RP – Applications 1.2.1

Application 1- Fonctionnement d'un moteur de recherche

Sur la page du moteur de recherche Google, on peut trouver à la fois des liens publicitaires et des liens non publicitaires.

1- A partir d'une requête librement choisie, on demande de donner des liens publicitaires qu'elle renvoie et d'indiquer les autres liens.

2- Quelles informations donne Google sur les autres liens ?

3- En quoi consiste l'indexation d'une page ? Pourquoi indexer des documents dans un système documentaire ?

Application 2- Syntaxe avancée des moteurs de recherche : le cas de Google

Pour Google, il est possible de faire des requêtes complexes. Lorsque plusieurs mots sont indiqués dans l'espace de recherche Google, le moteur cherche des pages qui contiennent l'ensemble de ces mots. On peut faire évoluer des types de requête :

- Pour indiquer que la recherche des mots ne doit porter que sur l'URL : allinurl:tutoriel kerio
- Pour chercher des fichier avec un certain type : filetype:pdf par exemple. Ce type de requête ne cherchera que des fichiers de type pdf.
- Pour chercher des pages qui contiennent des mots-clés hors de l'URL, du titre ou des liens : intext:connectivite limitee ou inexistante
- Trouver des pages qui contiennent un lien vers une autre page : link:http://www.elysee.com
- On peut enchainer des mots clés de recherche : | correspond à OU : Education nationale | oscar Dujardin

Proposer et tester des requêtes Google pour trouver :

A- L'ensemble des pages qui contiennent "comptabilite" dans leur url.

B- L'ensemble des fichiers pdf qui contiennent "comptabilite" dans leur url et sont de type pdf.

C- L'ensemble des ressources pdf qui traitent de la philosophie platonicienne

D- L'ensemble des pages qui contiennent bing dans leur texte et non pas dans leur URL.

E- Trouver l'ensemble des pages qui contiennent des liens vers la page Wikipedia sur la comptabilité.

A noter qu'au delà des requêtes traditionnelles possibles, les moteurs de recherche ont évolué pour prendre en compte d'autres aspects et faciliter la recherche d'information :

- prise en compte des erreurs de frappe.
- Proposition d'une liste de recherches à partir de la saisie de quelques lettres, de quelques mots.

- Prise en compte des dates dans l'actualité : mise en avant d'une série de liens d'actualité.
- Possibilité de modification de l'ordre des résultats (par pertinence, par date...)

Peut on obtenir les dernières pages modifiées sur un sujet en utilisant Google ?

Application 3 Moteurs de recherche et fonctionnement

Dans le domaine des systèmes informatiques, il existe toute une série de moyens d'accéder à l'information. Le moteur de recherche reste le moyen principal. Il existe des moteurs leaders et repérés comme tels. Les principaux moteurs de recherche actuels :

- Google
- Yahoo. Yahoo a d'abord été une collection classée de sites Web. Lors d'une première phase de son développement, utilisation de Google. A partir de 2004 : création de son propre moteur de recherche. fondé sur les technologies de ses acquisitions.
- Microsoft : Moteur MSN Search qui est le plus récent. Au début, utilisation des autres moteurs. Version beta en 2004 : Bing aujourd'hui
- Baidu : moteur de recherche chinois.

Le fonctionnement des moteurs de recherche :

- Des agents logiciels ou robots / bots explorent le web de manière itérative, ils indexent des quantités considérables de pages Web. Au vu de la taille du Web, il existe bien sûr des "stratégies"-algorithmes particuliers de parcours.
- Constitution de base de données massives qui contiennent les références des pages Web visitées.
- A partir de l'exploration, de la base constituée et du traitement des requêtes, les moteurs peuvent se distinguer.

Ces moteurs fonctionnent tous sur la base d'une innovation initialement introduite par Google : le pageRank. Celui-ci consiste à considérer qu'une page est d'autant meilleure qu'un grand nombre de pages pointent sur elle. L'innovation qui consiste à utiliser le pageRank est utilisée par la majeure partie des moteurs de recherche désormais. Cette Innovation a été introduite par Larry Page et Sergey Brin (Google) dans le cadre de leurs travaux de recherche de thèse à Stanford.

- Le principe :
  - chaque lien est considéré comme un vote pour le document vers lequel il pointe.
  - Les votes depuis les pages elle-même fortement pondérées sont plus importants.
- Le pageRank s'exprime sous la forme d'une pondération de 0 à 10 de chaque page.

De manière générale, soient A, T1,... Tn des pages Web, avec les pages T1...Tn qui pointent sur A. On note PR(B) le page rank associé à la page B, C(B) le nombre de liens sortant de la page B. Le page Rank est alors :

$$PR(A)=(1-d)+d\,(\frac{PR(T_1)}{C(T_1)}+....+\frac{PR(T_n)}{C(T_n)})$$

1- Qu'appelle t'on lien sortant ?

2- En quoi consiste l'indexation des pages Webs ? Donner un exemple.

3- Question autour du pageRank :

A- Explquer la notion de lien sortant

B- Que signifie le facteur d (voir la formule pour d=1 notamment).

C- Quel est le page Rank d'une page qui n'a aucun lien qui pointe sur elle ?

D- Que se passe t'il si le nombre de liens sortants d'une page qui pointe sur la page A augmente ? Comment peut-on l'interpréter ?

4- Comment pourrait on faire artificiellement augmenter le pageRank d'une page ? Les moteurs de recherche peuvent ils prévoir des stratégies pour s'en prémunir ?

A noter que Google n'affiche plus le pageRank associé à ses recherches, ainsi que les autres moteurs de recherche. Ils sont les seuls à disposer du pageRank final des pages qu'ils affichent. On pourra utiliser l'outil suivant pour évaluer le pageRank : http://www.pagerank.fr/

5- Tester le page rank de quelques pages et interpréter les résultats observés.

6- Pourquoi les moteurs de recherche tendent à ne plus afficher le page Rank ?

7- Comparer les principaux moteurs de recherche au travers de quelques requêtes particulières. Tendent ils à donner les mêmes résultats ? Peut-on trouver une requête telle que les moteurs de recherche ne donnent pas les mêmes résultats (ie ne donnent pas les mêmes pages Webs)

8- Quelles sont les attaques ou les problèmes dont a été ou dont peut être victime Google ? (Google Bombing et autre)

9- Combien de requêtes traitent quotidiennement les différents moteurs de recherche ? Combien d'utilisateurs ?

10- Quels sont les mots clés les plus recherchés sur Google ? Cela évolue t-il ?

Application 4 Au delà des moteurs de recherche traditionnels (ce qui suit est librement adapté du Net)

Au delà des moteurs de recherche leaders, il existe d'autres outils :

- Les répertoires ou annuaires : appelés aussi "répertoires", ils proposent une sélection de sites Web (et non de pages). Les sites sont décrits (titre, phrase de présentation, adresse URL) et organisés de manière hiérarchique en catégories et sous-catégories thématiques selon le principe de l'arborescence. Ce travail est réalisé par des humains, et non des machines, ce qui augmente la pertinence de la recherche (par rapport à un moteur) mais entraîne également un manque d'exhaustivité et des mises à jour peu rapides. http://www.aef-dmoz.org/
- Les portails : il s'agit d'une porte d'entrée sur Internet, constituée d'une **page d'accueil d'un site** mettant à disposition un **large ensemble de ressources et de services**, intérieurs ou extérieurs au site. Il existe différentes catégories de portails : généralistes ou spécialisés, professionnels ou grand public, émanant d'entreprises ou d'institutions publiques. Ex. : le portail de l'Union européenne : http://europa.eu/index_fr.htm
- Les sites et catalogues de bibliothèques : ces sites donnent accès à des ressources souvent invisibles pour les robots, mais surtout à des ressources validées, sélectionnées par des spécialistes, selon des procédures offrant des garanties scientifiques.

Pour des moteurs de recherche qui proposent des démarches particulières :

- Exalead http://www.exalead.fr/search Ce moteur français propose des mots-clés complémentaires (termes associés) et renvoie vers des annuaires, des fichiers multimédias… Grâce à cette organisation de mots-clés, le bruit documentaire (surabondance de d'informations peu ou pas pertinentes) est limité.
- Mozbot http://www.mozbot.fr Il a le même index que Google, soit environ 10 milliards de pages.
- Kartoo http://www.kartoo.com/flash.php3 (métamoteur) Les métamoteurs sont des « super moteurs » qui **interrogent simultanément plusieurs moteurs et annuaires** à partir d'une même requête. Ils éliminent ensuite les doublons et organisent généralement les résultats selon des modes spécifiques

1- Tester ces différents outils et noter leurs spécificités

2- Voir l'histoire d'Exalead : ce qui a conduit à son apparition et le débat autour de son modèle économique, ainsi que la volonté de se doter d'un moteur de recherche d'information qui soit français / européen dans l'optique d'une politique d'indépendance.

3- Reformuler le principe du méta-moteur et donner des outils correspondant.

4-

Application 5 Question ouverte

Qu'est ce qui permet de s'assurer de la qualité des informations d'une ressource Web ? Comment connaître le détenteur de telle URL ? Voir Whois

Application 6 Déterminer la qualité d'une page Web / statistiques du Web

En voyant l'URL d'une page Web, on peut rapidement dégager quelques éléments sur cette page, ainsi que sur sa qualité.

1- Comment se lit une URL ? De gauche à droite ? De droite à gauche ? On peut prendre plusieurs exemples :

A- http://www.elysee.fr

B-http://itunes.stanford.edu/

C-tempsreel.nouvelobs.com

2- Pour les principaux noms de domaine, indiquer à qui ils correspondent et quelle est l'autorité responsable de ce nom de domaine.

On peut également rechercher d'autres informations sur une page Web donnée :

3- La date de mise en ligne ou la date d'indexation sont des éléments difficiles à déterminer. Par contre, pour certains sites : il est possible de voir quelle est la date de dernière modification du site. Une fois que l'on est sur la page du site, on saisit : **javascript: alert (document.lastModified)** dans la barre d'URL. Attention cela ne fonctionnera pas sur tous les sites, mais on peut quand même avoir des informations sur certains sites : voir www.elysee.fr par exemple.

4- On peut chercher à déterminer qui détient telle URL. Pour cela, on pourra faire une requête whois qui donnera des informations sur la page que l'on est en train de consulter. On pourra par exemple aller sur le site de l'AFNIC et tester quelques requêtes Whois autour de blog ou de site de référence. A tester sur quelques sites de référence.

5- A noter que l'on peut aussi tenter de déterminer à qui appartient telle ou telle adresse IP : http://whatismyipaddress.com/ip, là aussi on pourra tester avec quelques adresses IP pour voir ce qui apparaît.

Statistiques du Web

De plus en plus de sites existent qui donnent une image du trafic et de la fréquentation des sites. Les interêtes commerciaux sont bien sûr évident, mais cette démarche est également intéressante pour l'analyse du Web et la compréhension des flux d'information. On pourra notamment voir : http://urlespion.co/

6- Tester cela avec quelques sites d'importance. Quel type d'information peut on glaner ?

A noter que le site Google Analytics est la référence, mais ce site est payant, même s'il offre certains services gratuits de

base.

7- Existe t-il d'autres sites permettant d'obtenir des statistiques de base sur le Net ?

Questions type DCG

Pour ces questions, elles peuvent demander de simplement consulter le cours ou de consulter Internet ou bien de développer une reflexion à partir des éléments du cours. Les réponses peuvent être longues ou courtes.

1- Pourquoi l'information constitue t'elle un "lubrifiant" de l'organisation ?

2- Comment appelle t-on une information résultant d'un traitement (calcul, tri...)

Application 1 (VS 5.2)

Dans le tableau ci-dessous, indiquez pour chaque besoin d'information exprimé si sa source est une obligation légale, une contrainte de coordination interne ou une nécessité liée à la préparation d'une décision

Besoin	Demandeur	Obligation légale	Coordination interne	Préparation décision
Plan de production de la semaine	Chef d'atelier			
Bilan social	Comité d'entreprise			
Liste des livraisons du jour	Service logistique			
Coût prévisionnel d'un futur projet éventuel	DG et cellule de réflexion stratégique			
Chiffre d'affaires du mois pour la déclaration de TVA	Service comptable			
Informations sur les produits d'un nouveau concurrent	Direction marketing			
Information sur divers investissements possibles	Direction de la production			