmdmarine.com.tn

Revendeur 3S
http://www.gnet.tn/~cefi/

Revendeur 3S
http://www.gnet.tn/~cisencomputer/

Revendeur 3S
http://www.gnet.tn/~computerhouse/

Revendeur 3S
http://www.gnet.tn/~msiconsultants/

Revendeur 3S
http://www.gnet.tn/~penta/

Revendeur 3S
http://www.gnet.tn/micro.island/

Habitation Moderne
http://www.habitationmoderne.com.tn

Hôtel Chems
http://www.hotelchems.com.tn

Société Beach Confection
http://www.gnet.tn/sbc/

ICARE Informatique
http://www.icare.com.tn

Société Ifriqia
http://www.ifriqiya.com.tn

Icom
http://www.icom.com.tn

Société IMPACT MEDIA
http://www.impact.com.tn

Najar Chaâbane-Renault
http://www.najar-chaabane-renault.com.tn

Repec
http://www.repec.com.tn

Service Assistance Matériels Informatiques
http://www.sami.com.tn

Secit immobilier
http://www.secitimmobilier.com.tn

Sinter-Souani
http://www.souani.com.tn

SDTS Hôtels
http://www.sdts.tourism.tn

Sfax Huile
http://www.sfaxhuile.com.tn

Société Générale d'Electroménager
http://www.sgewhirlpool.com.tn

Société industrielle d'amortisseurs
http://www.siaam.com.tn

Sepra
http://www.sepra.com.tn

S.I.L
http://www.sil.com.tn

Société Informatique Raja (SIR)
http://www.sir.com.tn

S.M.C
http://www.smic.com.tn

SOS Assistance / Allo taxi
http://www.sosassistance.com.tn

Starmedia
http://www.starmedia.com.tn

Société Tunisienne de Medecine Intrene
http://www.stmi.org.tn

Société Tunisienne des Industries Mécaniques
http://www.stim.ind.tn

Staedtler
http://www.staedtler.com.tn

Studi
http://www.studi.com.tn

Stusid
http://www.stusid.com.tn

Synergie informatique
http://www.synergie.com.tn

Synergie informatique
http://www.synergie.com.tn/~alliance

Synergie informatique
http://www.synergie.com.tn/~misfat

Tiffany
http://www.tiffany.com.tn

Tunisia poterie
http://www.tunisia-poterie.com.tn

Tunisie Cables
http://www.tunisie-cables.com.tn

Tunisiefactoring
http://www.tunisiefactoring.com.tn

Zone Franche de Zarzis
http://www.zfzarzis.com.tn

CONCEPTION ET IMPLEMENTATION D'UN MOTEUR DE RECHERCHE

Réalisé par

Sami Turki

Sabri Ben Amara

TS5- Télécommunications

Ce projet de fin d'étude a pour objectif la conception et l'implémentation d'un moteur de recherche pour 3S GlobalNet.

Dans ce rapport, on s'est intéressé à :

- ✓ Donner une idée sur les différentes architectures client-serveur ;
- ✓ Présenter les techniques de référencement ainsi que les outils de recherche sur Internet ;
- ✓ Expliquer le choix des techniques du développement les plus évolués tels que JSP et Oracle.

Ce moteur de recherche doit évoluer surtout côté rapidité et précision de la réponse pour pouvoir rivaliser au moins ces concurrents dans le monde arabe.

Mots clés : Moteur de recherche, Java, Référencement, Indexation, Internet, Serveur.

Année Universitaire : 2001- 2002

INTRODUCTION GENERALE

L'informatique a, par le passé, souvent été considérée comme une technologie. Mais, nous revendiquons pour elle le statut de discipline, statut qui lui est maintenant reconnu. C'est une science jeune. Elle a besoin de faire sa place. Elle l'a faite en certains lieux.

Quant aux télécommunications, c'est une discipline plus traditionnelle avec laquelle l'informatique a toujours entretenu des liens. La part de plus en plus importante du logiciel et le développement des réseaux ont, ces dernières années, conduit à une synergie de ces deux disciplines.

Pour nous, **l'essentiel est la modélisation**. Longue à développer, celle-ci ne s'arrête pas à des plates-formes et des stations de travail. Elle exige un savoir-faire et beaucoup de compétences, qui méritent d'être transmises. Elle exige un talent qui ne s'arrête pas aux aspects techniques, mais relève aussi de "l'esprit chercheur". C'est presque un art, comme l'architecture. Nous produisons et assemblons des codes exécutables pour aboutir à un objet final, le programme d'application. Ce travail d'architecte, d'urbaniste, a une finalité pratique, mais aussi une finalité esthétique qui va de pair avec l'efficacité du résultat.

Les systémiciens et les mathématiciens sont eux aussi des modélisateurs. Mais c'est l'informatique qui modélise les données et les traitements pour aboutir à la construction des programmes, pour partir de l'analyse et aller jusqu'à la programmation, en choisissant les bons outils et en sachant les mettre en oeuvre. Un véritable travail d'artisan.

Pour cela on a choisit ce projet qui s'inscrit dans un contexte purement informatique au prés d'une des plus grandes sociétés d'informatique en Tunisie. Dans ce rapport, on va vous présenter:

- 3S GlobalNet .
- Les outils de recherche .
- Le modèle de conception, de développement et les testes d'évaluation de notre moteur.

CHAPITRE A
PRESENTATION DU PROJET

I- Pourquoi Le Choix De Ce Projet :

Ce projet a pour but l'implémentation et la conception d'un moteur de recherche interne pour le compte de 3S GlobalNet. Ce projet est appelé à être optimisé pour devenir un moteur de recherche international.

Ce choix est basé sur de multiples raisons. En voici les plus importantes :

Ceci constitue une première pour tout le monde arabe. Et là on vous invite à consulter la liste des, soit disant, moteurs de recherche arabe présent actuellement sur la toile et qui ne sont en fait que soit des annuaires, soit des façades pour d'autre moteurs de recherche étranger .

3S GlobalNet est une grande société, leader dans le domaine de l'informatique, et le fait de travailler avec est sûrement une grande opportunité pour nous (avec tout ce que cela représente en terme de bon encadrement matériel et d'expérience) .

C'est un bon projet pour commencer la vie d'un programmeur qui rêve d'accéder au SUP'COM.

Tous ces raisons nous on poussé à revoir notre approche et ainsi à songer à subdiviser le projet en trois étapes à la fois distinctes et complémentaires :

Développer un moteur de recherche interne propre au site de 3S GlobalNet (qui est notre projet).

Développer un moteur de recherche pour les sites tunisiens .

Optimiser le travail pour avoir un moteur de recherche international.

Comment alors aborder ce sujet ?

Quels sont les outils de recherche sur Internet ?

A quel stade sont arrivés les concepteurs et développeurs dans ce domaine?

Quels outils utiliser ?

Une infinité de questions qui peuvent certainement nous guider à mieux comprendre le sujet et bien évaluer la complexité de la tâche.

II) Présentation du 3S GlobalNet :

1- Historique :

3S GlobalNet est une société de grande renommée qui a vu le jour dans un pays jusque là dépossédé de la « culture du Net » et loin de prétendre détenir une infra structure apte à de telle activités. Mais malgré son début dans cet environnement encore hostile, elle peut se valoir aujourd'hui d'avoir une réputation en or et d'être classée parmi les premiers.

Et ceci est un petit résumé de sa glorieuse traversé :

- 1988: Fondation de Standard Sharing Software ou 3S, par une équipe de docteurs en informatique .
- 1989: Création de la bibliothèque 3S constituant la base des logiciels assurant la portabilité de nos applications sur toute plateforme .
- 1990: Développement d'AVICENNE logiciel de gestion des pharmacies portable sur tout système. 3S devient le premier distributeur en Tunisie des produits SCO et est encore à ce jour l'unique SCO UNIX Partner, partenaire privilégié de SCO en Tunisie .
- 1991: Développement d'applications de gestion d'entreprises portables sur tout système AVERROES pour la gestion commerciale et MAESTRO pour la gestion comptable .
- 1992: Lancement d'un projet majeur : AQARIX, le premier progiciel pour la gestion complète d'hôtellerie. Développement d'une application de gestion d'entreprises portable sur tout système pour la gestion du personnel, de la paie et de la présence TASK .

- 1993: 3S ouvre un axe Réseaux et Communications et devient rapidement CISCO Systems Partner en Tunisie .
- 1994: Développement d'une application de gestion d'entreprises portable sur tout système SINDBAD pour la gestion des transporteurs, des transitaires, des dépositaires. 3S s'allie à IBM pour promouvoir la gamme RISC/6000 sous UNIX .
- 1995: Mise en place d'un service cabling spécialisé en câblage informatique et en équipements d'interconnexion .
- 1996: Développement d'une application de gestion d'entreprises portable sur tout système SAFARI pour la gestion des agences de voyage. 3S devient le partenaire privilégié de PARADYNE .
- 1997: 3S sélectionné par le Ministère des Communications devient Provider Internet et fournit, depuis le 15 Septembre 1997, les abonnements et services Internet en Tunisie. 3S ouvre une agence Internet, change de nom et de logo et devient 3S GlobalNet .
- 1998: 3S GlobalNet développe des sites Web et forme les abonnés à tous les services d'Internet .
- 1999: 3S travaille à la mise en place d'un centre de support et de formations agréé CISCO Systems, SCO et IBM .

2- Production :

3S GlobalNet est le premier opérateur en Tunisie en matière de développement de progiciels de gestion multi-utilisateurs, multi-postes et multi-environnements. Dès leur conception, tous les progiciels sont certifiés au passage à l'an 2000.

Tous les progiciels sont développés par une équipe d'ingénieurs assistés par des spécialistes de la gestion. Ils sont écrits en langage C afin de leur assurer une vitesse d'exécution et une portabilité à toute épreuve. Ainsi tous nos progiciels fonctionnent indifféremment sous DOS, UNIX, NT ou NOVELL.

Il est à noter que tous les progiciels 3S GlobalNet sont mis à jour tous les six mois pour être en adéquation avec les nouvelles méthodes de gestion, la nouvelle législation ou pour répondre aux demandes de nos clients. Le développement de nouveaux modules enrichi nos produits et assure la continuité d'exploitation par nos clients. En fait acquérir aujourd'hui un système 3S Global Net, ce n'est pas simplement s'équiper, c'est investir pour le futur.

3S GlobalNet dispose aujourd'hui d'une gamme complète de progiciels couvrant plusieurs domaines :

Gestion des entreprises (AVERROES, TASK II, MAESTRO).

Gestion d'hôtels (AQARIX).

Gestion de pharmacie (AVICENNE).

Gestion d'agences de voyage (SAFARI, TASK II, MAESTRO).

Gestion de transitaires (SINDBAD, TASK II, MAESTRO).

En outre 3S GlobalNet dispose de progiciels de gestion d'entreprises indépendamment de leurs secteurs d'activité :

- MAESTRO (Finance) est composé de quatre modules :
- MAESTRO Base (comptabilité générale et auxiliaire) .
- MAESTRO Immo (Gestion des immobilisations et des amortissements) .
    - MAESTRO Plus (Contrôle et justification des comptes) .
    - MAESTRO Budget (Budget).
- AVERROES (Stock et gestion commerciale) qui est composé de deux modules :
    - AVERROES Base .
    - AVERROES Prodix (Gestion de la production).
- TASK II (Gestion de la paie, du personnel et de la présence) est composé de trois modules :
    - Paie .
    - Personnel .
    - Présence.

Tous les progiciels 3S GlobalNet sont caractérisés par :

- Un très haut niveau de paramétrage permettant de s'adapter aisément à toutes les spécificités évolutives de votre organisation sans contrainte de dépendance de la taille de la société ou nécessité d'intervention de votre part.
- Une extrême simplicité d'utilisation et ce grâce aux techniques de programmation et d'interfaçage intelligent : Menus arborescents, multi-fenêtrage, accès multicritère, saisie corrélée, choix par défaut, etc…
- Un système de sécurité discriminatoire fiable et puissant. La définition des accès et des droits des utilisateurs par action engendre de manière dynamique le changement des menus.

3- Les Partenaires De 3S GlobalNet :

De part ses compétences, 3S GlobalNet est depuis 1993 partenaire avec CISCO en Tunisie. CISCO Systems Inc. est le leader mondial des réseaux pour Internet et le géant de l'interconnexion avec plus de 50% du marché mondial toutes technologies confondues. La gamme des produits CISCO inclue les routeurs, les LAN, les switches ATM, les serveurs d'accès et les softs d'administration réseaux. Ces produits sont intégrés grâce au

logiciel IOS (Internetworking Operating System) de CISCO qui permet de servir plusieurs sites distants LAN, WAN et réseaux IBM.

Aujourd'hui, l'environnement réseau étant à la fois complexe et dynamique, les acteurs du marché réseaux et communications doivent rester compétitifs, professionnels et performants.

En tant que leader mondial en technologies réseaux, CISCO Systems sélectionne avec soin dans le monde entier des partenaires privilégiés capables de distribuer ses solutions réseaux, de conseiller les entreprises et de fournir un support technique de qualité. CISCO certifie ses partenaires et si nécessaire leur apporte une aide technique.

CISCO Systems et 3S GlobalNet ont pour objectif de permettre aux entreprises d'utiliser leurs ressources informatiques de manière plus efficace en les interconnectant. Dès 1986, CISCO lançait son premier routeur multi-protocoles, un produit permettant à tous les types d'ordinateurs du marché de partager leurs ressources sur le même réseau. Dès 1993, CISCO met en place le premier réseau intégrant plus de 1000 routeurs. Aujourd'hui, la technologie CISCO permet à des ordinateurs hétérogènes, stations de travail, mini-ordinateurs et systèmes centraux de communiquer à travers des réseaux distants ou sur des réseaux locaux à haut débit. L'IOS CISCO (Internetwork Operating System) a été créé, non seulement pour les produits CISCO, mais également pour les équipements de plus de 20 constructeurs. L'IOS CISCO est devenu le standard de facto de l'industrie de l'interconnexion.

Pour toujours mieux répondre à vos besoins, les ingénieurs de 3S GlobalNet suivent le programme CISCO Certified Internetwork Expert (CCIE) et sont en cours de certification expert.

C'est l'un des plus haut niveau de certification mis en place par CISCO Systems. Le statut CCIE prouve, si besoin est, un haut niveau technique de nos ingénieurs en matière d'interconnexions, de communications, de réseaux et d'Internet.

4- Palmarès :

3S partenaire CISCO systems, c'est 90% du marché tunisien de l'interconnexion de réseaux. Prés de 40 réseaux locaux sont installés par nos soins chaque année. Ainsi, 3S a installé à ce jour prés de 120 réseaux locaux, totalisant plus de 850 postes de travail.

A titre d'exemple, 3S a réalisé, entre autres, les réseaux suivants :

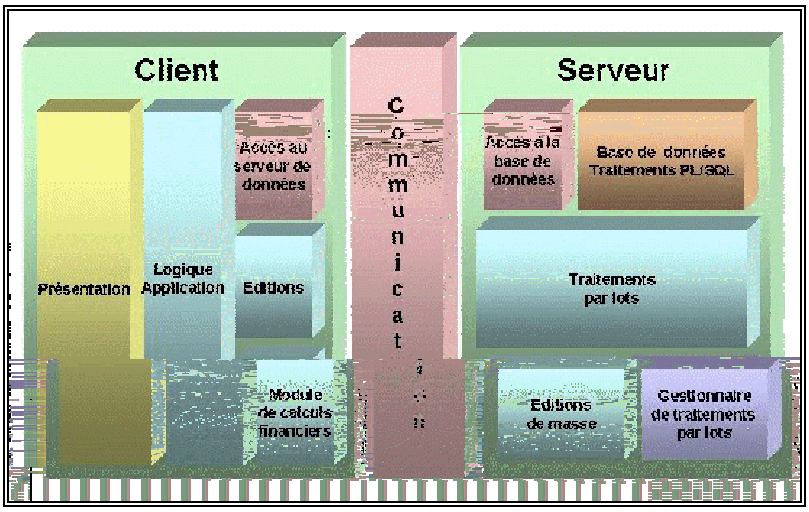
- Présidence de la République .
- Premier Ministère .
- Tunisie Télécom .
- Arab Tunisian Bank .

- Réseau National .
- Ministère de l'Economie .
- 36 réseaux locaux interconnectés à travers le réseau national X25 Tunipac desservant plus de 120 stations .
- Secrétariat d'Etat à la Recherche Scientifique et à la Technologie.
- Réseau National de la Recherche et de la Technologie : 12 noeuds basés sur des routeurs multi-protocoles CISCO .
- Ministère de la Santé Publique .
- Plus de 30 routeurs et plus de 1000 connexions TCP/IP sur des serveurs IBM, Data General, SUN, HP.
- Agence Tunisienne de l'Emploi .
- Subdivisions câblées et reliées à travers le réseau national totalisant 400 connexions TCP/IP sur des serveurs IBM via plus de 50 routeurs CISCO .
- Agence Tunisienne de l'Internet .
- L'ensemble du Backbone Internet en Tunisie a été équipé par nos soins en produits CISCO .
- Office du Commerce Tunisien .
- 100 prises, 9 km de câbles, des rocades fibre optique, des armoires, des routeurs sur deux bâtiments, en moins de 15 jours.

CHAPITRE B

NOTION CLIENT/SERVEUR

I) Introduction :



Dans un contexte décentralisé, une entité logiciel (au sens large : application, système ou processus) commande des traitements à une autre. La première est appelée client, et la seconde serveur. ces deux notions rejoignent celles de maître/esclave que l'on retrouve souvent dans la littérature informatique.

Dans le cadre d'une base de données, les traitements commandés par le client sont exprimés, sous forme d'ordres d'accès aux données, au moyen d'un langage puissant : SQL. Ensuite, le serveur chargé de gérer une base de données, exécute ces ordres et envoie les résultats des requêtes au client. Évidemment, un serveur doit être capable de traiter simultanément des ordres en provenance de différents clients.

Les moyens de communication entre les machines logiques affectées à la même machine physique sont divers : les messages, les "pipes" (au sens UNIX), les zones de mémoire communes . alors que les moyens de communication des machines physiques sont: les lignes de communication et d'une façon générale, les réseaux informatiques. Un serveur est donc une machine logique qui :

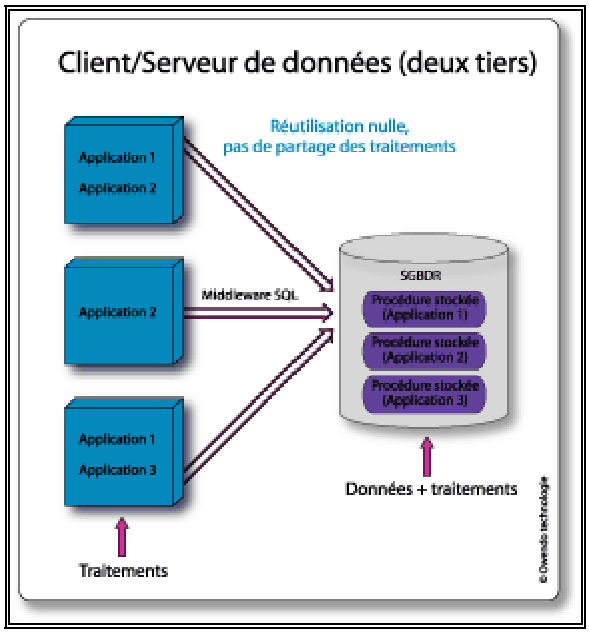
- Possède une base de données .
- Gère, grâce à un noyau d'exécution, une base de données .
- Est capable de gérer des accès concurrents.

Ceci implique que la machine physique sur laquelle le serveur s'exécute, possède un système d'exploitation mettant en œuvre un mécanisme de multitâches.

II- Les différentes architectures :

1) Présentation De L'architecture à 2 niveaux :

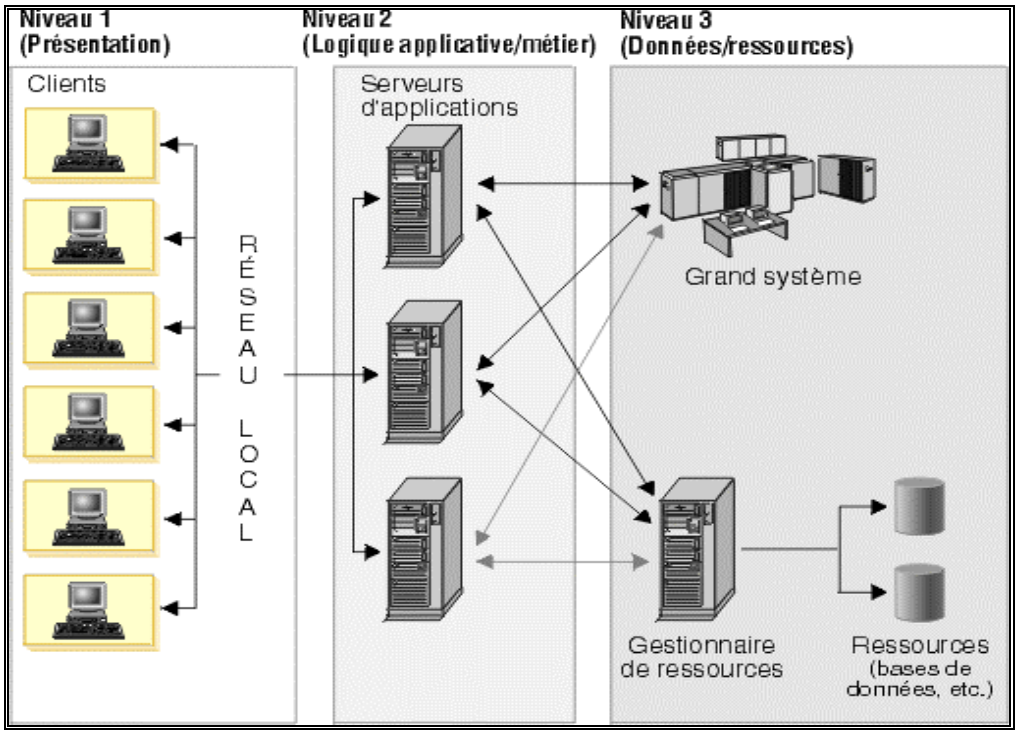
L'architecture à deux niveaux (aussi appelée architecture 2-tiers, tier signifiant étage en anglais) caractérise les systèmes clients/serveurs dans lesquels le client demande une ressource et le serveur la lui fournit directement. Cela signifie que le serveur ne fait pas appel à une autre application afin de fournir le service.



2) Présentation de l'architecture à 3 niveaux :

Dans l'architecture à 3 niveaux (appelées architecture 3-tiers), il existe un niveau intermédiaire, c'est-à-dire que l'on a généralement une architecture partagée entre:

1. Le client: le demandeur de ressources .
2. Le serveur d'application (appelé aussi middleware): le serveur chargé de fournir la ressource mais faisant appel à un autre serveur .
3. Le serveur secondaire (généralement un serveur de base de données), fournissant un service au premier serveur.



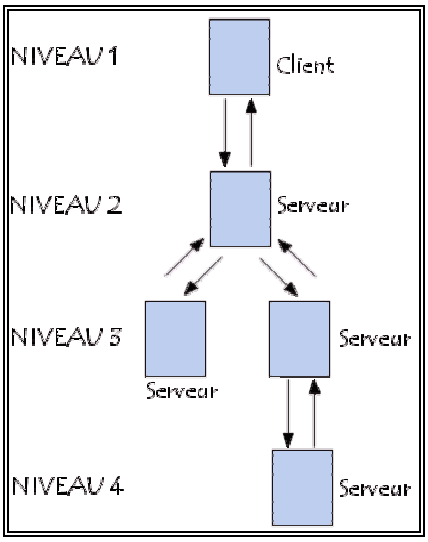
L'architecture à deux niveaux est donc une architecture client/serveur dans laquelle le serveur est polyvalent, c'est-à-dire qu'il est capable de fournir directement l'ensemble des

ressources demandées par le client. Dans l'architecture à trois niveaux par contre, les applications au niveau serveur sont délocalisées, c'est-à-dire que chaque serveur est spécialisé dans une tâche (serveur web/serveur de base de données par exemple). Ainsi, l'architecture à trois niveaux permet:

- une plus grande flexibilité/souplesse .
- une plus grande sécurité (la sécurité peut être définie pour chaque service) . de meilleures performances (les tâches sont partagées).

3) L'architecture multi-niveaux :

Dans l'architecture à 3 niveaux, chaque serveur (niveaux 1 et 2) effectue une tâche (un service) spécialisée. Ainsi, un serveur peut utiliser les services d'un ou plusieurs autres serveurs afin de fournir son propre service. Par conséquence, l'architecture à trois niveaux est potentiellement une architecture à N niveaux...



CHAPITRE C

REFERENCEMENT ET OUTLIS DE RECHERCHE

I- Introduction :

L'objectif de ces outils est de mettre à la disposition du public une constellation de sites divers et variés traitant aussi bien de l'actualité que des sciences et des technologies ou encore de l'Art et de la culture. Dans ces moteurs de recherche ou annuaires, se côtoient des sites de tout horizon. Ils peuvent être aussi bien personnelles que professionnelles, de grands consortiums industriels que de petites et moyennes entreprises.

L'offre étant pléthorique et la concurrence de plus en plus forte, il devient considérablement difficile d'obtenir le référencement dans les outils de recherche. Pour parvenir à se référencer auprès de ces derniers, une bonne connaissance de leur fonctionnement et de leurs méthodes d'indexation est indispensable.

D'abord, la plupart des internautes interrogent des moteurs de recherche afin de trouver un site correspondant à leurs requêtes, et partant, pour obtenir des réponses à leurs questions. Ainsi, 80% des sites Web sont visités suite aux réponses données par les outils de recherche. Par ailleurs, 67% des internautes ne saisissent que deux mots clés pour effectuer une recherche. Ensuite, pour 75% du public, seulement les dix premières réponses (soit la première page) sont consultées et 90% ne vont pas au-delà de la deuxième page.
D'où, l'importance de bien maîtriser les mots clés par lesquels le site sera découvert parmi une multitude. Choisir judicieusement des mots clés et un titre contribuera indéniablement à garantir un succès dans le travail de référencement.

D'autre part, il semble qu'actuellement les outils de recherche ne parviennent pas à absorber la formidable croissance des sites sur le Web. L'ensemble des moteurs ne référence que la moitié des documents estimés sur Internet, représentant plus de 17 millions de sites (Source Netcraf). De plus, la plupart de ces outils n'indexeraient au total que 42% des 2,2 milliards de pages présentes sur Internet (Source Cyveillance). All The Web serait le premier moteur de recherche puisqu'il référencerait à lui seul environ 35% des pages Web, suivi par Altavista avec moins de 25% et Northern Light avec près de 20%.

Au niveau de la popularité, l'annuaire Yahoo détient la palme avec un taux de fréquentation de 39%. Le moteur de recherche Altavista prend la seconde position avec un peu plus de 18%, suivi par Excite et Infoseek avec 14%, les autres se contentant du reste. En conséquence, il est parfaitement inutile de s'escrimer sur des moteurs de recherche à petite fréquentation et plutôt se focaliser sur les outils majeurs d'Internet, ceux dont la base de données et la fréquentation sont importantes. Puisque 65% des usagers de l'Internet passent par eux pour trouver une information, les moteurs de recherche et les annuaires demeurent le principal vecteur de visibilité d'un site sur le World Wide Web.

III- Référencement :

Aujourd'hui soumis aux aléas d'une évolution rapide d'Internet et de tous ses outils, le référencement fait l'objet d'une véritable science sans laquelle sa réalisation devient extrêmement délicate et laborieuse . Surtout si on tend à atteindre une place parmi les dix premières réponses des moteurs de recherche. Néanmoins, de nombreux prestataires proposent leurs services pour le référencement des sites Web. Vous libérant ainsi d'un travail récurrent et pénible.

Après le développement d'un site, si l'on désire obtenir une audience confortable et tirer les bénéfices de notre travail en terme de rentabilité, de notoriété et d'image, alors on n'a pas d'autre alternative que de lui faire un coup de publicité.

C'est pourquoi, la solution réside dans le référencement auprès des moteurs de recherches tel que Voila et Altavista ou des annuaires comme Yahoo et Nomade afin d'acquérir une certaine popularité et une visibilité maximale auprès du public.

1) Historique :

Les techniques de référencement évoluent sans cesse. Voici les grandes étapes que nous avons identifiées.

- 1994: apparition des premières listes de liens et inscriptions grâce à des contrats informels .
- 1995: référencement des Home Pages dans les annuaires comme Yahoo !
- 1996:début du référencement manuel de plusieurs dizaines de pages du site web dans les moteurs de recherche comme Altavista et Excite .
- 1997: contrôle grossier des résultats et mis en place de tableaux de bords en se basant sur les classements sur quelques mots-clés .
- 1998: analyse fine du référencement et contrôle de l'évolution dans le temps, référencement dans les agents intelligents, début des campagnes publicitaires à base de référencement (et non de bandeaux publicitaires) .
- 1999: apparition de stratégies complexes de référencement (serveurs parallèles, création de liens artificiels pour rendre un site plus populaire aux yeux des moteurs qui mesurent la popularité des sites en fonction des liens qui pointent vers eux) .
- 2000: veille concurrentielle permanente et généralisation du référencement dans les agents intelligents.

2) Les techniques du référencement :

Une opération de référencement dite de qualité respecte généralement les étapes suivantes avec quelques variantes selon les prestataires.

- Le travail sur la structure du site : Pour des raisons budgétaires et de délais, le travail d'optimisation ne peut se faire que sur un nombre de pages limité. Un moteur peut très bien indexer toutes les pages non dynamiques d'un site mais l'effort ne portera que sur quelques unes à placer en bonne position. L'emplacement de ces pages devra être mûrement réfléchi car si la politique de référencement est réussie, ces pages deviendront un point d'entrée pour une part non négligeable des visiteurs. Pour faciliter le référencement sur certains moteurs, certaines techniques, tel que l'utilisation des frames, peuvent éventuellement être remise en cause.

- L'optimisation du contenu des pages : La plupart des sites obtiennent une rémunération, de la part de leurs sponsors, sur la base du taux de fréquentation. Plus l'audience est élevée, plus les gains deviennent importants. **Par esprit mercantile et d'arrivisme, des webmasters se laissent tenter par certains artifices illicites**.

3) La création de pages satellites :

Les pages satellites sont des reproductions généralement fidèles de la page d'accueil qui sont créées spécialement pour optimiser le référencement. En effet, une page ne peut être optimisée que pour nombre limité de mots clés. Souvent le visiteur va pénétrer sur le site par une page satellite sans même s'en rendre compte. Les pages satellites peuvent être dédiées à un moteur de recherche spécifique ou à un thème de recherche ou secteur d'activité précis. Chaque moteur ayant ses propres modes de fonctionnement, il peut être intéressant de créer une page sur-mesure adaptée à ses exigences techniques ou à ses méthodes de calcul de pertinence. Les sites proposant des services ou produits variés peuvent rendre nécessaire la création de pages satellites spécialisées sur chaque activité.

4) Ne pas essayer de tromper les robots :

Si divers trucs existent pour aider les robots à trouver votre site attractif et pour promouvoir divers mots clé, il n'est cependant pas correct de tricher. En effet divers webmasters choisissent de faire confiance à ces méthodes:

Utilisation de fichiers uniquement pour les robots avec un contenu uniquement composé de répétition de mots clé ou de mots très utilisés sur Internet (éventuellement même par un public assez particulier...) mais sans rapport avec le contenu de votre site .

Répétition de parties du code HTML (comme par exemple du titre) ce qui ne change pas l'affichage mais trompe certains robots .

Répétition de mots clé dans le contenu, mais avec une couleur ou une taille prévue pour rendre le texte invisible au lecteur et donc seulement pour les robots .

Utilisation de diverses pages avec le même contenu, pour augmenter le nombre d'occurrences des mots sur le site .

Référencement de diverses pages d'entrée juste avec un contenu de base (bouton pour entrer sur le site) ou chargement automatique de la page principale. Pour augmenter le nombre de présences dans les moteurs.

Il faut savoir que de plus en plus de robots utilisent maintenant des technologies étudiées pour contrer les tricheries et les annuaires effectuent aussi manuellement des visites de contrôle. Les sanctions prévues deviennent de plus en plus sévères (radiation du site) et liste noire ou encore pire une action judiciaire peut vous être intentée dans le cas de concurrence déloyale... En effet, par l'utilisation de leurres vous trompez les visiteurs sur la qualité et le contenu de votre site, et les moteurs de recherche qui essayent d'attirer les utilisateurs par un contenu de qualité se doivent de parer à cette fraude.

5) Aidez les robots :

Puisque l'objectif demeure dans l'absolu, celui d'améliorer l'audience d'un site, vous n'avez certainement aucun intérêt à vous retrouver embourber dans une sombre histoire de fraude, mais plutôt à tenter par tous les moyens légaux et moraux d'atteindre un taux de fréquentation escompté. L'honnêteté, le respect des acteurs du Web, la connaissance des techniques de référencement, la richesse et la qualité de votre site vous garantissent un succès et une notoriété sur Internet.

6) Les Mots Clés :

Les mots clés sont particulièrement sensibles puisque :

C'est par leurs intermédiaires que le site sera trouvé et visité .

Ils déterminent le classement dans les moteurs de recherche .

Ils contribuent à valider l'inscription dans les outils de recherche.

A cause de ces trois raisons majeures, on est tenu de déterminer scrupuleusement chacun des mots clés des pages. Ils sont présents pour valoriser le site et non pour le vouer aux gémonies.

Le but principal de ces éléments est de permettre une visibilité maximale du site. Idéalement, chacune des requêtes de recherche effectuées sur un des mots clés doit aboutir sur le site ou du moins à afficher le site dans les dix premières réponses.

a) Titre :

Le titre est une des pièces maîtresses de la promotion d'un site, alors il ne doit pas être négligé et surtout il faut garder à l'esprit que si un usager potentiel ne se rend pas sur le site lors de la première approche, il y a de fortes probabilités pour qu'il soit définitivement perdu même si le contenu semble être meilleur que celui du concurrent bénéficiaire.

Lorsqu'un utilisateur effectue une recherche par mots clés, le moteur de recherche applique la méthode du scoring pour trouver et classer des sites dans une liste contenant

plusieurs centaines de réponses. Le scoring évalue un prétendant aux nombres de répétition dans une page du mot clé spécifié. Plus ce dernier sera localisé à un niveau supérieur du document, plus il aura de poids et inversement, plus il sera bas, moins il sera influent.

b) Les balises META :

Les balises <*META...*> fournissent aux moteurs de recherche des indications primordiales et d'autres informations sur les pages d'un site.

En premier lieu, elles accueillent la description et les mots clés. Ces derniers sont des éléments capitaux lors du référencement puisqu'ils seront à l'origine de la visibilité de votre site Web .

En second lieu, elles déterminent l'activité des robots sur le site en lui ordonnant ou lui interdisant d'indexer des documents et de suivre les pages liés ou encore en lui spécifiant la fréquence de ses visites .

Ensuite, elles renseignent sur les caractéristiques du site en matière de date de création et d'expiration, de langue usitée, de date de modification, etc…

Enfin, elles donnent des renseignements sur les responsables du site en précisant les noms des auteurs, du webmaster, les adresses e-mail, la localisation, etc…

Si le document en question ne contient pas de balises <*META...*>, alors le moteur de recherche se contentera des premières lignes du texte en guise de description et les mots clés seront déterminés par le nombre de répétitions dans le document. Ainsi, il y aura risque de perte en visibilité et peut être en crédibilité.

La plupart des robots des moteurs de recherche analyseront en priorité cet entête pour en extraire les informations nécessaires au référencement. Subséquemment, ces informations seront comparées au reste de la page pour valider la qualification du site et partant d'entamer son indexation dans la base de données du moteur.

Voici quelques unes des nombreuses balises <*META...*> :

```
<HEAD>
<META NAME="description" CONTENT="fournisseur d'Internet privé">
<META NAME="keywords" CONTENT="Internet, tunisie, informatique, globalnet">
<META NAME="Robots" CONTENT="index, nofollow">
<META NAME="revisit" CONTENT="15 days">
<META NAME="createDate" CONTENT="15/06/1998">
<META NAME="Language" CONTENT="FR">
<META NAME="LastUpdated" CONTENT="26/11/2001">
<META NAME="revision" CONTENT="15/01/2000">
<META NAME="contact" CONTENT="faker@gnet.tn">
```

<META NAME="contactPhoneNumber" CONTENT="0021671700100">
<META NAME="contactNetworkAddress" CONTENT="193.95.73.30">
<META NAME="Generator" CONTENT="webexpert 2000">
<META NAME="Owner" CONTENT="Chakroun">
<META HTTP-EQUIV="Content-language" CONTENT="fr">
<META HTTP-EQUIV="Content-Type" CONTENT="text/html. charset=ISO-2022-JP">
<META HTTP-EQUIV="Title" CONTENT="3S GlobalNet">
<META HTTP-EQUIV="refresh" CONTENT="60. index.html">
<META NAME="DC.Coverage" CONTENT="Tunisie, Tunis">
<META NAME="DC.Publisher" CONTENT="Raouf">
<META NAME="DC.Title" CONTENT="3S GlobalNet">
<META NAME="htdig-email-subject" CONTENT="facker@gnet.tn">
<META NAME="htdig-noindex">
</HEAD>

Les balises <*META...*> ne sont pas prise en compte par tous les moteurs de recherche. La seule parade est alors de spécifier les mots clés dans le corps du document par l'intermédiaire du texte et des commentaires voire des attributs *ALT* de la balise d'insertion d'image <*IMG...*>.

III- Les Annuaires :

Les annuaires ont un fonctionnement très simple puisqu'ils proposent des sites classés par catégorie contrairement aux moteurs qui fournissent une liste de sites par rapport à une requête par mot clé d'un internaute.

La page d'accueil de ces outils de recherche affiche les rubriques principales sous forme de thèmes génériques telles que Actualité et Média, Enseignement, Santé, Education, etc…

Chaque site indexé est accessible quelque part au sein des catégories principales dont chacune comporte plusieurs sous-répertoires. Cette structure permet à partir d'une expression générale de descendre peu à peu vers une rubrique précise où est localisée l'information recherchée. L'accès à l'information devient plus aisé et intuitif mais nécessite un certain nombre de clics de souris et une concentration importante afin d'éviter un fourvoiement lors d'une recherche.

La proposition d'un site dans un annuaire demande au préalable un examen approfondi de ses catégories et de ses sous-catégories afin de déterminer le plus précisément possible la rubrique la plus appropriée.

Egalement appelés les répertoires, les annuaires référencent manuellement les sites qui leurs sont proposés. L'ajout d'un site dans un annuaire nécessite de la part du webmaster de

compléter un questionnaire plus ou moins long comprenant notamment un titre, une description, des mots clés, etc...

Contrairement aux moteurs de recherche, **ce n'est pas un robot qui viendra évaluer le site mais un humain**. L'inscription dépendra en grande partie de la justesse des renseignements fournis et surtout de la qualité du site.

1) Principe de fonctionnement:

Contrairement à ce que l'on pourrait croire, les annuaires ne référencent pas des pages, mais des sites. Ces sites sont généralement classés par catégories. Lorsque l'on consulte un annuaire, on a donc deux possibilités:

- Utiliser une recherche par mots clés, comme pour les moteurs de recherche .
- Parcourir des menus, des catégories et des sous-catégories, jusqu'à ce que l'on trouve le site qui nous intéresse.

Ce second mode de consultation est moins efficace lorsque l'on recherche une information précise. Mais il permet d'avoir des listes de sites classés par thèmes. Le nombre de sites présents dans les annuaires est nettement moindre que dans les moteurs de recherche. Les annuaires ne viendront pas parcourir vos pages pour voir s'il y est des sites qu'ils ne connaissent pas encore.

Les annuaires sont moins riches que les moteurs de recherche. Ils ne gardent qu'une fiche par site avec un nombre limité de mots clés et une description dépassant rarement vingt-cinq mots. Les mots clés qui ne figurent pas sur la page d'accueil même ne seront pas pris en compte.

Le classement en catégories nécessite une intervention humaine. Le site sera visité avant inscription dans l'annuaire. Tous les annuaires se réservent le droit d'accepter ou de refuser les sites qui leur sont soumis. Si le site ne réponds pas aux critères moraux des éditeurs d'un annuaire, il sera purement et simplement refusé.

L'indexation humaine nécessaire pour référencer un site fait que le délai de référencement est en général plus long pour les annuaires que pour les moteurs de recherche. Le délai de référencement est le temps qui s'écoule entre la soumission de votre site et son inscription effective dans l'annuaire.

Annuaires et moteurs de recherche sont complémentaires. Lorsqu'un annuaire ne trouve pas de sites correspondant aux mots clés indiqués par l'utilisateur, il passe la main à un moteur de recherche.

Par exemple, si Yahoo!, un annuaire, ne peut répondre à votre demande, il la transmettra automatiquement à Inktomi, un moteur de recherche.

IV- Portail :

Il est important de ne pas confondre portails ou outils de recherche. Yahoo ou Altavista sont des portails. Dmoz ou Google sont des outils de recherche.

Un portail est un espace de médiation virtuelle bénéficiant d'une très forte part de marché sur l'ensemble des internautes ou sur une cible très précise et s'appuyant sur l'agrégation de contenus, de services, d'outils de groupware et d'outils de recherche permettant aux utilisateurs de localiser rapidement les ressources en ligne dont ils ont besoins.

Pour comprendre le concept de portail, il faut connaître l'historique des outils de recherche sur Internet. Pendant longtemps, les seuls revenus des annuaires de recherche comme Yahoo ou Excite provenaient de l'affichage de bandeaux publicitaires. Ils enregistraient, en effet, un trafic élevé car chaque internaute se connectait plusieurs fois par mois sur ces outils. Malheureusement, les visiteurs ne consultaient que deux ou trois pages en moyenne à chaque visite. Ceci ne permettait d'afficher que deux ou trois bandeaux publicitaires, ce qui correspondait à des revenus publicitaires de 50 à 60 centimes. Les outils de recherche tentent donc de multiplier les services pour inciter les visiteurs à rester plus longtemps et à visualiser davantage de pages pour générer plus de revenus publicitaires. La diffusion des dépêches d'agences comme l'AFP ou Reuters est l'un des premiers services offerts, bientôt suivi par un service d'informations boursières, d'informations sportives… D'autres services de recherche viennent s'intégrer au service de base : Yahoo, annuaire de sites web, se dote d'un moteur de recherche en texte intégral (par un premier partenariat avec Altavista), d'un service de pages jaunes et racheté un annuaire d'adresse e-mail offrant même des e-mails gratuits… Enfin, la manne du commerce électronique permet aux sites portails de compléter leurs revenus : ils inaugurent tous les uns après les autres des espaces shopping.

Les revenus sont générés principalement par :

La publicité (sous forme de bandeaux ou de sponsoring) .

Les commissions suite aux ventes générées (programme d'affiliation) .

L'inscription dans les annuaires (qui tend à devenir payante).

La vente de services (Yahoo.com propose ainsi la création de boutiques en ligne).

Tous les outils de recherche ont convergé vers le modèle du portail si bien qu'il est de plus en plus difficile pour le néophyte de distinguer les sites qui étaient à l'origine des annuaires web (comme Yahoo) de ceux qui étaient des moteurs de recherche en texte intégral (comme Altavista, Excite, Webcrawler, Hotbot ou Lycos).

Les véritables sites portails présentent les caractéristiques suivantes :

Ils enregistrent un très fort trafic (relativement aux populations qu'ils visent) .

Ils fidélisent leurs visiteurs .

Ils ont quasiment toujours un service de recherche très performant .

Ils ont une taille critique qui leur permet de se positionner en tant que fédérateurs dans leur spécialité .

Ils sont bien des portes d'entrée qui permettent d'évoluer plus facilement sur Internet .

Leur modèle économique est complexe et ne repose pas uniquement sur la publicité.

Le concept du portail a été largement repris et décliné comme cela se produit souvent lorsqu'un modèle économique émerge sur Internet, si bien que de nombreux sites prennent désormais l'appellation de sites portails sans en présenter toutes les caractéristiques de bases. Très souvent, des sites agrègent un certain nombre de services avant même d'avoir un trafic important, ce qui fait perdre toute sa pertinence à la démarche.

V- Moteur De Recherche :

Les moteurs de recherche parcourent le Web (toile) à l'aide de spiders (araignées) ou robots et indexent automatiquement les pages des sites qui leurs sont soumises. En suivant les liens, les pages sont indexées les unes après les autres. Si bien que toutes les pages d'un site pourraient être finalement enregistrées dans la base de données du moteur. De même, des sites externes rattachés à vos documents par des liens peuvent faire également l'objet d'une indexation.

En fait, à partir d'une centaine de pages référencée, un moteur de recherche a la capacité de parcourir plusieurs centaines de milliers, voire même des millions, de documents rattachés par des liens hypertextes ou hyperimages. Les pages reliées sont toutes susceptibles de recevoir la visite d'un indexeur. Cela pouvant provoquer des problèmes, notamment au niveau de la liberté, il existe un moyen de contrer ce phénomène par le fichier *robots.txt* contenant des commandes spécifiques (Disallow:/répertoire/) qui indiquent aux robots de ne pas visiter certains répertoires ou certains fichiers de votre site. En effet, en plaçant ce fichier texte à la racine de votre site, vous avez la possibilité de contrôler l'accès à vos documents.

Par ailleurs, l'indexation automatique échoue systématiquement sur certaines pages dont la construction est particulière. En effet, des pages structurées autour des cadres (frames) ou d'une programmation spécifique à l'image de l'ASP (Active Server Pages), du JavaScript, du Flash, etc., constituent une pierre d'achoppement pour le référencement.

Enfin, les moteurs de recherche demeurent extrêmement susceptibles en matière de spamindexing. Les tentatives fallacieuses dans le but d'obtenir un meilleur classement sont sévèrement réprimées. Parfois, la frontière entre l'optimisation et le spamindexing semble bien ténue.

1) La conception de nouveaux moteurs de recherche :

En juillet 1998, le moteur AltaVista a récupéré 170 millions de pages (sur un total estimé à 350 à cette période), dont 125 millions ont été indexées, ce qui représente environ 800 gigaoctets de texte brut (un gigaoctet, ou milliards d'octets, correspond au contenu en texte de l'intégrale d'une grosse encyclopédie), soit, après traitement, à un index de près de 250 gigaoctets installé sur plus d'une vingtaine de machines de très haut de gamme (dix processeurs et dix gigaoctets de mémoire vive par machine). Ces machines sont interrogées 37 millions de fois par jour, en semaine, avec un temps de réponse moyen de 0,6 seconde. **Ces quelques chiffres aident à apprécier la difficulté de la conception et de la mise en œuvre d'un moteur de recherche généraliste.**

Afin d'améliorer la pertinence des documents retournés, les moteurs de recherche mettent en œuvre plusieurs solutions. L'une concerne les requêtes elles-mêmes, par exemple la correction orthographique ou la détection automatique des phrases. Ainsi, un moteur fournira de meilleures réponses si la requête est l'expression «effet de serre» plutôt que les trois mots «effet», «de» et «serre», car l'expression est plus discriminante que les trois mots pris séparément. Ce fait est bien connu des linguistes : le sens des textes est en grande partie contenu dans les groupes nominaux. Depuis l'été 1998, AltaVista reconnaît automatiquement les phrases, ce qui améliore sensiblement la qualité du moteur.

Cependant, le principal levier dont dispose l'architecte d'un moteur de recherche reste encore l'amélioration de l'algorithme d'évaluation de pertinence (le ranking). Sur le réseau, les algorithmes fondés sur la mise en correspondance des mots des requêtes et des mots contenus dans les documents trouvent rapidement leurs limites : documents volontairement biaisés par les «spammers», polysémie importante, nombre de documents trop élevé, duplication anarchique des documents. Tout concourt à faire échouer les algorithmes classiques d'évaluation et de pertinence. Pour ces raisons de nouveaux algorithmes, tels ceux utilisés par Clever ou Goggle, sont élaborés avec des premiers résultats prometteurs.

Un autre élément important de la conception d'un moteur de recherche réside en la prise en compte de la polysémie, problème particulièrement épineux sur l'Internet. Un acronyme comme «BSE» signifie, entre autres, Bovine Spongiform Encephalopathy («encéphalopathie spongiforme bovine»), Breast Self Examination («auto-examen des seins»), ou encore Bombay Stock Exchange («Bourse de Bombay»). Lorsqu'un utilisateur

effectue la requête «bse», il est donc essentiel que le moteur lui fournisse un moyen de préciser de manière simple l'objet réel de sa recherche. C'est ainsi qu'une équipe de l'École des mines de Paris a mis au point la technique Cow9, utilisée par le moteur AltaVista depuis 1997.

Cette technique permet d'affiner les requêtes à l'aide d'une liste thématique construite automatiquement à partir des résultats des recherches. Les thèmes non pertinents peuvent être exclus d'un simple clic de souris, tandis qu'un zoom est possible sur les thèmes jugés les plus pertinents. Une approche similaire, permettant le classement des résultats des recherches dans des dossiers thématiques (dont la liste est établie manuellement), a plus récemment été déployée sur le moteur de recherche NorthernLight.

D'autres approches sont utilisées sur certains moteurs pour aider les utilisateurs dans leurs recherches. Citons par exemple la fonction What's Related de Netscape, qui propose des liens vers des pages au contenu proche d'une page donnée.

Des approches différentes de celles qui sont suivies par les moteurs améliorent aussi la recherche d'information pertinente, comme la recherche par nom de marques de RealNames, la reformulation de requêtes en questions aux réponses connues, voie suivie par AskJeeves, l'approche des anneaux qui consiste à relier entre eux par des liens hypertextes les sites aux contenus voisins (ce qui ne résout toutefois pas le problème de trouver un premier site situé dans l'anneau), ou encore les méta-moteurs qui interrogent en parallèle plusieurs moteurs de recherche classiques et fusionnent ensuite de manière intelligente les résultats.

Sachant que le taux de couverture d'un moteur (c'est-à-dire la proportion de documents qui se trouve effectivement dans la base de données du moteur) est au maximum de 30 pour cent, l'utilisation de méta-moteurs peut sembler intéressante. Malheureusement, la pratique montre que la valeur ajoutée de ces outils est bien faible et que, malgré son taux de couverture limité, un moteur classique estime mieux la pertinence des documents qu'un méta-moteur.

Le monde des outils de navigation est en perpétuelle mutation, et les défis auxquels sont confrontés les architectes des moteurs de recherche sont légion. D'abord, comment espérer maintenir, à un coût raisonnable, un taux de couverture du Web suffisant quand le nombre de documents disponibles augmente de six pour cent par mois? Comment fournir des réponses pertinentes à un utilisateur qui fournit très peu d'indications sur ce qu'il recherche vraiment? Et comment amener, sans le décourager, l'utilisateur à en dire plus?

La prise en compte des liens hypertextes, comme avec Clever, améliore la pertinence des documents retournés, mais ces solutions sont loin d'être parfaites et on ne fera pas

l'économie d'une réflexion approfondie sur les impacts socio-économiques et culturels de ces méthodes, notamment les impacts liés au risque de consolidation des positions dominantes et au risque de perte du «service universel» (c'est-à-dire la garantie de pouvoir se faire entendre sur le réseau) pour le citoyen et les minorités. Il faudra d'autres approches fondées sur la collaboration.

2) Composition d'un moteur de recherche :

Le moteur de recherche est composé d'un robot, d'une base de données, d'un agent.

a) Les robots:

Ils sont appelés des "wanderers" (du verbe to wander : vagabonder, errer), des "crawler" (du verbe to crawl : ramper, se traîner) et aussi des "spiders" (araignées). Ce sont des programmes informatiques qui parcourent le WEB (toile) pour référencer les liens qui existent dans les pages. Un robot se comporte comme un visiteur, pas comme un virus. Il démarre d'une page de liens et suivra de façon récursive tous les liens qu'il trouvera à partir de cette page initiale. Ces robots utilisent le protocole http (hypertext transfer protocole) pour repérer les documents chez les serveurs (les nouveaux sites), indexer l'espace pour la recherche par mots - clés, rechercher les liens morts pour la maintenance des sites ajour. Leur fonction est d'indexer, de valider le texte en html (hypertext mark up language), les liens, les nouveautés, de créer des sites miroirs. Ils font une liste chronologique des URL (Uniform Ressource Locator), repèrent les documents qui ont des liens, les listes, les annuaires de nouveautés, les best of. Ils parcourent Internet constamment de façon automatique, ils suppriment les doublons. Chaque robot travaille à sa manière, certains travaillent sur des ressources plus nombreuses que d'autres. C'est la qualité de la démarche du robot lorsqu'il parcourt la toile qui détermine la qualité et la quantité des informations ramenées pour alimenter sa base de données. Ces robots rapatrient l'information trouvée à un instant T (stockage sur plusieurs Gigaoctets du disque dur de l'ordinateur). Lorsque vous interrogez ensuite les moteurs de recherche, le système recherche sur son disque dur les termes correspondants puis vous propose à l'affichage une description du site et le lien vers le serveur Web qui contient l'information. Or il se peut qu'entre-temps l'information originale ait été modifiée ou ait été déplacée du serveur Web. Vous ne pourrez donc pas forcément retrouver l'information concernée (messages d'erreurs du type ERROR The requested URL could not be retrieved ou File Not Found, The requested URL /scd-bib-histoire.html was not found on this server).

b) La base de données:

Les données ramenées par les robots sont indexées dans des catalogues qui contiennent les listes de notion repérées : adresse, titres, sous-titres, mots des premières lignes des

textes, résumés, éventuellement texte intégral. Ces données sont stockées dans la base de données du moteur avec une adresse qui localise les documents. Par des techniques heuristiques d'auto - apprentissage, le robot recherche, trouve et indexe les meilleurs sites. La taille de la base de données détermine la couverture de la recherche. Lycos, par exemple, a plusieurs bases de données : plus un moteur de recherche a de liens, plus il obtient de réponses et devient populaire.

c) L'agent :

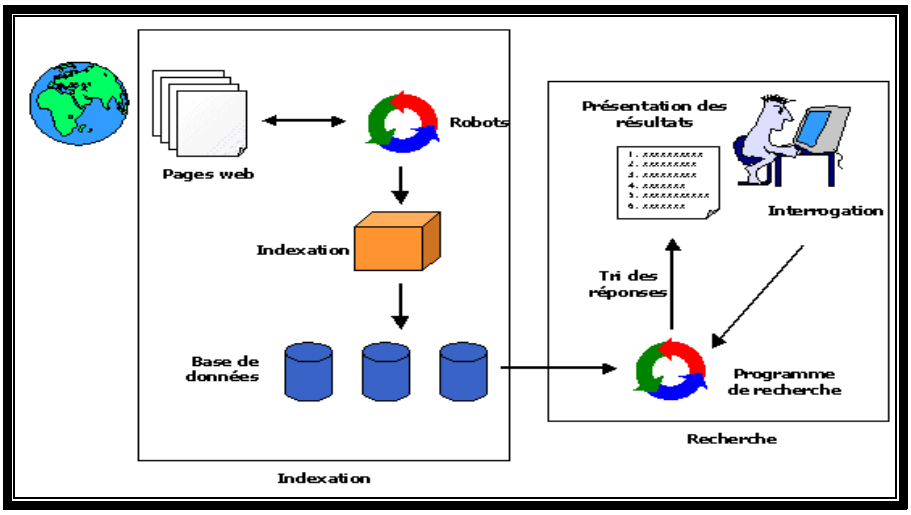
Il effectue la recherche pour l'utilisateur et propose une liste de réponses classées, dans un certain ordre de pertinence… Les moteurs de recherche affichent les adresses des documents qui mentionnent le plus fréquemment le mot-clé recherché. Parmi ces systèmes, on trouve :

- La solution d'agent simple : On en cite Telescript de General Magic ou Topic de Verity (tous deux en perte de vitesse à cause de problèmes commerciaux), mais aussi des systèmes universitaires tels que Softbot, SIMS, ou KSE qui proposent un modèle agent pour la recherche d'informations.

Autonomy, très inspiré de l'intelligence artificielle (développé par l'université de Cambridge), est un moteur à raisonnement dynamique qui suggère des notions voisines afin d'être le plus précis possible. Mais ces agents ne coopèrent pas.

- La solution multi agents : Les multi agents ont une capacité à agir et à décider. Leur principe de travail est basé sur la coopération, l'interaction et la résolution de conflits. Compte a leur capacité, elle se base sur les croyances et la capacité à raisonner sur autrui: c'est à dire sa capacité à apprendre de manière dynamique. La distribution est un des aspects essentiels des systèmes multi agents. Parmi les agents coopératifs, on peut citer InfoSleuth, basé sur la technologie Carnot, qui développe une modélisation sémantique permettant de décrire les ressources d'information et de promouvoir l'utilisation des agents. Avec InfoSleuth, on utilise des agents sémantiques pour effectuer des recherches coordonnées. ils coopèrent pour transformer les données trouvées en données compréhensibles.

3) Principes de fonctionnement :



a) La collecte des informations ou indexation automatique:

L'objectif des moteurs de recherche est de rapatrier et organiser l'information (contenue dans les pages HTML essentiellement mais aussi des sites GOPHER, FTP, des News).

La première étape consiste à parcourir le réseau. Elle est effectuée par des logiciels appelés robots (ou crawler, spider ou worm). Ils parcourent le réseau grâce aux liens hypertextes rencontrés. Il faut (en théorie) à l'heure actuelle environ 15 jours pour cela. Les fréquences de mises à jour varient selon les robots.

La seconde étape consiste à rapatrier tout ou partie de l'information trouvée et à l'indexer. Selon les robots, tout ou partie de l'information trouvée va être rapatriée : certains robots rapatrient uniquement les titres des fichiers HTML ou les premiers paragraphes et les mots les plus fréquents. D'autres rapatrient le contenu complet des fichiers HTML.

La parade mise en œuvre par les moteurs : un logiciel peut être mis en œuvre dont le rôle consiste à vérifier quasiment en permanence la validité des adresses URL. (Notez que les annuaires font de même pour éliminer les liens inactifs).

Les moteurs n'indexent "que" le contenu des pages HTML, des serveurs GOPHER, FTP et des News. Vous n'aurez pas accès par exemple au contenu des bases de données, aux références bibliographiques contenues sur les serveurs commerciaux, aux références bibliographiques contenues dans les catalogues de bibliothèques.

b) Le signalement des informations :

Tout ou partie du contenu des fichiers HTML est donc rapatrié vers le site du moteur de recherche, est indexé et devient un critère de recherche.

Le scoring et le ranking : Généralement la recherche produira des résultats qui seront affichés dans un ordre quelconque ou encore selon certains critères de tri reliés à la pertinence de l'information trouvée.

Le scoring est un système de notation et le ranking une méthode classificative. Cela permet d'évaluer un document cible, à la récurrence et au positionnement des mots clés en

son sein. Un pourcentage est calculé en fonction de ces données et permet de classer les sites par ordre du plus élevé vers le moins élevé.

La majorité des moteurs de recherche analysent l'intégralité d'une page Web en dénombrant chacune des répétitions du ou des mots clés spécifiés et en donnant plus d'importance aux mots situés au sommet du document et une moindre à ceux qui se trouvent en bas. D'autres ne tiennent compte que du titre ou des premières lignes du document ou encore de tout, hormis les commentaires, etc…

Les moteurs de recherche tiennent compte non seulement de la position des mots dans la page et de leur fréquence, mais aussi de leur mise en forme comme la taille des caractères (SIZE), la mise en gras (<B>), en italique (<I>) ou en souligné (<U>) ou encore les niveaux de titre (<H1>, <H2>, etc.)...

Malheureusement, les moteurs de recherche ne possèdent pas le même algorithme de calcul, si bien qu'une page pourra bénéficier d'un très bon scoring dans l'un et d'un très mauvais dans l'autre. Ces résultats contradictoires dans les moteurs de recherche vous obligent à composer en conséquence.

4) Différents moteurs et stratégies de recherche d'information :

Les moteurs de recherche permettent différentes manières pour rechercher l'information. Par exemple, Excite permet le langage naturel (descriptif), plus il y aura de mots pour exprimer la requête, plus les résultats seront bons. Magellan, au contraire, doit être interrogé de façon très simple. D'autres moteurs autorisent l'utilisation d'opérateurs.

	AltaVista	**HotBot**	**Excite**	**Lycos**	**Web Crawler**	**Google**
Date de lancement	Décembre 1995	Mai 1996	Octobre 1995	Juin 1995	Avril 1994	1998
Taille de l'index (millions de pages)	350	200	250	50	2	200
Délai de rafraîchissement de l'index	4 à 6 semaines	4 semaines	6 semaines	2 à 3 semaines	1 semaine	4 à 6 semaines
Nom du spider	Scooter	Slurp	ArchiText Spider	T-Rex	ArchitText Spider	Googlebot

Moteurs/noms	**Auteurs**	**Couverture**
AltaVista : indexation large sur texte intégral	Digital Research Laboratory	Le robot, Scooter, parcourt 2,5 millions de sites par jour

Lycos (nom d'araignée dangereuse)	Carnegie Mellon University	Ecrit en C, il explore de 5000 à 10000 documents par jour.
WebCrawler	Department of Computer Science and Engineering at Washington University (Seattle). Les éditeurs de GNN (Global Navigator Network) ont réalisé le catalogue	Tout Internet
WWWW (Worm) : un pionnier, très difficile d'accès, temps de réponse long		Découverte de sites nouveaux, inclut les sites GIF
Open Text : bon temps de réponse	développé et maintenu par Open Text Corporation. Yahoo a acheté la licence du logiciel pour renforcer sa recherche	Liens intéressants : permet de rechercher sur résumé ou texte intégral
Excite : utilise un moteur de recherche et une classification en catégories, propose de nombreux services		Couvre les sites les plus populaires, recherche plus des services que sur mots-clés
Infoseek : payant	Infoseek Corporation (California)	Index des pages, accès aux news, donne accès à des bases non accessibles par d'autres : textes intégraux de presse spécialisés
CUIW3	centre Universitaire d'Informatique de l'Université de Genève et l'institut d'informatique et de mathématiques appliquées de Bern	banque de données de mots-clés faite à partir des listes spécialisées et des catalogues-matières de certaines bases entièrement faite à la main

Les résultats des recherches effectuées au moyen de moteurs de recherche sont relativement difficiles à utiliser. Certains moteurs proposent des facilités comme le ranking à partir des mots de la requête, mais les choix sémantiques demeurent mystérieux. la notion de relevance feedback permet à l'utilisateur d'exprimer s'il aime ou non le résultat obtenu. Live Topics d'AltaVista fournit une aide à la formulation de la question en présentant une vision synthétique et globale des résultats. L'utilisateur vote et choisit de nouveaux termes qu'il associe à sa question (accède au contenu des résultats).

5) La Grosse Faille Des Moteurs De Recherche :

Les moteurs de recherche ont aujourd'hui un énorme défaut: ils n'analysent que le contenu des pages HTML et ne vont jamais explorer les bases de données. Or la plupart des sites professionnels sont organisés autour de bases de données. Ce problème n'est pas

mineur. Exemple : Vous créez un service de petites annonces de ventes de Bateaux. Votre service n'a pas beaucoup de succès et vous affichez en tout vingt annonces, dont deux pour des voiliers. Ces annonces seront présentées sous forme de pages HTML et les moteurs les prendront en compte.

A présent, supposons que le même type de service est beaucoup plus ambitieux. Vous avez six mille annonces, dont mille concernent des voiliers. Ces annonces sont enregistrées dans une base de données. Les pages sont générées à la demande en fonction de critères précisés par l'utilisateur. Comme les annonces sont stockées dans une base de données et non sur des pages HTML, elles ne seront pas prises en compte par le moteur de recherche.

En effet, celui-ci n'ira pas plus loin que la page HTML proprement dite, à savoir le formulaire de saisie.

IMPORTANT : Aucun des moteurs de recherche n'affiche ouvertement ses critères de tri de liste. C'est un sujet tabou. Et ce pour deux raisons :

La raison officielle est que ces critères de tri soient susceptibles d'évoluer rapidement .

La raison officieuse est que, en ne donnant pas ses critères de tri, et donc de choix, on évite les multiples plaintes de concepteurs de sites qui s'estimeraient mal placés.

Et pour certains moteurs, le fait de tenir le mode de classement secret permet de jouer sur celui-ci. Pour remonter, par exemple, la cote d'un site qui viendrait juste de signer un contrat publicitaire

6) Le Fameux Google :

En possession d'algorithmes de recherche puissants, d'un système d'indexation unique en son genre, à base de liens, et de plus de 3500 serveurs d'Architecture Intel®, Google propose aujourd'hui le moteur de recherche le plus rapide du Net.

a) Profil de l'entreprise :

Google est l'une des entreprises les plus novatrices du Web. Doté d'une technologie et d'une infrastructure originales, son moteur de recherche est capable d'apporter des réponses ultrarapides aux questions les plus complexes. Mis au point en 1998 par Larry Page et Sergey Brin, deux étudiants de Stanford alors en doctorat, Google ne tardera pas à s'imposer en douceur et bénéficie aujourd'hui d'une popularité grandissante sur l'Internet.

Malgré un budget publicitaire et marketing restreint, Google satisfait sur-le-champ plus de treize millions d'investigations quotidiennes, alors qu'il n'en totalisait guère plus de 500000 par jour en juin 1999. Le célèbre site NetCenter de Netscape en fait son moteur de

recherche privilégié et fournit aussi à certains homologues les services de Google à titre gratuit ou payant.

b) Opportunité, Une fructueuse rapidité :

Le terme concentration suffit à lui seul à résumer la stratégie de Google. Tandis que les poids lourds Yahoo! et AltaVista se transforment en portails multiservices qui proposent messagerie électronique et salons de chat, Google continue à peaufiner et à élargir sa technique fondamentale de recherche. Sa concentration exclusive sur son cœur de métier commence à porter ses fruits sous la forme de contrats lucratifs avec les constructeurs OEM et d'une forte croissance de la demande servie par un très efficace bouche-à-oreille du public.

Les recettes de la structure Google sont constituées d'une part par les licences d'utilisation de sa technologie accordées à des clients tels que Netscape NetCenter, Red Hat, le Washington Post et Virgin Net, et d'autre part par la publicité affichée sur les pages de résultats de recherche. La restriction de ces annonces à de simples slogans lui permet de rester fidèle à sa mission première qui est d'informer.

Il relève de l'évidence qu'un chercheur quelconque est incapable de se mesurer au service ultrarapide de Google. Les conclusions des bancs d'essais communiquées par des publications indépendantes attestent invariablement de la promptitude inégalée de ce moteur de recherche. De surcroît, sa méthode particulière d'évaluation de la pertinence des résultats d'après le nombre de liens pointant sur un site semble fournir des références curieusement appropriées. Il en résulte qu'une fois initiés à l'emploi de Google.com, les internautes réitèrent sans relâche leurs consultations.

« De conception modulaire, notre solution fait appel à de petits serveurs multi-redondants et ultrarapides grâce à la fonction d'équilibrage de charge. » Jim Reese, Ingénieur responsable de l'exploitation chez Google.

Google a reçu le titre de meilleure Cybertechnique de l'année 1999 (Best Cybertech of 1999) décerné par le magazine Time. Un an plus tard, la revue Time Digital l'a placé en tête des dix meilleurs sites. En 1999, il s'est également vu décerner le prix d'excellence technique (Technical Excellence Award) par PC Magazine. Au cours du dernier trimestre 1999 et du premier trimestre 2000, NPD Online Research a mené des enquêtes destinées à mesurer la satisfaction et la fidélité des utilisateurs. A deux reprises, ce moteur de recherche a été classé premier parmi treize sites de recherche et portails. Cependant, le ralliement massif du public a du même coup confronté Google à une croissance vertigineuse qui requiert une gestion avisée.

Il va de soi que tous ces témoignages d'approbation n'auraient aucune incidence positive si Google se ruinait en matériel. C'est la raison pour laquelle Jim Reese, ingénieur responsable de l'exploitation chez Google, a choisi des serveurs d'Architecture Intel® sous Linux* qui conviennent à merveille au plan d'action de l'entreprise : « L'utilisation de Linux sur la plate-forme Intel rend le rapport prix/performances imbattable. Cet avantage est tout simplement incomparable. »

c) Une solution disponible à grande échelle :

Google comprend mieux que ses concurrents l'intérêt de l'évolutivité horizontale. Il exploite donc la technologie RAIS (Redundant Arrays of Inexpensive Servers), capable d'administrer un volumineux index d'un à deux téraoctets ou 1000 gigaoctets comportant plus de 200 millions de pages Web. Pour obtenir des résultats de recherche ultrarapides à partir d'une base de données aussi consistante, Google déploie un gigantesque parc informatique constitué de plus de 3 500 serveurs monoprocesseurs d'Architecture Intel®. Ces machines, reliées par des cartes Ethernet Intel® PRO/100, sont dotées de 256 Mo à 1 Go de RAM, fonctionnent sous Linux et exploitent une série de solutions propriétaires.

« L'utilisation de Linux sur la plate-forme Intel® rend le rapport prix/performances imbattable. Cet avantage est tout simplement incomparable. » Jim Reese, ingénieur responsable de l'exploitation chez Google.

« L'application de Google ne nécessite pas l'acquisition d'un matériel onéreux », affirme Jim Reese. « Contrairement à un site d'e-Commerce, nous n'avons nul besoin de consacrer un gros budget à un gros système et à un réseau SAN. De conception modulaire, notre solution fait appel à de petits serveurs multiredondants et ultrarapides grâce à la fonction d'équilibrage de charge. Nous bénéficions en outre d'une tolérance exceptionnelle aux pannes et, même en cas de défaillance, nos serveurs fonctionnent toujours parfaitement. »

Dans l'environnement de Google, les performances d'E/S des disques sont essentielles, mais restent tributaires du coût élevé des sous-systèmes SCSI haut de gamme. Google adopte donc la technologie IDE, plus avantageuse économiquement, et équipe tous ses serveurs de deux disques internes d'une capacité de stockage de 22 ou 40 Go chacun : « Après avoir réalisé très tôt de nombreux bancs d'essais, nous nous sommes aperçus que, pour obtenir un bon rapport qualité/performances, il fallait installer deux disques durs IDE, chacun sur un contrôleur distinct », poursuit Jim Reese.

L'impressionnant index de recherche de Google est distribué et mis en miroir sur environ 7000 disques. Grâce à ce dispositif, un logiciel de répartition modulable de la charge dirige les requêtes vers les disques et les serveurs les plus disponibles. La connectivité réseau est assurée par les cartes Ethernet Intel PRO/100 équipant les serveurs,

eux-mêmes reliés à une dorsale Ethernet Gigabit. Ce mode d'organisation donne ainsi lieu à une réactivité exceptionnelle.

La technologie RAIS met à la disposition de Google une infrastructure performante, très largement évolutive et économique, capable de faire face aux besoins croissants des utilisateurs. Google déploie **chaque jour 30 nouveaux serveurs** pour renforcer son parc informatique et répondre à la demande. Les serveurs se répartissent sur deux sites d'hébergement de la baie de San FranCISCO et sur un troisième situé sur la côte EST des Etats-Unis. Google envisage de se doter d'infrastructures en Asie et en Europe afin de réduire le temps de latence des recherches effectuées par les utilisateurs de ces deux continents.

d) Des applications sur mesure :

Avec un parc informatique qui devrait compter 10000 serveurs d'ici à la fin de l'année 2001, Google assume des frais de locaux non négligeables. L'entreprise californienne a tout de suite travaillé en étroite collaboration avec plusieurs constructeurs OEM pour obtenir des unités centrales et des racks plus compacts. Google pense pouvoir abriter 8 serveurs dans un espace d'environ 2 mètres de haut sur 60 centimètres de large et 75 centimètres de profondeur. Il diversifie ses sources d'approvisionnement en matériel pour parer aux ruptures de stock et s'assurer des livraisons toujours à point nommé. La société Rackable Systems, constructeur spécialisé dans les serveurs compacts, se charge de monter les configurations.

Certes, la gestion d'un parc de serveurs aussi important représente un défi unique en son genre. Google a élaboré sa propre solution d'administration à distance et d'équilibrage de charge. Cette entreprise développe d'ailleurs elle-même la plupart de ses logiciels, notamment ses consoles d'administration à distance et ses moteurs d'équilibrage de charge. Enfin, elle a conçu une méthode rationalisée de configuration des serveurs qui permet la mise en ligne rapide des nouveaux systèmes.

« Nos machines sont toutes modulaires. A partir d'un matériel de base et du système d'exploitation Linux, nous les avons, disons-le, " googlelisées " en leur adjoignant du code maison », reprend Jim Reese. « A ce stade, elles sont toutes interchangeables. Il nous est possible d'ajouter des serveurs Web en cas de besoin, car toutes nos machines possèdent des configurations identiques. »

Dans le même temps, Google continue de perfectionner ses services. En outre, le bouton « J'ai de la chance », présent sur chaque page de recherche, conduit directement l'utilisateur au résultat le plus pertinent. En résumé, grâce à la rapidité du processus de configuration du

système et à sa haute modularité, Google est à même de faire évoluer son infrastructure à mesure du déploiement de nouveaux services.

e) Synthèse :

Bien qu'en possession d'un parc de plus de 3500 serveurs d'Architecture Intel® qui bourdonne comme une ruche laborieuse dans les locaux de son hébergeur, rien ne semble arrêter Google dans sa course à l'expansion. Son taux de croissance atteint 25 % par mois et devrait connaître un essor encore plus accéléré compte tenu de la multiplicité des sites qui utilisent son service gratuit ou payant de recherche sur le Web.

Néanmoins, le public est assuré que Google est prêt à s'équiper de serveurs d'Architecture Intel® supplémentaires pour relever le défi. En bâtissant une infrastructure hautement évolutive, Google tire le meilleur parti de sa technologie avancée, et possède un allié puissant et incomparable. A mesure du développement du site www.google.com par paliers, l'entreprise agrandit sa base opérationnelle et conforte sa position de moteur de recherche le plus rapide du Web.

CHAPITRE D

CONCEPTION ET IMPLEMENTATION

I- CONCEPTION :

1) Etude de l'existant :

A ce jour, il n'existe aucun moteur de recherche tunisien digne de ce nom. Et voici quelques défauts flagrants des plus populaires des sites tunisiens :

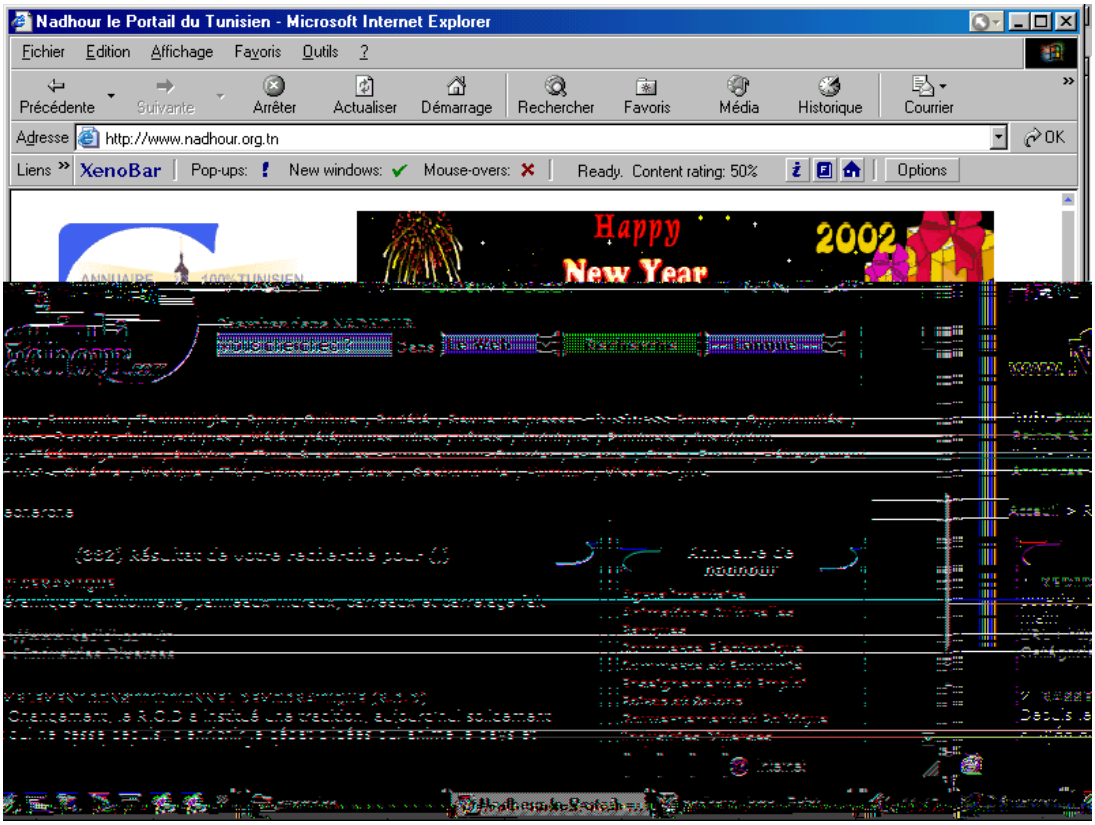
Commençant par le site du leader des fournisseurs Internet sur le marché : Planet

Le fait qu'il a choisit de faire appel à Lycos malgré son armada d'ingénieurs et de développeurs expérimentés montre du doit la difficulté que représente la conception d'un moteur de recherche même pour « Planet ».

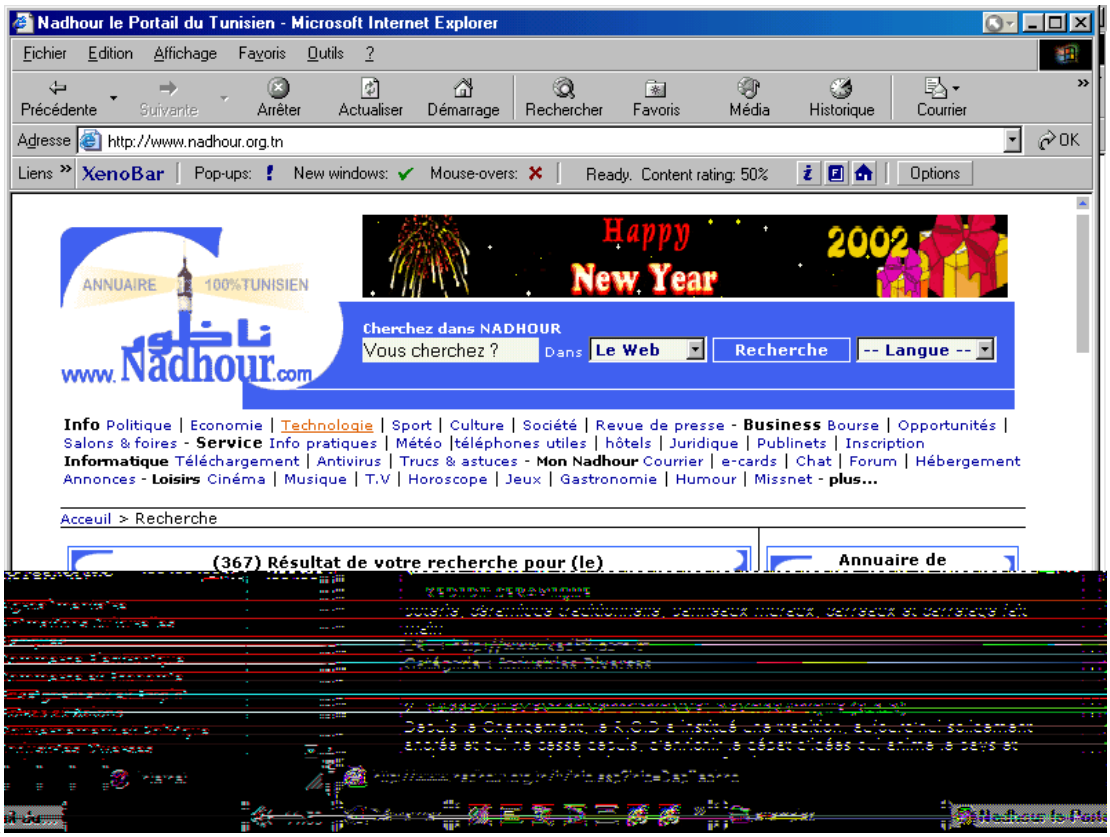
Continuant avec celui qui a été le plus remercié de tous les sites tunisiens : prix présidentiel, titre d'Oscar Web de l'année 2001 : « www.nadhour.org.tn »

Ceci est le résultat de la recherche d'un « s ».

Et ce ci pour un « , ».



Enfin, pour « le ».



Voici la liste des annuaires de recherche tunisien et arabe :

- Hahooa : La version arabe de voila par wanadoo.
- AltaVista : Le célèbre moteur de recherche arabisé et adapté au Monde arabe.
- Al-Taweel : Moteur de recherche en arabe pour les sites arabes.
- Al-Murshid : Moteur de recherche en anglais.
- Art Arab : Moteur de recherche en arabe, spécialisé dans les services gratuits et l'Internet du Monde arabe.
- Swaah 2000 : Moteur de recherche en arabe avec des liens intéressants vers d'autres sites.
- Yasalaam : Moteur de recherche en arabe et en anglais.
- Saudi Links : Moteur de recherche par thème, assez complet, sur le Monde arabe, proposé par Saudi Links. En arabe.
- AlDalil : Moteur de recherche par thème, sur le Monde arabe, proposé par Sakhr. En arabe ou en anglais au choix.
- Arabnet : Moteur de recherche par pays.
- Arabseek : Moteur de recherche par domaine.
- Maghrebnet : Moteur de recherche par domaine sur le Maghreb.
- 1001 Sites : Moteur de recherche par domaine ou mots-clés sur les pays arabes.
- Konouz : Moteur de recherche en arabe.
- Alsaha : Moteur de recherche en arabe.
- Arabic2000 : Moteur de recherche en arabe.
- Country information : Moteur de recherche d'informations sur les pays arabes.

II- Outils utilisés :

1) Java :

Java, développé par une équipe de Sun Microsystems dirigée par James Gosling1, est le dernier né des langages de programmation par objet. Fondé notamment sur le principe d'indépendance du support d'exécution, il permet de générer du pseudo-code (appelé couramment bytecode) interprété par une machine virtuelle.

L'essor de Java a été largement catalysé et médiatisé par le développement de l'Internet et du World Wide Web. Une des caractéristiques les plus connues du code Java est en effet de pouvoir être téléchargé et exécuté par des navigateurs Web. L'impact est tel que l'on estime aujourd'hui le nombre de développeurs Java à plus de 500 000 à travers le monde.

La mise en œuvre dans les réseaux privés d'entreprise des technologies de l'Internet que constitue l'Intranet ouvre, grâce à Java, de nouvelles perspectives qui révolutionnent la manière de concevoir les architectures Clients-Serveurs déployées jusqu'à aujourd'hui. Les postes clients s'allègent et représentent des coûts d'administration moins importants. Les

serveurs hébergent non seulement les données de l'entreprise mais aussi les applications écrites en Java, qui sont écrites une fois et une seule, quelle que soit leur support final d'exécution. Elles peuvent ainsi être mises à jour de manière centralisée.

Les librairies de classes Java spécialisées, qui intègrent en particulier l'interrogation des bases de donnés, ainsi que l'apparition des Network Computers, font de Java une technologie clé du déploiement de ce nouveau système d'information de l'entreprise, centré sur le réseau.

a) Pourquoi le choix de Java?

Java a été conçu dès son origine pour permettre le développement d'applications pour des environnements hétérogènes communicants. Le projet original, connu sous le nom de Oak, avait notamment pour but de proposer un environnement de développement logiciel pour des applications destinées aux équipements électroniques intégrés à des réseaux, tels que les assistants personnels numériques, les décodeurs de télévision numérique (Set Top Boxes) ou autres pagers.

L'objectif était donc de développer une plate-forme distribuée capable d'exécuter en temps réel du code à la fois compact, robuste, sécurisé et portable. On dit parfois que les cafétérias d'entreprise sont le lieu de réunions fructueuses. Java y a peut être quant à lui trouver son nom, qui en américain désigne une bonne tasse de café noir. De nombreux produits associés à Java sont de la même veine, jusqu'aux mesures de performance des interpréteurs Java qui s'expriment en CaffeineMark !

Si la syntaxe de ce nouveau langage est pour une large part empruntée à C++, sa conception et son architecture s'inspirent également de nombreux autres langages comme Eiffel, Smalltalk, Objective C et Cedar/Mesa.

Le résultat que nous connaissons aujourd'hui se révèle idéal pour le développement d'applications interactives, distribuées, sécurisées et intégrées au réseau, pouvant s'exécuter sur un large ensemble de plate-forme, du World Wide Web au poste de travail individuel en passant par des appareils électroniques grand-public.

Java est en effet un langage interprété et orienté-objet, permettant de générer du code dynamique et distribué. Il satisfait aux exigences de simplicité, de robustesse, de sécurité, de portabilité et d'indépendance vis-à-vis des architectures matérielles. Performant, il est aussi capable de gérer l'exécution de tâches en parallèle au sein d'une même application. Java est soutenu par un grand nombre de très grand noms de l'informatique : SUN, IBM, ORACLE...

b) Avantages du Java :

- **Familiarité** :

A première vue, si l'on s'en tient à la syntaxe, un programme Java ressemble à un programme C++. Un programmeur familier des langages C ou C++ se sentira donc en terrain de connaissance. Cette familiarité fait de Java un langage dont l'apprentissage est facile. Il ne nécessite pas un investissement de temps important pour une équipe de développeurs expérimentés et ses mécanismes sont faciles d'accès pour le débutant. Pour reprendre un exemple souvent cité dans les manuels de programmation, voici un prototype de programme Java qui ne fait qu'afficher une chaîne de caractère.

- **Simplicité :**

Java est, en outre, plus facile à comprendre et à maîtriser que C++, qui intègre des fonctionnalités complexes, dont certaines sont, en pratique, peu utilisées. En comparaison, Java réduit le nombre de structures de données utilisables au minimum et élimine les zones de recouvrement. Java délaisse des fonctionnalités comme l'héritage multiple ainsi que la surcharge d'opérateurs et abandonne des primitives comme #define ou typedef.

Pour le programmeur, la simplicité de Java s'exprime aussi en grande partie par le fait qu'il est libéré de la gestion explicite des adresses mémoire des objets qu'il manipule. Cette disposition évite les erreurs de programmation les plus fréquentes et les plus délicates à corriger. En contrepartie, des fonctionnalités ont été ajoutées, comme le «garbage collector», qui assure les tâches de libération de mémoire pour les objets dont la vie vient à expiration.

- **Neutralité vis-à-vis des architectures matérielles et logicielles :**

Java utilise un double mécanisme de compilation et d'interprétation. Ce mécanisme permet au code généré par le compilateur Java, appelé bytecode, d'être indépendant des architectures matérielles et des systèmes d'exploitation. Le bytecode Java est en effet chargé et interprété par une machine virtuelle, qui est elle même portée sur la plupart des systèmes d'exploitation du marché.

Cette machine virtuelle fait notamment partie intégrante des navigateurs Web compatibles Java (Netscape et Hotjava de Sun en sont les premiers exemplaires). L'utilisation de Java et de sa librairie graphique AWT (Abstract Window Toolkit) permet donc d'écrire des applications directement exécutables sur Solaris et la plupart des systèmes UNIX, Windows 95, Windows NT, Mac OS ainsi que des systèmes temps-réel comme OS9 et VxWorks.

- **Portabilité :**

La machine virtuelle Java, qui, disponible sur toutes les plates-formes informatiques du marché, dissimule les architectures matérielles et logicielles vis-à-vis des applications, est en soi un premier gage de portabilité. Il restait à définir les types de base du langage (entiers, flottants, etc), selon des formats indépendants des différents systèmes d'exploitation, pour atteindre une portabilité totale. L'approche suivie par Java consiste donc à figer la représentation de ces types. Les choix qui ont été faits sont les suivants : Le système Java lui-même est entièrement portable. Le compilateur Java est écrit en Java. Le runtime, écrit en C ANSI, est conforme POSIX. Il n'existe aucun point dépendant de l'implémentation dans les spécifications du langage.

- **Robustesse et sécurité :**

Le principe général de sécurité mis en oeuvre par la machine virtuelle Java consiste à délimiter l'espace dans lequel les applications vont pouvoir agir. Elle se comporte ainsi un peu comme des parents qui laissent leur enfant en bas âge jouer dans un bac à sable.

Deux étapes de vérification du code sont effectuées. Elles ont lieu au moment de la compilation et avant l'exécution du programme assurée par l'interpréteur. Java étant un langage fortement typé, une vérification stricte du programme est faite au moment de la compilation. Cette vérification est d'autant plus stricte que Java impose au développeur des règles d'écritures rigides. Les déclarations de variables par exemple doivent être explicites. De plus, le bytecode étant un code symbolique, lui même typé, certaines vérifications effectuées à la compilation sont répétées au runtime. L'éditeur de liens comprend le système de typage, et répète certaines phases de vérifications de types pour résoudre les problèmes d'incohérences entre versions de logiciels.

Java élimine en outre la possibilité d'écrire dans une zone de mémoire déjà allouée (ou d'écrire dans une zone défendue), et de corrompre des données. Le langage substitue à l'arithmétique sur les pointeurs l'utilisation de véritables tableaux et chaînes de caractères. L'interpréteur Java est à même de vérifier les index sur ces tableaux et ces chaînes, ce qui élimine les erreurs par débordement des écritures en mémoire.

De plus, le programmeur ne peut transformer un entier en une référence à un objet par simple conversion de type (casting). Si Java n'a pas la prétention de résoudre tous les problèmes liés à la qualité du logiciel, il n'en reste pas moins que ces dispositions, qui éliminent une catégorie entière d'erreurs de programmation, facilitent d'autant les phases de test et d'assurance qualité.

- **Java est un langage interprété et dynamique :**

L'interpréteur Java exécute directement les bytecodes objets Java sur tout système sur lequel est disponible le runtime Java. Les phases de vérifications étant réalisées avant la phase d'interprétation du code, l'interpréteur lui-même fonctionne à pleine vitesse car il est purement dédié à l'exécution du programme.

Java est un langage dynamique, à savoir que les liens entre les objets sont résolus lors de l'exécution, et non au moment de la compilation. C'est une caractéristique primordiale dans un contexte où il est nécessaire pour un objet de pouvoir envoyer des messages à un autre objet sans en connaître a priori le type. L'édition de liens dynamiques apporte la flexibilité maximale au moment de l'exécution des applets ou des applications. Il existe néanmoins deux exceptions au modèle, qui sont utilisables dans des contextes spécifiques.

La première concerne la possibilité d'écrire des méthodes natives, dans un autre langage que Java (C ou C++ par exemple), mais invocables au coeur d'un programme Java. Le contexte d'utilisation est alors en général local, la méthode employée étant pré-compilée et dépendante de la plate-forme matérielle. Il peut s'agir d'exploiter un périphérique particulier ou plus simplement d'exploiter des bibliothèques écrites dans d'autres langages, la partie Java pouvant être dans ce cas une simple interface utilisateur.

La seconde n'a pas à passer le bytecode par l'interpréteur pour son exécution, mais par un compilateur. Ce compilateur, appelé souvent «Just In Time Compiler», génère un code machine directement exécutable sur le système d'exploitation natif de la machine hôte.

Cette opération permet des gains de performances importants. Les temps d'exécutions sont alors quasi identiques à ceux obtenus avec des programmes C++ équivalents. Java apporte tout de même le bénéfice du garbage collector, qui gère les tâches de libération de mémoire de manière optimisée. Après tout, Java est aussi un langage de programmation généraliste, dont les caractéristiques permettent notamment le prototypage rapide d'applications. Celles-ci n'ont pas nécessairement de rapport avec le Web !

- **Vers une nouvelle architecture Client-Serveur :**

La mise en oeuvre d'applications Java dans un réseau d'entreprise permet, en s'appuyant sur les APIs d'extension telles que JDBC ou JavaIDL, de déployer des applications client-serveur. Une application peut maintenant être stockée sous forme d'applets sur un serveur Web interne, et téléchargée simplement à travers le réseau. L'interface d'accès aux applications est un browser Web, ou un logiciel qui en reprend les fonctions principales (c'est par exemple le cas de **Hotjava** de Sun). Au moment de l'exécution sur le poste client, cette application peut à son tour se connecter à un serveur de bases de données, communiquer avec des objets distribués, selon la norme CORBA, ou encore coopérer avec des applicatifs comme des agents de recherche intelligents.

On peut dès aujourd'hui bâtir un véritable Intranet en s'appuyant sur les standards du Web. En intégrant les possibilités de relier simplement les applications Java au système d'information de l'entreprise, on est maintenant capable de déployer des applications distribuées qui héritent naturellement des qualités intrinsèques de Java. Parmi celles-ci, la portabilité, primordiale, permet aux applications de s'exécuter sur tous les types de postes clients existants (PC, Mac et stations UNIX). La seule contrainte de ces postes de travail est alors d'héberger un browser Web compatible Java. Un programme Java est écrit une seule fois, quel que soit le nombre de postes clients différents sur lesquels il est susceptible de s'exécuter. Cette approche transforme l'architecture classique des applications clients /serveur. Il fallait jusqu'à présent installer individuellement les applications sur chacun des postes clients, les mettre à jour et les maintenir. La complexité de cette tâche augmentait avec le nombre de plate-forme hétérogènes que l'on devait gérer.

Grâce à la méthode d'accès uniforme offerte par l'Intranet, et au fait que les serveurs sont désormais capables d'héberger à la fois les données et les applications clientes sous formes d'applets Java, la gestion du système d'information de l'entreprise devient extrêmement flexible. L'économie d'échelle réalisée sur les coûts d'administration des postes clients est un bénéfice de ce nouveau modèle, centré sur le réseau. L'arrivée prochaine des Network Computer ne fait qu'ajouter un gain supplémentaire dans les économies réalisables pour catégorie entière d'applications client-serveur.

c) les Servlets :

- **Qu'est qu'une servlet Java ?**

Les servlets sont aux serveurs ce que sont les applets aux browsers. Ils correspondent à des programmes Java normaux qui utilisent des modules supplémentaires (ainsi que les classes et les méthodes associées) figurant dans l'API des servlets Java. Les servlets s'exécutent sur une machine de serveur Web à l'intérieur d'un serveur compatible avec Java. Ils permettent l'extension des fonctions du serveur.

Une servlet peut être chargé automatiquement lors du démarrage du serveur Web. Elle peut également être chargée lorsque le premier client demande les services de la servlet. Une fois chargées, les servlets restent actives dans l'attente d'autres requêtes du client.

Les servlets permettent l'extension des fonctions du serveur grâce à la création d'un environnement de prestation de services de requête/réponse via le Web. Un client envoie une requête au serveur. Ce dernier transmet au servlet les informations relatives à la requête. La servlet crée ensuite une réponse que le serveur renvoie au client.

Dans la mesure où il s'agit d'un programme Java, la servlet peut utiliser toutes les fonctions du langage Java lors de la création de la réponse. La servlet peut également

communiquer avec des ressources externes telles que des fichiers ou des bases de données, ou avec d'autres applications (également écrites en langage Java ou dans d'autres langages), afin de créer la réponse et éventuellement de sauvegarder des informations relatives à l'interaction requête/réponse.

La réponse envoyée au client peut donc être une réponse dynamique et unique conçue pour une interaction particulière et non une page HTML statique existante.

- **Quel est le cycle de vie d'une servlet ?**

  Le servlet est chargé (automatiquement au démarrage du serveur, ou lors de la première requête du client) .

  Le serveur crée une instance du servlet .

  Le serveur appelle la méthode *init()* du servlet .

  Une requête du client parvient au serveur (elle figure déjà sur le serveur si la requête du client a lancé le chargement du servlet) .

  Le serveur crée un objet *Request* spécifique à cette requête .

  Le serveur crée un objet *Response* spécifique à cette requête .

  Le serveur appelle la méthode *service()* du servlet pour transmettre les objets *Request* et *Response* .

  La méthode *service()* reçoit de la part de l'objet Request des informations concernant la requête, traite cette dernière, puis utilise des méthodes de l'objet *Response* pour renvoyer la réponse au client. Il se peut que la méthode *service()* appelle d'autres méthodes pour traiter la requête, par exemple, *doGet()* ou *doPost()* ou de nouvelles méthodes écrites par vous-même .

  Pour chaque requête supplémentaire du client, reprenez la procédure à l'étape 4 .

  Lorsque le servlet n'est plus requis par le serveur, ce dernier appelle la méthode *destroy()* du servlet.

- **A quoi servent les servlets ?**

Les servlets exécutent un grand nombre de fonctions, par exemple :

- Une servlet peut créer et renvoyer une page Web HTML complète dont le contenu dynamique dépend de la nature de la requête du client .
- Une servlet peut simplement créer une partie d'une page Web HTML qui est intégrée à une page HTML statique existante .
- Une servlet peut communiquer avec d'autres ressources du serveur, y compris des bases de données, d'autres applications Java et des applications écrites dans d'autres langages .

- Une servlet peut traiter les connexions avec plusieurs clients en acceptant les données en entrée de plusieurs clients et en diffusant à ces derniers des résultats. Une servlet peut, par exemple, correspondre à un serveur de jeux électroniques faisant intervenir plusieurs joueurs .
- Une servlet d'un serveur peut établir une connexion distincte avec une de la machine du client et maintenir la connexion, ce qui permet plusieurs transferts de données par le biais de la même connexion. Ces performances permettent une communication facile entre le client et le serveur. Cette communication peut se faire via un protocole personnalisé ou via une norme telle que IIOP.

Vous pouvez développer une servlet qui permet à un client de télécharger un agent vers la machine du serveur qui exécute la servlet. Dans la mesure où cette dernière se situe à proximité des données et des autres ressources dont elle a besoin, le trafic du réseau est réduit et un petit ensemble de résultats est renvoyé.

➢ **Pourquoi le choix des Servlets :**

Les servlets sont le seul produit d'interconnexion entre Web et base de données qui :

- est indépendant des OS (Unix ou NT) .
- est indépendant des serveurs Web (Apache, IIS ou Netscape) .
  peut produire de l'HTML côté client (notamment pour la consultation de la base), sur la base d'HTTP .
  peut dialoguer avec des applets Java côté client avec un protocole à objets distribués de type RMI .
  s'appuie sur un langage vraiment standard : Java (et non pas Java script ou Visual Basic).
  fait mieux que CGI en prenant en charge les connexions des utilisateurs en multi-thread automatiquement.

Servlets, scripts CGI, ISAPI et NSAPI?

La plus grosse différence entre les CGI et les servlets est la performance. Il n'y a qu'une seule machine virtuelle Java qui tourne sur le serveur, et la servlet est placée en mémoire une fois qu'elle est appelée. Elle n'est pas remise en mémoire jusqu'à ce que la servlet change, et une servlet dont le code a été modifié peut être réactivé (c'est à dire replacé en mémoire) sans redémarrer le serveur ou l'application. Les servlets résident en mémoire, n'utilisent qu'un seul processus pour toutes les instances (mais plusieurs threads). De ce fait leur exécution est très rapide et moins gourmands en ressources (CPU, Mémoire). L'information statique peut être donc partagée par plusieurs invocations de la servlet, vous

autorisant ainsi de partager cette information entre plusieurs utilisateurs et une servlet pouvant supporter plusieurs requêtes concurremment et les synchroniser.

Pour pallier les faiblesses de CGI, Microsoft et Netscape ont mis au point leurs propres API, afin de permettre aux développeurs de créer des applications serveur sous forme de bibliothèques partagées. Ces bibliothèques sont conçues pour être chargées dans le même processus que le serveur Web. De plus, elles peuvent traiter de multiples requêtes sans avoir recours à plusieurs processus. Elles peuvent être chargées lors de démarrage du serveur Web ou en cas de besoin. Si elles ne sont pas utilisées pendant certain temps, le serveur les décharge de la mémoire.

Même si ces bibliothèques in-process enrichissent le serveur web, elles présentent certains défauts :

Ces API dépendent d'une plate-forme spécifique. C'est pourquoi les programmes écrits à l'aide de celles-ci ne peuvent être utilisées que sur la plate-forme correspondante. Il est difficile, voire impossible, de placer ces programmes dans d'autres environnements .

De nombreux utilisateurs accèdent simultanément à ces bibliothèques. L'exécution parallèle des threads doit être fiable. Cela signifie que les utilisateurs doivent être très prudents lorsqu'ils accèdent aux variables globales et statiques .

Si un programme serveur provoque une violation d'accès, il est possible que le serveur Web dans son ensemble « plante » car ce conflit survient dans le même processus.

Les servlets sont modulaires, chaque servlet peut accomplir une tâche spécifique et ainsi vous pouvez les rendre communicantes. Les servlets peuvent être indépendantes de l'OS (Unix, NT…) et du serveur Web. (Les servlets sont supposés être exécutables sans modification, d'une plateforme à l'autre, tel que l'annonce Sun : "**Write once, Run anywhere**".

d) Java Server Page :

PHP- JSP :

- PHP est un langage procédural avec possibilité de faire de l'objet.

-JSP (donc java) est un langage full objet avec tous les avantages que cela comporte en terme de rapidité de développement plus de nombreux IDE.

- En terme de système d'information, java possède beaucoup plus d'Api que PHP pour tous types de connexions à d'autres applicatifs ou protocoles : cela permet de se greffer plus facilement à tout système d'information : CICS, MQSeries, SNMP, tivoli, bases de données,

Corba, RMI, XMLRPC, FTP... Notamment, java permet de faire tourner de réelle applications derrière le serveur applicatif.

- PHP peut se connecter à toutes les bases de données du marche mais les interfaces sont propriétaires : Pour ouvrir une connexion avec MYSQL : « mysql_connect », pour ouvrir une connexion avec Oracle : « OCILogonJava » à une interface d'abstraction de connexion au bases : « JDBC » (Java Data Base Connectivity).

Avantages par rapport à PHP :

Lorsque l'on change de base de données, normalement, il n'y a pas besoin de changer le code de l'application.

- Toujours concernant les bases de données : PHP gère en natif un système de pooling de connexion. Ce système, quoi que relativement performant n'est pas aussi performant que les systèmes de pool de connexion que l'on peut avoir en java (plus de possibilité de configuration du pool).
- Les serveurs applicatifs java (respectant les spécifications servlet 2.2) permettent beaucoup plus de choses : meilleure gestion des sessions, gestion de contexte meilleure gestion du système de request.
- Les serveurs applicatifs java peuvent se greffer sur tous les serveurs WEB du marché, **PHP** ne doit que fonctionner avec Apache et IIS.
- les serveurs applicatifs java peuvent fonctionner "in process" ou "out process" : c'est a dire dans le même processus que le serveur web ou dans un processus sépare. Ce qui permet (dans le cas du out-process) de placer le serveur web et le serveur applicatif sur des machines différentes : cela permet d'avoir un système complet N-Tiers qui assure notamment du load balancing et de la haute disponibilité a tous les niveaux. PHP ne permet pas cette souplesse d'architecture, car il tourne "in-process", (je crois que PHP possède maintenant un module lui permettant de faire de l'out-process aussi, mais lorsqu'on l'utilise, je crois que certaines fonctionnalités tel que la gestion des sessions ne fonctionnent pas).
- Au niveau du développement de l'application, java permet de manière plus simple de respecter de design patern MVC (Model View Controler). D'ailleurs, un grand nombre de framework de développement existent pour faciliter et respecter ce paradigme (voir les framework STRUTS et TURBINE de l'Apache Software Foundation). Implémenter un MVC en PHP est pratiquement impossible, car pas de possibilité de faire du code autrement que des pages PHP, alors que java permet de créer des servlets (Controler), des javaBeans (Model) et des JSP (View) ce qui confère une grande souplesse de développement (partages nette des taches et des compétences sur un projet).

- Comme son nom l'indique (PHP = Personnal Home Page) permet le développement plus rapide de petites applications web, mais en aucun cas de vrai applications web scalable et fortement connectées. Dés lors que le projet commence à devenir gros, le PHP devient complexe à mettre en oeuvre.
- PHP nécessite un peu moins de ressource qu'un serveur applicatif java. Si on résonne en revanche en terme de coût : les coûts induit (pour de gros projet) par des développements PHP dépasse très rapidement le coût marginal de la puissance machine nécessaire pour faire tourner la même chose en java. En terme de rapidité d'exécution, les temps de réponses de java sont aujourd'hui équivalent a ceux de PHP, même si la ressource machine et mémoire est supérieur pour java.
- Enfin, mais ça a son importance, Java est soutenu par un grand nombre de très grand noms de l'informatique : SUN, IBM, ORACLE...

L'avenir du Java : Les seuls qui doutent encore du Java sont ceux qui ne le connaissent pas. Citons en terminant Bill Gates, l'homme le plus riche de la terre, écrivant au sujet du Java : "This scares the hell out of me !".

Remarque : Ce qui fait la force de JSP, ce n'est pas JSP lui même, mais c'est JAVA, dont JSP est un moyen simple de fabriquer des pages HTML. Car en terme pure, JSP tout seul n'a pas grand avantages par rapport à PHP puisque globalement ils font la même chose : mettre des lignes de programme dans une page HTML qui sera exécutée par le serveur.

e) L'installation De L'environnement De Développement Java :

Pour cela, il suffit d'utiliser le JDK1.3 fournit par SUN. Si vous disposez d'un kit de développement java d'une version ultérieure ou égale à la version 1.1, vous pouvez ignorer cette étape. Il est téléchargeable au format tar.gz à L'URL suivant : *http://java.sun.com/j2se/1.3/download-linux.html*. Une fois ce fichier téléchargé, vous devez le décompresser et le désarchiver dans le répertoire de votre choix (par exemple */usr/local/jdk1.3*) grâce à la commande suivante :

Tar xfvz j2sdk-1_3_0-linux.bin -c /usr/local

Une fois ceci effectué, déplacez vous dans le répertoire $*java_home/bin* ou *$java_home* de la manière suivante :

Cd /usr/local/jdk1.3/bin

Lancez alors la commande :

./javac

Si vous obtenez un message vous indiquant toutes les options disponibles pour le compilateur java, c'est gagné. Afin de faire bénéficier toutes vos applications susceptibles

d'utiliser java (comme Tomcat) de l'environnement installé, il faut ajouter les outils du langage dans votre path comme ceci :

Export path=$path:/usr/local/jdk1.3/bin

En considérant toujours que vous avez installé le JDK dans */usr/local/jdk1.3.*

Maintenant que l'environnement de développement pour java est installé, nous allons passer à l'installation du moteur de servlets Tomcat.

- **L'installation De Tomcat :**

Nous allons maintenant procéder à l'installation du moteur de servlets. Pour cela vous devez d'abord télécharger la version 3.2.1 de Tomcat disponible à L'URL suivant: *http://jakarta.apache.org/builds/jakarta-tomcat/release/v3.2.1/bin/*. Cette version n'est pas la dernière version de Tomcat. En effet le développement de Tomcat est vraiment très actif, et de nouveaux concepts sont introduits régulièrement.

Par exemple, en ce qui concerne le protocole de communication entre le serveur web Apache et le moteur de servlets Tomcat, un nouveau procède est apparu depuis la version 3.2.1 : JK. Mod_JK (le module dynamique chargé par apache au démarrage pour prendre en compte JK) est destiné, à moyen terme, au remplacement de Mod_JSERV (le protocole de communication entre le serveur web et le moteur de servlets). JK apporte un certain nombre d'avantages comme le support de plusieurs serveurs web différents via la même interface (IIS, NETSCAPE SERVER, APACHE 2.X, etc.), la gestion correcte du protocole https (protocole http sécurisé grâce a l'utilisation de SSL). De plus, JK n'est pas une adaptation de d'Apache JSERV au contraire de Mod_JSERV qui comporte donc du code inutile. De plus, Mod_JSERV est encore largement plus répandu que Mod_JK.

Afin d'installer Tomcat, il faut télécharger les sources de la dernière version stable à L'URL suivant : *http://jakarta.apache.org/builds/tomcat/release/v3.2.1/src/.* La dernière version stable est la 3.2.1.

Vous devrez ensuite télécharger d'autres outils :

- **JSSE :** pour Java Secoure Socket Extension: permet d'établir des connexions sécurisées au travers d'un réseau (a l'aide de SSL par exemple) entre deux machines. Disponible à *http://java.sun.com/products/jsse/.*
- **ANT :** c'est un gestionnaire de compilation conditionnelle, au même titre que make sous unix. Il permet, en écrivant des fichiers de description de compilation, de ne recompiler que les fichiers qui ont changé. C'est très utile pour compiler de gros programmes (comme Tomcat). Disponible à :

http://jakarta.apache.org/builds/ant/release/v1.2/src/jakarta-ant-src.tar.gz.

- **JAXP :** pour Java Api for Xml Processing : permet d'analyser des fichiers au format xml. Disponible à http://java.sun.com/xml/download.html.
- **Servlet api :** les classes qui implémentent l'api des servlets. Disponible a http://java.sun.com/products/servlet/download.html.

Une fois ces outils téléchargés, vous devez créer un répertoire de base pour la création d'une version binaire de Tomcat : nous utiliserons */usr/local/tomcat-dist* :

Mkdir /usr/local/tomcat-dist

Décompressez ensuite tous les fichiers téléchargés dans ce répertoire :

Tar xfvz jakarta-tomcat-3.2.1-src.tar.gz -c /usr/local/tomcat-dist

Unzip jsse-1_0_2-gl.zip -d /usr/local/tomcat-dist

Mkdir /usr/local/tomcat-dist/jakarta-ant

Tar xfvz jakarta-ant-src.tar.gz -c /usr/local/tomcat-dist/jakarta-ant

Unzip jaxp-1_0_1.zip -d /usr/local/tomcat-dist

Unzip servlet-2_2b.zip -d /usr/local/jdk1.3/jre/lib/ext

Installons maintenant les packages java nécessaires à la compilation de Tomcat. Nous avons déjà copié le package des servlets dans */usr/local/jdk1.3/jre/lib*, il faut faire de même avec les fichiers jaxp.jar et parser.jar de la version de JAXP que vous avez récupéré et avec les fichiers jnet.jar, jsse.jar et jcert.jar que vous avez récupéré dans l'archive de JSSE. Ce procédé vous évite d'ajouter le chemin vers ces fichiers dans la variable d'environnement classpath, mais vous pouvez également utiliser cette méthode.

Maintenant, il faut compiler ANT. Pour cela, allez dans le répertoire dans lequel vous l'avez décompressé et entrez :

./build.sh

La compilation devrait se dérouler sans aucun problème. Si une erreur de compilation mentionne qu'une classe ou qu'un package est introuvable, vous devriez vérifier que les .jar suscites sont à un emplacement correct (ou que la variable d'environnement classpath contient de bonnes valeurs.).

Il faut maintenant compiler Tomcat. Pour cela positionnez vous dans le répertoire dans lequel vous l'avez décompressé, et lancez :

./build.sh dist

Cette commande va construire une distribution binaire identique à celle qui est téléchargeable sur le site de la fondation Apache mais adaptée à votre configuration. Une fois la compilation effectuée, vous pouvez tester le fonctionnement de Tomcat.

Pour cela, allez dans le répertoire *bin* de votre distribution de Tomcat qui devrait être *$tomcat_home/build/tomcat/bin* si *$tomcat_home* est */usr/local/tomcat-dist*. Entrez en tant qu'utilisateur `root :

./startup.sh

Vous devriez voir une série de messages apparaître mentionnant le démarrage de Tomcat. Ouvrez ensuite un navigateur web, et demandez d'accéder a L'URL suivant : *http://127.0.0.1:8080/.* Si vous obtenez la page de garde du serveur Tomcat, tout fonctionne parfaitement.

A ce moment précis, Tomcat fonctionne en mode indépendant, ce qui n'est pas ce que nous souhaitons. Nous allons maintenant décrire son association avec le serveur web Apache de façon à en faire un moteur de servlets externe. Rassurez vous, cette partie de l'installation de l'environnement de travail est beaucoup plus simple que la précédente.

Tout d'abord, vous devez télécharger une version d'Apache supérieure à la 1.3.9, la dernière si possible (1.3.14) à l'url : *http://httpd.apache.org/dist/apache_1.3.14.tar.gz.*

Ensuite, vous devez la décompresser dans le répertoire de votre choix, par exemple */usr/local/src* comme ceci : **Tar xfvz apache_1.3.14.tar.gz -c /usr/local/src**

Ce qui devrait créer le répertoire */usr/local/src/apache-1.3.14*. Allez dans ce répertoire avec la commande : **Cd /usr/local/src/apache-1.3.14**

Et lancez la configuration du script de compilation comme ceci : **./configure -enable-module=so**, ceci va activer le support des modules chargeables dynamiquement, ce qui sera nécessaire pour la communication entre Apache et Tomcat. Une fois le script de configuration termine, lancez la commande suivante : **Make install** qui va compiler et installer le serveur web Apache dans le répertoire */usr/local/apache*. Passons maintenant à la prise en charge de Tomcat.

Configuration de mod_jserv :

Pour permettre à Apache d'utiliser Mod_jserv afin de communiquer avec Tomcat vous devez compiler le module chargeable mod_jserv.so. Pour cela, repositionnez vous dans le répertoire contenant les sources de ce module : **Cd /usr/local/tomcat-dist/jkarta-tomcat-3.2.1-src/src/native/apache/jserv**

La configuration de mod_jk :

Mod_jk est lui aussi un module permettant la communication entre Apache et Tomcat, il fonctionne donc selon le même principe que mod_jserv dont nous venons d'aborder la configuration de base. Vous devez donc suivre la même démarche pour installer ce module. Les seules modifications à apporter concernent les fichiers à manipuler. Vous devrez ainsi exécuter dans l'ordre les commandes suivantes :

Cd /usr/local/tomcat-dist/jkarta-tomcat-3.2.1-src/src/native/apache1.3

/usr/local/apache/bin/apxs -o mod_jserv.so -c *.c ../jk/*.c -i ../jk-I

/usr/local/jdk1.3/include -i /usr/local/jdk1.3/include/linux

Cp mod_jserv.so /usr/local/apache/libexec

Et inclure le fichier */usr/local/tomcat-dist/build/tomcat/conf/mod_jk.conf* dans le fichier de configuration d'apache *httpd.conf*. Redémarrez Tomcat et Apache comme précisé dans la section précédente et effectuez le même test.

2) Oracle :

Oracle est avant tout un SGBD relationnel. Sa fonction première est de gérer d'une façon intégrée l'ensemble de données d'une entreprise et de les rendre accessible à un nombre important d'utilisateurs et d'applications tout en garantissant leur sécurité, leur cohérence et leur intégrité.

a) Caractéristiques :

Oracle veille à ce que tout ensemble d'opération de mise à jour de la base de données constituant une unité logique de traitement soit exécutée dans sa totalité ou pas du tout (principe de tout ou rien). Cette unité logique de traitement est appelée transaction .Oracle permet la définition, la validation et l'annulation, de transaction.

Pour rendre plus efficace cette gestion des transactions, oracle permet également la définition de sous transaction. Cette notion de sous transaction évite d'annuler la totalité de la transaction dans les cas où seule une partie de la transaction doit être annulée. Les opérations constituant une transaction peuvent être des opérations de mise à jour concernant des données réparties entre différentes bases locales. Dans ce cas, oracle utilise le mécanisme de validation en deux phases pour assurer la cohérence des données au niveau globale.

La confidentialité des données est assurée dans oracle au moyen des concepts de privilèges, de rôles et de vues. Chaque utilisateur peut avoir des privilèges pour se connecter à la base via l'un des outils d'oracle, créer et manipuler différents objets (tables, vues, …)

Il peut se voir attribuer un privilège pour effectuer un opération sur un objet qui ne lui appartient pas, et éventuellement le droit d'attribuer ce même privilège à d'autres utilisateurs. A fin de faciliter cette gestion des privilèges, oracle a introduit la notion de rôle. Un rôle est un ensemble pouvant être attribué à un groupe d'utilisateurs ou à un autre rôle.

On peut ainsi définir une hiérarchie des rôles. La portabilité d'oracle sur une très grande variété de plates-formes matérielles et OS, la compatibilité aux normes internationales et son architecture répartie font de lui un SGBD à architecture ouverte. Oracle est écrit en une

très forte portion, en langage C et il est de ce fait disponible sur une très grande variété de plates-formes et OS.

Cette portabilité offre une souplesse supplémentaire aussi bien en phase de développement qu'en phase d'exploitation. L'environnement de développement peut être différent de celui de l'exploitation.

III Développement :

Comme nous avons déjà vu, un système de catalogage et d'indexation de données possède trois éléments de base : une méthode permettant de collecter des informations sur les données, une base de données servant à les stocker et une méthode d'accès sélectif aux données utilisant les informations.

Pour notre cas ces 3 éléments de base sont :

Une application java appelé **Araignée** pour collecter les données et les insérer dans une base de données .

Une base de données oracle .

Une servlet appelée **ServletRecherche** pour la recherche dans la base de données et l'envoi du résultat au client.

Pour illustrer le fonctionnement de ce type de système, voici un petit examen du code de notre outil de recherche.

1) L'Araignée :

Pour le cas du notre moteur de recherche, nous avons choisit de faire l'indexation des pages se trouvant dans un même répertoire racine dans le choix sera laissé à l'utilisateur au début de l'exécution.

L'application s'exécute en lui passant l'adresse du répertoire principal comme paramétre, il est placé dans la file d'attente. L'araignée commence son parcours par la lecture du contenu de ce répertoire et l'indexation des pages html/htm qu'elle détecte. Chaque fois qu'elle rencontre un sous-répertoire, elle le stocke dans la file d'attente. Quand elle finit, elle l'élimine de la file d'attente et passe au prochain. Cette action est répétée jusqu'à l'élimination du dernier répertoire. Pour extraire les informations de ces pages html, l'araignée utilise ce qu'on appelle un parseur.

Le parseur est une machine à états finis qui permet d'analyser le code html. Il lit, l'un après l'autre, les caractères figurant sur une page, en indiquant la signification de la position en cours. Par exemple, à la réception d'un caractère <, l'état signale « à l'intérieur de la balise ». Si le caractère suivant est un A, l'état devient « à l'intérieur de la balise d'ancre ». Les instructions **case** structurées permettent d'extraire les informations requises. Le code de ce type de parseur se complique rapidement, d'autant que non seulement les balises html

possèdent de nombreuses variantes, mais aussi des pages qui ne contiennent pas que du code html valide.

Grâce à Java, il suffit de joindre les opérations requises à des parties de code html. Les classes principales se trouvent dans *javax.swing.texte.html.* L'Api contient la classe **parser**, mais celle-ci ne nous concerne pas directement. Les données de la page web sont acheminées via un *delegator*, qui coordonne l'analyse. A l'appel de la méthode *parse* de *ParserDelegator*, il faut lui fournir un objet *ParserCallback*, qui contient les opérations à effectuer en cas de détection de données et de balises précises. Dans notre programme, nous avons fait dériver la classe *ParserCallback*, pour obtenir *Hparser*, où nous avons surchargé les méthodes de la classe parent pour pouvoir exécuter nos opérations.

ParserCallback contient quatre méthodes à surcharger : *handleSimpleTag*, *handleStartTag*, *handleEndTag* et *handleText*. La première est appelée lorsque le parseur rencontre une balise simple (comme <P> ou <HR>). Ici, seul <META>nous intéresse. A la détection de cette balise, nous appelons l'une de nos méthodes, en transmettant les attributs (nom et contenu) qui se trouve dans la balise.

L'extraction de la description et des mots clés est réalisée par la méthode *handleMeta.* Alors que celle qui s'occupe du texte du titre ou du lien est *handleText*.

A présent, voici les étapes suivies de leurs codes d'insertion des données dans la base :

Chargement du pilote de connexion avec une base de données oracle :

DriverManager.registerDriver (new oracle.jdbc.driver.OracleDriver());

Création d'un objet de type Connection :

Connection connexion=DriverManager.getConnection ("jdbc:oracle:thin:@adresse-ip_du_serveur:port:gnet", "non_utilisateur", "mot_de_passe")

Insertion des données collectées par le parseur dans notre base de données à l'aide de l'objet PrepareStatement :

```
PrepareStatement inserer = connexion.prepareStatement("INSERT INTO page_gnet(url,
titre, description, keywords, text_brut, text_recherche, taille, id_page)
inserer.clearParameters();
inserer.setString(1, url + tree[i].substring(repertoire_debut.length(),tree[i].length()) );
inserer.setString(2, " " + organiser(hparse.getTitle()) + " ");
inserer.setString(3, " " + organiser(hparse.getDescription()) + " ");
inserer.setString(4, " " + organiser(hparse.getKeywords()) + " ");
inserer.setString(5, hparse.getPagetext());
inserer.setString(6, " " + organiser(eleminer(hparse.getPagetext())) + " ");
inserer.setLong(7, tree[i].length()/1024);
```

```
inserer.setLong(8, ++rang);
inserer.executeUpdate();
```

2) Base de données :

Notre base de données adaptée à notre système d'indexation ne nécessite pas plus qu'une table pour héberger les données fournies par le parseur. Les champs de cette table sont :

Champs	**Type**	Taille
URL	Varchar2	**150**
TITRE	Varchar2	**150**
KEYWORDS	Varchar2	**200**
DESCRIPTION	Varchar2	**200**
TEXT_RECHERCHE	Varchar2	**4000**
TEXT_BRUT	Varchar2	**4000**
TAILLE	Nombre	**-**
ID_PAGE	**Nombre**	**-**

Voici le script utilisé pour la création de cette table sous Oracle :

```
Création De La Table
SQL>
SQL> create table gnet_table
 2  (url varchar2(150),
 3  titre varchar2(150),
 4  keywords varchar2(200),
 5  description varchar2(200),
 6  text_recherche varchar2(4000),
 7 text_brut varchar2(4000),
 8  taille number,
 9  id_page number NOT NULL);
Table créée.
Création De La Séquence
SQL> create sequence SEQ_gnet_table
 2  INCREMENT BY 1
 3  START WITH 1
 4  MAXVALUE 1000
```

```
 5  NOCYCLE;
Séquence créée.
Création Du Déclencheur
SQL>
SQL>  create trigger TRG_gnet_table
 2   BEFORE INSERT
 3  ON courage FOR EACH ROW
 4    DECLARE
 5       iCounter gnet_table.RANG%TYPE;
 6        cannot_change_counter EXCEPTION;
 7    BEGIN
 8       IF INSERTING THEN
 9          Select SEQ_gnet_table.NEXTVAL INTO iCounter FROM Dual;
10           :new.id_page := iCounter;
11
11           IF :new.id_page IS NULL THEN
12              :new.id_page := 0;
13           END IF;
14        END IF;
15        IF UPDATING THEN
16           IF NOT (:new.id_page = :old.id_page) THEN
17              RAISE cannot_change_counter;
18           END IF;
19        END IF;
20     EXCEPTION
21         WHEN cannot_change_counter THEN
22            raise_application_error(-20000, 'Cannot Change Counter Value');
23     END;
24  /
Déclencheur créé.
```

3) ServletRecherche :

Lors de l'envoie du navigateur du client d'une requête de recherche, le serveur crée une instance du servlet et exécute la méthode *init()* qui va se charger d'établir la connexion avec la base de données. Puis crée un objet *Request*, et un objet *Response* .

Ensuite, il appelle la méthode *doGet()* qui reçoit de l'objet *Request* des informations concernant la requête. Cette méthode va sélectionner les sites qui contiennent le mot à chercher en utilisant des requêtes SQL et puisque aucun des moteurs de recherche n'affiche ouvertement ses critères de tri de liste parce que c'est un sujet tabou, comme nous l'avons indiqué précédemment, on ne peut pas donner plus de détails mais on aura l'occasion de défendre la performance du choix de notre méthode de scoring le jour de la soutenance.

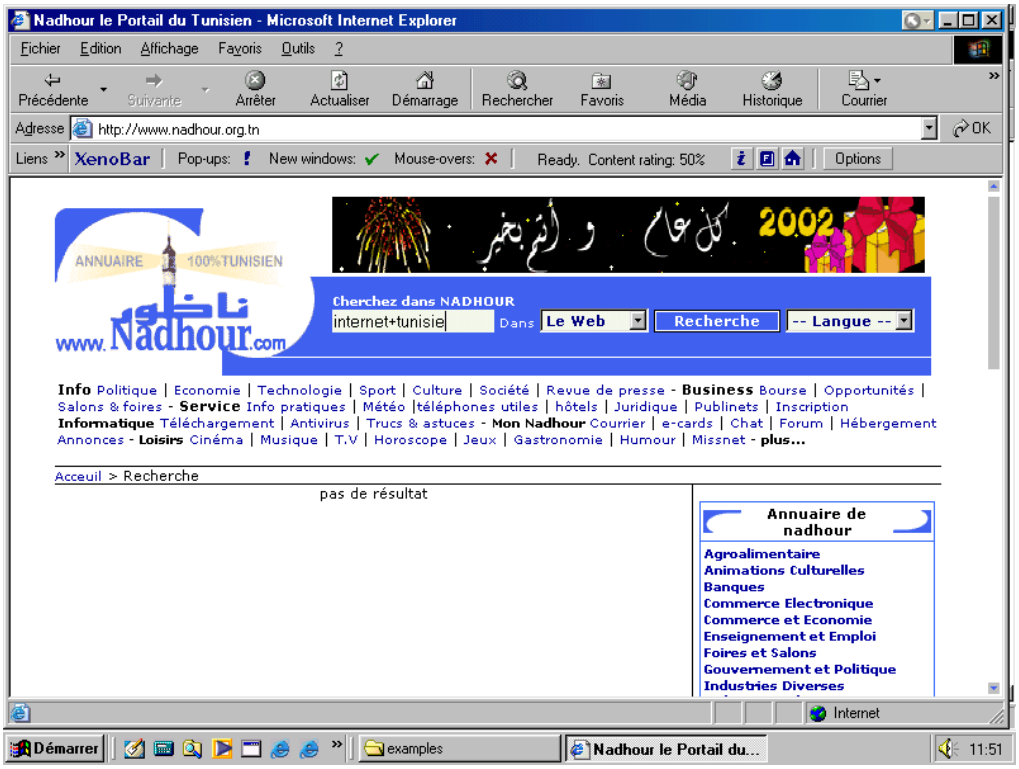
Enfin, on a utilisé un objet *PrintWriter* pour envoyer le résultat au client sous forme d'une page html.

IV- Test et évaluation:

Certains moteurs distinguent entre les majuscules et les minuscules alors que d'autres non. Par exemple, Altavista est très sensible aux accents. Pour notre moteur, On a pu résoudre ce problème .

On a utilisé une méthode de scoring qui nous permet d'avoir un bon classement des résultats alors que Nadhour n'a aucune méthode de scoring .

Notre moteur accepte les recherches contenant l'opérateur « + ». Par exemple, « Internet+ tunisie ». Voici le résultat du Nadhour :



REMARQUES GENERALES

- Les moteurs cherchent à être exhaustifs (dans la réalité, aucun ne l'est) .
- Il existe une grande quantité d'information sur le WEB. Ce qui existait hier, existe encore peut-être aujourd'hui mais peut-être que demain il n'y sera plus .
- Il n'existe pas de répertoire complet de l'information. Aucun engin de recherche ne peut prétendre tout posséder dans ses bases de données .
- La qualité de l'information doit toujours être revalorisé. Ce n'est pas parce qu'on le trouve sur Internet que c'est nécessairement vrai .
- Il faut essayer de bien cibler le type de recherche afin d'éviter la surabondance .
- La notion de taux de précision versus taux de rappel :

  Le taux de précision est la quantité de résultats pertinents parmi l'ensemble des résultats obtenus. On doit maximiser le taux de précision dans une recherche. L'utilisation d'un ET logique dans la recherche permettra généralement d'augmenter le taux .

  Le taux de rappel est la quantité de résultats pertinents parmi l'ensemble des documents contenus dans la base de donnée. Encore une fois on doit maximiser ce taux. L'utilisation d'un OU logique dans la recherche permettra généralement d'augmenter ce taux.
- Alors, quand faut-il utiliser un annuaire ou un moteur ? Il n'est pas possible de répondre car tout dépend de votre recherche !!!
  - Pour la formulation de votre requête, gardez à l'esprit que : lorsque vous utilisez un annuaire, c'est comme si vous consultiez un catalogue de bibliothèque .
  - lorsque vous utilisez un moteur, c'est comme si vous recherchiez directement dans les ouvrages et revues que possède la bibliothèque.

- Les seuls qui doutent encore du Java sont ceux qui ne le connaissent pas. Citons en terminant **Bill Gates**, l'homme le plus riche de la terre, écrivant au sujet du Java : "This scares the hell out of me !".

CONCLUSION

L'histoire a montré que dans la vie toute chose qui stagne finit par mourir. L'informatique ne fait pas exception à cette règle. Puisque tout produit quoi qu'en soit son importance, s'il n'évolue pas il va finir par être rejeté par tous et ne pourra ainsi reprendre sa place que s'il est entièrement remit en question.

Et que dire lorsque le programme est en rapport avec Internet. Là où les outils de navigation sont en perpétuelle mutation, et les défis auxquels sont confrontés les concepteurs sont légion.

D'autre part, il semble que les outils de recherche actuels ne parviennent pas à absorber la formidable croissance des sites sur le Web. L'ensemble des moteurs ne référence que la moitié des documents estimés sur Internet, représentant plus de 50 millions de sites (Source Netcraf). De plus, la plupart de ces outils n'indexeraient au total que 42% des 2,2 milliards de pages présentes sur Internet (Source Cyveillance).

Cependant, si l'on veut pouvoir un jour songer à concurrencer ces vétérans de l'Internet, il faudra bien qu'on y travail encore pour en faire un véritable portail consultable par tous les internautes. Chose qu'on espère faire avec nos partenaire de 3S GlobaNet prochainement. D'autant plus que cela sera certainement plus éducative compte tenus de l'expérience professionnel que l'on peut encore acquérir.

Finalement, même si l'on a équipé notre moteur de recherche des meilleurs techniques de développement tel que JSP et Oracle, il nous est très difficile de prétendre pouvoir un jour détrôner Yahoo, l'annuaire le plus populaire au monde, ou encore Google qui dispose de plus de 10000 serveurs à son actif.

BIBLIOGRAPHIE

www.oracle.com

www.sun.com

www.javasoft.com

www.apache.org

www.docsdunet.com

www.developer.com

www.codehound.com

www.lbb.org

www.epita.net

www.toutjavascript.com

www.webdeveloper.earthweb.com

www.php.net

www.sabee.com

www.servlet.com

www.scripts-fr.com

www.hec.ca

www.agora21.org

www.cnam.fr

www.abondance.com

www.adverline.com

www.guidewebmaster.net

www.guidewebmaster.net

www.intel.fr

www.ondelette.com