Université Mohamed Boudiaf de M'sila
Faculté des mathématiques et de l'informatique
Département d'informatique
Module : RechInf **Durée : 02 – 03 séances**

SÉRIE N°01

NB :

- *La préparation de la série au préalable est obligatoire*
- *Donner une durée de 10 – 15mn aux étudiants pendant chaque séance pour leur permettre de résoudre les exercices de la série.*

Exercice 1 : le réseau Internet : Concepts, avantages et inconvénients, services et outils

1. Définir l'internet
2. Donner une architecture approchée de l'internet en présentant ses éléments principaux
3. Donner les avantages et les inconvénients de l'internet
4. Citer quelques services de l'internet
5. Définir une adresse IP et donner un exemple
6. Que signifie une adresse uniformisée de ressources
7. Définir un protocole de transfert et donner des exemples des protocoles les plus utilisés
8. Que signifie le terme DNS (Domain Name Space)?
9. C'est quoi un web invisible ?
10. Citer quelques outils pour surfer sur internet

Exercice 2 : Data Mining, Text Mining, Web Mining

1. Définir les concepts suivants :
   - Data Mining.
   - Text Mining.
   - Web Mining
2. Citer quelques types de recherche
3. Quels sont les différents types de documents en donnant des exemples ?

Exercice 3 : IR: Concepts de base, #Modèles, Types de requêtes, Représentation

1. Comment représenter un document ainsi qu'une requête ?
2. Citer quelques types de requêtes.
3. Quels sont les modèles de SRI ?
4. Donner quelques mesures de similarité.
5. Discuter les avantages et les inconvénients de chaque modèle de codage : le modèle binaire, le modèle basé sur TF, le modèle basé sur TF-IDF
6. Comment améliorer les résultats d'une recherche ?
7. Citer les taches de la phase de pré-traitement.
8. Etablir une comparaison entre un système de RI et un système QR
9. Que signifie le terme Spam en RI

Bon Courage

Resp.Module : Said KADRI

SÉRIE N°01- CORRIGÉ-TYPE

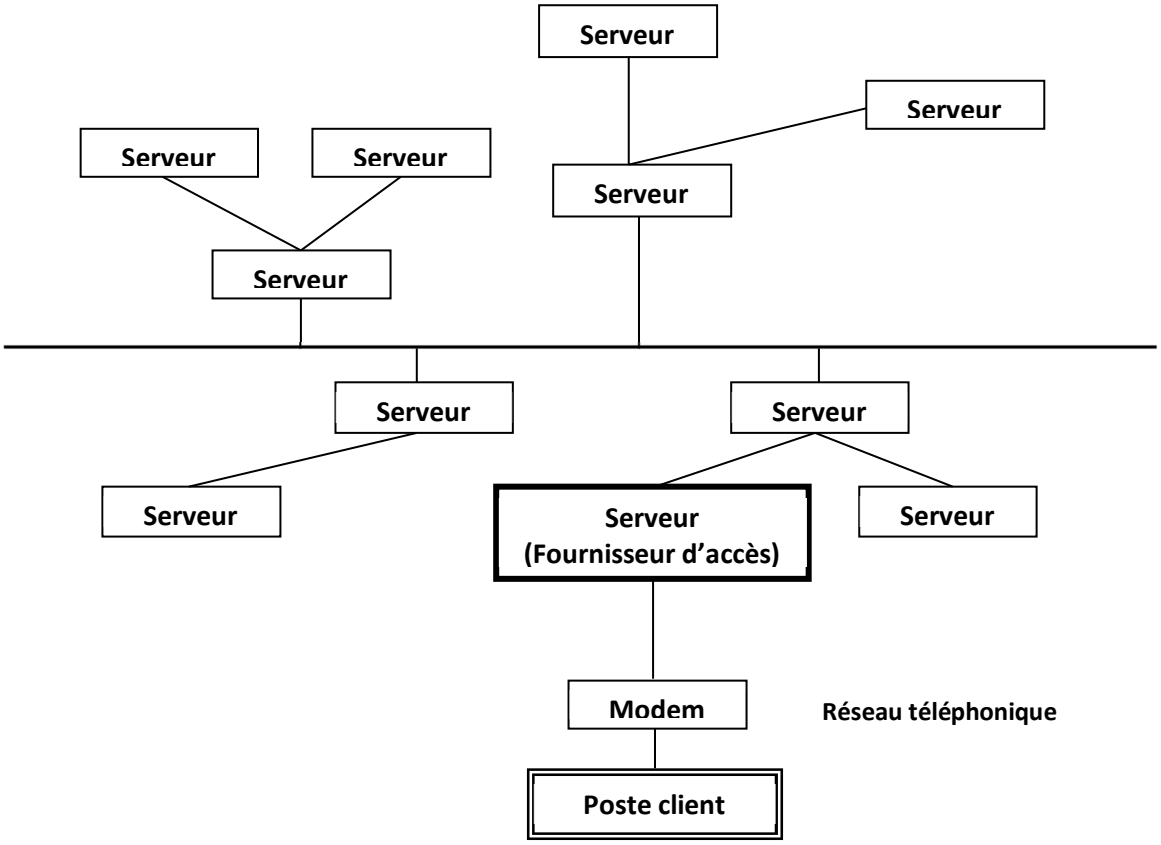
Exercice 1 : le réseau Internet : Concepts, avantages et inconvénients, services et outils

---

1. Définir l'internet

   (Prendre les définitions des étudiants et choisir la plus complète)

2. Donner une architecture approchée de l'internet en présentant ses éléments principaux



Les éléments principaux de l'architecture sont :

- Des stations de travail ou des ordinateurs personnels.
- Des postes serveurs et d'autres clients.
- Routeurs.
- Réseaux téléphoniques.
- Réseaux locaux.

3. Donner les avantages et les inconvénients de l'internet

a) Avantages :

- Exploitation d'une variété de ressources sur le Net.
- Connexion instantanée avec des sites de tout le globe terrestre.
- Possibilité d'échanger des informations de toute sorte entre personnes, géographiquement éloignés.
- Possibilité de partage: Forums de discussion, instant messaging, discuter, filmer, acheter,

b). Inconvénients :

- Connexions occupées ou lentes.
- Pas de méthode standard d'organisation.
- Des sites pourront être retirées (indisponible, non mis à jour, déplacés ….)
- Des sites peuvent avoir de contenus malicieux (virus, hackers, ….)
- Intervention possible dans la vie privée des utilisateurs (Contrôle difficile de l'utilisation des informations).
- Risque de déconnexion du monde social, « réel »….., de la vie …
- Tout le monde n'est pas connecté à Internet et Internet ne connecte pas à tout.

4. Citer quelques services de l'internet

- ***Messagerie***
- ***Le web (World Wide Web/WWW/W3)***
- ***Le service de transfert de fichiers par FTP***
- ***Le service News (news groups):***
- ***Internet téléphonie:*** Utiliser l'ordinateur comme un téléphone en utilisant:
- ***Gopher :*** est un service d'internet très riche (un site spécialisé) présentant les informations sous forme de menus successifs, et une interface très commune pour l'utilisation de grosses bases de données. Pour se connecter manuellement à un service gopher, il suffit d'entrer l'URL du site contenant le gopher sous la forme ***gopher://site:port***. souvent, les gophers sont localisés avec un port de communication spécifique représenté par un chiffre (ex : ***gopher://Mysite.calva-com.fr:250***)

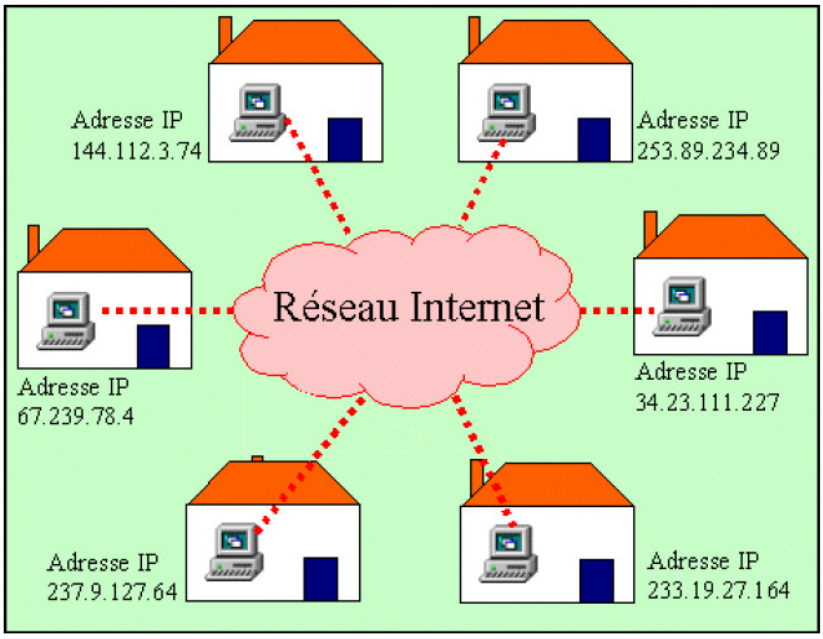
- ***WAIS (Wide Area Information Server):*** Un système de recherche d'information par mots-clés. Ou est une technologie d'interrogation de bases de données. A la différence d'ARCHIE ou de GOPHER, la recherche effectuée dans une base de données Wais porte non sur le titre uniquement, mais aussi sur le contenu des fichiers présents dans la base de données.

- ***Archie:*** c'est un système de recherche qui sert à consulter le répertoire des serveurs FTP pour chercher des fichiers. Ce système s'adresse surtout aux gens désireux de trouver des documents à transférer sur leur ordinateurs, et non de l'information lisible en ligne. (ex :le site ARCHIEPLEX).
- ***Real Audio:*** système de transfert de musique en temps réel.
- ***TelNet:*** permet à un utilisateur d'internet de se connecter et donc de d'utiliser à distance une machine comme si se trouvait face à elle.
- ***Le service chat (Instant Messaging/Chatting/IRC: Internet Relay Chat):*** permet de communiquer par message avec d'autres personnes à distance et en temps réel.
- ***Le service skype :*** permet de communiquer avec d'autres personnes en temps réel et par voie et photo (vidéo).

5. Définir une adresse IP et donner un exemple

C'est un groupe de quatre (04) nombres décimaux séparés par des points qui permettent de repérer une machine sur le réseau. Les quatre chiffres désignent successivement : ***adresse de la région, adresse du réseau, adresse de sous-réseau, et l'adresse de la machine***. Une adresse IP identifie une machine de façon unique, mais une même machine peut avoir plusieurs adresses IP.

(exemple: 113.2.28.101 représentée sur 04 octets)

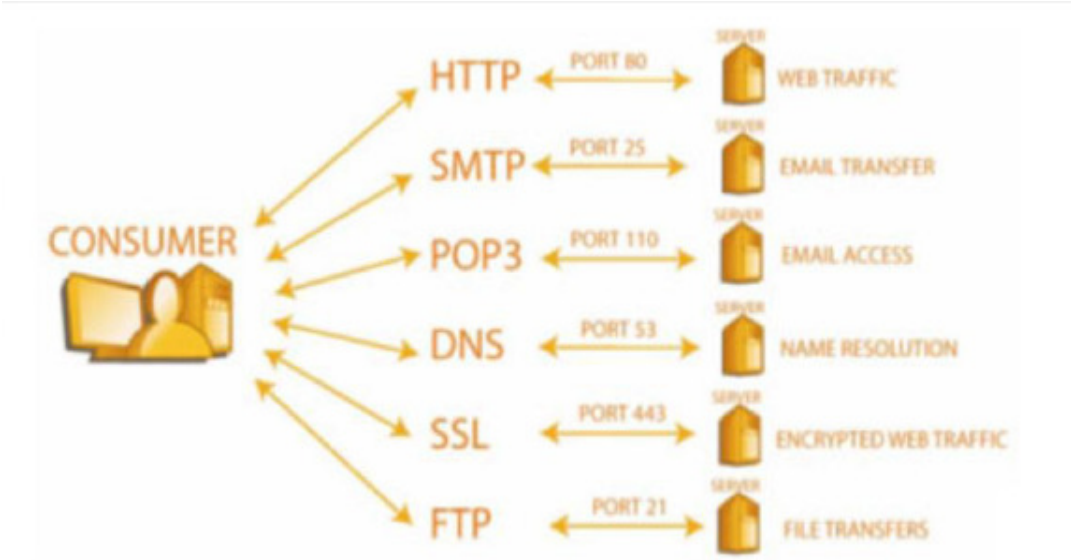


6. Que signifie une adresse uniformisée de ressources
   Une adresse uniformisée de ressources : est une adresse formée d'un ou de plusieurs mots séparés par des points, de sorte que le premier mot représente une entité (entreprise, université, personne, …) et le deuxième représente son type, son activité, le pays (exemple : Microsoft.com, Entv.dz, Mesrs.gov, …).
   On peut avoir également des adresses formées de trois mots telles que : FTP.Ibp.fr (nom du protocole, nom de l'entité, le pays), WWW.Microsoft.com
   pour le suffixe qui représente le pays, on peut avoir: Dz(Djazair), Fr(France), Uk(United Kingdoom), Ca(Canada), Sp(Spain), …
   L'adresse URL n'est pas suffisante pour localiser l'information sur le réseau internet, d'où votre ordinateur contacte des serveurs spécifiques sur internet appelés « annuaires de noms de serveurs DNS» qui l'aident à convertir cette adresse URL en une adresse IP.
   Chaque annuaire de noms de serveurs contient des tables d'informations lui permettant de convertir les noms de domaine en adresses Internet.

7. Définir un protocole de transfert et donner des exemples des protocoles les plus utilisés

   En simple terme, un protocole est un ensemble de règles et de procédures qui participent dans la réalisation d'une action sur le réseau, certains spécialistes le définit comme étant un langage de format standardisé, que deux ordinateurs (un client et un serveur) utilisent pour communiquer entre eux. Il existe des milliers de protocoles sur internet mais la plupart des utilisateurs n'en utilisent que quelques-uns, parmi les protocoles les plus courants:
   - HTTP:// pour le transfert des pages hypertexte (ex : http://site/fichier).
   - FTP:// pour le transfert des fichiers. (ftp://site/fichier).
   - MAILTO:// pour le transfert des courriers électroniques
   - NEWS:// pour afficher le contenu des groupes (transfert de New) (ex : news :groupe://serveur de news).
   - FILE:// pour l'affichage d'un fichier local (sur le disque dur du serveur).
     Exemples : * File://C:\rep\test.htm afficher la page test.htm du répertoire C:\rep du poste client
     * File://C:\ afficher le contenu de l'unité C dans l'explorateur.

- TELNET : pour visualisation de l'écran d'un ordinateur distant (ex : TELNET://site).
- Gopher : pour visualiser dans une page le contenu d'un service/site Gopher (ex : Gopher://site/fichier).
- WAIS : pour interroger une base de données WAIS (WAIS://site/Requête(paramètres)).
- STMP, POP3, IMAP : protocoles de transfert de messages électroniques sur internet
- SSL : protocole de transfert de données sécurisées (utilisé en e-business)



8. Que signifie le terme DNS (Domain Name Space)?

DNS (Domain Name System) : système de service de noms crée en 1983 permettant d'établir une correspondance entre une adresse IP et une adresse URL ou nom de hôte associé. Il a la forme d'un annuaire distribué sur internet.
Exemples : www.iut-blagnac.fr ↔ 193.54.227.200
ftp.univ-rennes.fr ↔129.20.245.1
www.w3.org ↔ 18.23.0.22
Le DNS de l'ENSIMAG ↔ 195.221.228.2
L'Amphi E de l'ENSIMAG ↔ 195.221.228.31

9. C'est quoi un web invisible ?

Les meilleurs moteurs de recherche indexent moins de 20% du web; En effet, les outils de recherche ne référencent pas :

- Les pages non html;
- Les pages non référencées;
- Les fichiers dynamiques;
- Les pages protégées par un mot de passe;
- Les bases de donnée

10. Citer quelques outils pour surfer sur internet

1. Navigateur Internet:

Un « navigateur » ou « un fureteur » ou aussi « un butineur »
Exemples : Mosaic, Internet explorer, Netscape, Mozila FireFox, Opera, Icab, google chrome, etc …

2. *Les annuaires ou répertoires*

Les annuaires d'un site Internet qui "indexe" ou "classe" dans une base de données des adresses URL selon le contenu des pages. Yahoo et La Toile du Québec sont deux exemples d'annuaires disponibles sur internet.

3. *Les portails*

Un site portail est une porte d'entrée à Internet. En général, la page d'accueil de ce genre de sites propose en plus d'un moteur de recherche, des hyperliens avec une foule d'informations et de services tels que : le courrier électronique gratuit, une sélection de salles de clavardage (de chat), les actualités, la météorologie, …etc.

Le portail est conçu pour guider les internautes, faciliter leur accès au réseau et les inciter à utiliser ce site comme point de départ pour le Web. Parmi les portails existants, mentionnons Canoë (www.canoe.ca) et Sympatico (www.sympatico.ca).

4. *Les moteurs de recherche*

Contrairement aux annuaires, les moteurs de recherche indexent eux même les adresses URL du réseau Internet. Des robots (aussi appelés Araignées) visitent chaque jour des milliers de pages et les ajoutent à leur base de données. Certains robots fouillent le contenu des pages HTML, d'autres se contentent de lire l'information contenue dans la description du site ou même seulement dans le titre de la page WEB. Google, Altavista, bing, boogle, lycos, voila, webcrawler, Hotbot, Ayne, … sont des exemples de moteurs de recherche sur internet.

Les méta-moteurs sont des sites qui utilisent plusieurs moteurs de recherche pour trouver de l'information.

- Ariane 6: est un méta-moteur francophone. (http://www.ariane6.com/)
- Métamoteur: est un moteur de recherche très complet. (http://www.metamoteur.net/)

Metamoteur.net

- AskJeeves: Méta-moteur où la recherche se fait par questions (en anglais)
  (http://www.askjeeves.com/).

Exercice 2 : Data Mining, Text Mining, Web Mining

---

1. Définir les concepts suivants:
   - Data Mining: une discipline de l'informatique qui consiste à utiliser des techniques avancées (de l'IA) pour extraire de l'information pertinente relative à un sujet bien déterminé à partir de large quantités d'informations.
   - Text Mining TM : Est une branche de DM dont la source des données est une collection de textes.
   - Web Mining WM : est aussi une branche de DM, mais la source de données est le web (des pages web). On note ici que le WM utilise les mêmes techniques de DM plus des techniques spécifiques.
2. Citer quelques types de recherche
   - Recherche de textes
   - Recherche d'images et de vidéos
   - Recherche de musique
3. Quels sont les différents types de documents en donnant des exemples ?
   - Les documents structurés (ex : BDR)
   - Les documents semi-structurés (ex : doc HTML, doc XML)
   - Les documents non structurés ou plats ou libres (ex : doc Word)

Exercice 3 : IR: Concepts de base, #Modèles, Types de requêtes, Représentation

---

1. Comment représenter un document et une requête ?
   - On représente un document/une requête en utilisant un ensemble de termes (un terme = mot, phrase, concept, racine, lemme, bi-gram, …).
   - La sélection des termes se fait soit d'une façon manuelle (par un expert humain) ou automatique (par le système).
   - La sélection des termes se fait en se basant sur des critères statistiques ou linguistiques.
2. Citer quelques types de requêtes.
   - Keyword queries.
   - Boolean queries
   - Phrase queries.
   - Proximity queries
   - Full document queries.
   - Natural language questions
3. Quels sont les modèles de SRI ?
   - Boolean model
   - Vector space model
   - Statistical language model
4. Donner quelques mesures de similarité.
   Euclidean, Cosine, Dice, Jaccard, Overlap, Okapi, Inner product, manhatan, Chebychev
5. Discuter les avantages et les inconvénients de chaque modèle de codage :
   - Le modèle binaire
   - Le modèle basé sur TF
   - Le modèle basé sur TF-IDF
6. Comment améliorer les résultats d'une recherche ?
   Sopwords removal, stemming, Relevance feedback technique
7. Citer les taches de la phase de pré-traitement.
   Sopwords removal, stemming, Frequency counts + TF-IDF
8. Etablir une comparaison entre un système de RI et un système QR
   - Question in QA vs. Query in IR
   - Exact answers vs. relevant documents for results
   - Require more complex natural language processing techniques
9. Que signifie le terme Spam en RI (un document spam)
   - First generation of spam: building documents with specific high-frequency terms, in order to appear first in the retrieval for some queries
     ⇔ Première génération de spams : construire des documents avec des termes spécifiques ayant des fréquences élevées de sorte que ce document apparait en premier rang suite à l'exécution d'une requête.
   - A doorway document is used to get highly ranked, but when accessed by a browser, it redirects the user to a spam.
     ⇔ Un document spécifique avec un rang élevé est utilisé (doorway document), mais lorsqu'il est accédé par un navigateur il redirige l'utilisateur vers un document spam.