Systèmes Documentaires
et
Nouvelles Technologies

Démarche adoptée

I- Documents Web et moteurs de recherche

II- Les flux RSS

III- Recherches par supports

IV- Recherches par thèmes

I- Documents Web et moteurs de recherche

- Quelques éléments sur le langage HTML : **H**yper**T**ext **M**arkup **L**anguage
- Un système hypertexte est un système contenant des nœuds liés entre eux par des hyperliens.
- Les liens Web.

I- Documents Web et moteurs de recherche

- Un moteur de recherche est un site Web qui permet d'effectuer des recherches sur les sites Web
- Il fonctionne grâce à un travail préalable et constant d'indexation.
- Un moteur de recherche parcourt constamment le Web via un parcours récursif des liens Webs par des crawler/web-bots

I- Documents Web et moteurs de recherche

- Pour chaque parcourue : on en conserve des mots clés. On a donc construit un index : pour chaque page, on a une série de mots clés qui lui est associée.
- Lorsqu'une requête est passée au moteur de recherche, l'index est utilisé pour rendre des résultats et renvoyer des pages Web, en fonction de la présence des mots de la requête dans la page, de la rareté des mots clés...

I- Documents Web et moteurs de recherche

- Comment les pages sont classées ? D'abord en fonction des mots-clés qu'elles contiennent, en fonction de leur rareté. Ensuite, grâce au page rank, l'innovation principale de Google.
- Innovation de Larry Page et Sergey Brin 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.

- Pondération de 0 à 10 de chaque page.

I- Documents Web et moteurs de recherche

- 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 :

$$PR(A) = (1-d) + d\left(PR\frac{(T_1)}{C}(T_1) + \ldots + PR\frac{(T_n)}{C}(T_n)\right)$$

Le PageRank peut être calculé en utilisant un simple algorithme itératif, et correspond au vecteur propre p

- pages considérés), on obtient effectivement une distribution de probabilité
- Pour une explication complète :

http://www.webmaster-hub.com/publication/L-algorithme-du-PageRank-explique.html

I- Documents Web et moteurs de recherche

- Seul Google dispose du page rank final d'une page.
- Des algorithmes ont raffiné le principe du pageRank
- D'autres critères sont pris en compte pour le classement des pages au final
- Un blog sur les moteurs de recherche : http://blog.abondance.com/

I- Documents Web et moteurs de recherche

- Il est possible de préciser des types de recherche en travaillant avec Google :
  - Type de fichier,
  - Langue
  - Etc...
- Pour se faire, on définit un certain nombre d'opérateurs.

I- Documents Web et moteurs de recherche

Opérateur	Description	Valeur / attribut
intext	Recherche un mot dans le texte de la page	mot ou expression
site	Limite la recherche ou à un domaine donné	une URL
intitle	recherche le mot dans le titre d'une page	mot ou expression
inurl	recherche le mots dans l'URL	mot ou expression
filetype	recherche les fichiers avec une extension donnée	extension de fichier
inachor	recherche un mot dans la description des liens	mot ou expression
daterange	rechreche les pages ayant éété indexées au cours d'une période donné	Deux dates séparées par un tiret, sans guillemets

- Le schéma est le suivant : opérateur:valeur
- Par exemple, on peut faire une recherche :
  - FASB filetype:pdf
  - intext:"education" site:www.lemonde.fr

I- Documents Web et moteurs de recherche

- D'autres exemples de requètes :
  - Intext: theorie intext: jeux intext: algorithme
  - intext: informatique intext: comptabilite
  - Normes comptables filetype:pdf
  - Rapports publics filetype:pdf
  - Sarkozy Dati site:www.slate.fr

I- Documents Web et moteurs de recherche

- A noter que des requêtes plus complexes sont possibles en utilisant des connecteurs logiques : OU/ET etc...
- A noter qu'il est possible d'accéder une partie des fonctionnalités via la recherche avancée de Google.
- Il est possible de classer les résultats obtenus
- Il existe également des Metamoteurs (Ixquick etc...)

I- Documents Web et moteurs de recherche

- Les principaux moteurs de recherche actuels :
- Google
- Yahoo. Yahoo a d'abord été une collection classée de sites W
- Microsoft : Moteur livre search qui est le plus récent. Au dé
- Baidu : moteur de recherche chinois

I- Documents Web et moteurs de recherche

- D'autres moteurs de recherche :
  - Exalead
  - All the Web
  - ...

II- Les flux RSS

- RSS est un format XML (ce format sera étudié plus avant en cours)
- RSS :
  - Rich site summary
  - Really simple syndication

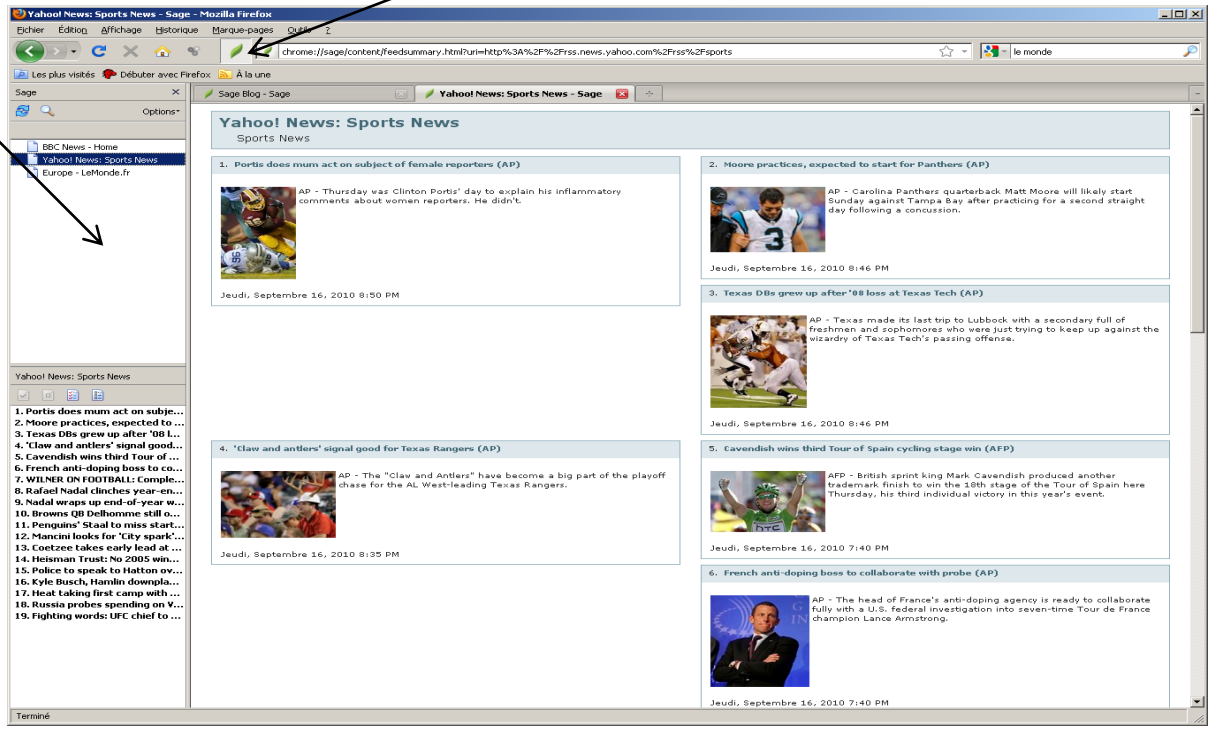
- On parle de syndication de contenu ou de fils d'information
- Sur les sites, des fichiers RSS sont disponibles qui contiennent l'ensemble des dernières nouvelles.

II- Les flux RSS

- On peut s'abonner à un ensemble de flux RSS en fonction des centres d'intérêts personnels :

II- Les flux RSS

- Des logiciels permettent de faire de récupérer du contenu depuis différents site et d'actualiser les informations en continu. Exemple de SAGE à installer sur Mozilla Firefox (le rechercher en saisissant "SAGE rss" sur Google)
- Après la relance de Mozilla, une nouvelle icône est apparue :
- Pour ajouter des éléments, on fait un clic droit, on choisit "Nouveau Marque page" et on copie l'url des fichiers rss dans la liste des flux



III- Recherches par supports

- Un Métamoteur de recherche de video : video.google.com
- Pour les images : les moteur de recherche proposent des moyens pour rechercher des images.
- Recherche de textes / livres : Google books
- Recherche de brevets http://fr.espacenet.com/

IV- Recherche par thèmes

- Pour chaque thème, on peut voir des blogs de références, mais aussi les sites d'actualité, s'abonner au newsleters.
- Informatique et gestion : 01Entreprise, 01Net, zdnet
- Sécurité http://www.securite-informatique.gouv.fr/, http://www.secuser.com/
- http://www.indicateur.com/