Web Mining: introduction

source *Modeling the Internet and the Web*
P. Baldi, P. Frasconi, P. Smyth

L. Ralaivola & C. Capponi

15 janvier 2008

Web Mining: introduction

source *Modeling the Internet and the Web*
P. Baldi, P. Frasconi, P. Smyth

L. Ralaivola & C. Capponi

15 janvier 2008

Plan

Organisation du cours

Intervenants & horaires

- A. Habrard : mercredi à partir de 10h30
- LR : mercredi à partir de 8h30

Notation

- M2 pro : TPs + exposé orienté pro + examen
- M2 recherche : exposé orienté recherche + examen

Plan

Contexte

Gisement d'information

- Énorme quantité d'information facilement accessible
- Couverture thématique large et diverse
- Tous types de données (tables structurées, textes, multimedia, etc.)
- Beaucoup d'informations sont semi-structurées (HTML, XML)
- Beaucoup d'informations reliées les unes aux autres (intra et inter-sites)
- Redondance des informations
- Données bruitées (mélange de plusieurs catégories d'informations) : contenu principal, publicités, panneaux de navigation, notes de copyright, etc.

Points de vue sur le Web

Différentes perceptions

- Web de surface et web profond
  - Web de surface : pages accessibles via un navigateur
  - Web profond : bases de données accessibles seulement via des interfaces de requêtes paramétrées
- Web des services
- Le web est dynamique (contenu, topologie)
- Le Web : société virtuelle
  - Données, informations et services
  - Interactions entre personnes
  - Organisations et systèmes automatiques
  - Communautés

Web mining : large champ d'études

Web structure mining (WSM)

Découverte de connaissances utiles à partir de la structure du web issue des hyperliens

- ranking de pages (HITS, PageRank) – LR, aujourd'hui
- crawling – M2 BDA

Web mining : large champ d'études

Web structure mining (WSM)

Découverte de connaissances utiles à partir de la structure du web issue des hyperliens

Web content mining (WCM)

Fouille, extraction et intégration de données utiles (informations, connaissances) à partir du contenu des pages web

- fouille de textes – LR, 2 séances
- utilisation de la structure des documents (cf. WSM) – AH, 1 séance

Web mining : large champ d'études

Web structure mining (WSM)

Découverte de connaissances utiles à partir de la structure du web issue des hyperliens

Web content mining (WCM)

Fouille, extraction et intégration de données utiles (informations, connaissances) à partir du contenu des pages web

Web usage mining (WUM)

Découverte de patterns d'accès et de navigation à partir de logs

- comportement de navigation – AH, 1 séance
- analyse du comportement de recherche – M2 BDA

Plan

Qu'est-ce que le Web ?

Principaux aspects

- Définition conceptuelle (plutôt que physique) : pas d'évolution temporelle trop rapide
- Le web est la combinaison de :
  - Ressources : principalement les pages web
    - Sites ou fichiers (sous formats variés)
    - Mais aussi : bases de données, bases de services
  - Identificateurs de ressources : chaînes de caractères représentant des adresses généralisées
  - Protocoles de transfert
- Toute ressource peut contenir plusieurs références vers d'autres ressources – codées par des identificateurs – : réseau d'interconnexions.
- Différence entre internet (et ses protocoles) et Web

Plan

HTML et SGML

Normes descriptives

- HTML est une application de SGML (depuis 2000 : XHTML sur XML)
- HyperText Markup Language (descriptive *vs* procedural)
- eXtensible Markup Language (XML, 1996) = sous-ensemble de SGML adapté pour le Web
- Document Type Definition (DTD) : spécifie la structure générique d'un ML (*aka* grammaire formelle)

Tag img

```
<!ATTLIST   IMG
src         %URI;     #REQUIRED   - URI of image -
alt         %Text;    #REQUIRED   - description ->
```

```
<img src="monimage.png" alt="une photo de porte">
```

Structure générale d'un document HTML

Ingrédients principaux

- Trois composantes [Raggett et al., 1999] :
  1. Information de version
  2. En-tête (balise `<head>`)
  3. Corps du document (balise `<body>`)

Structure générale d'un document HTML

Ingrédients principaux

- Trois composantes [Raggett et al., 1999] :
  1. Information de version
  2. En-tête (balise `<head>`)
  3. Corps du document (balise `<body>`)
- Dans le corps, l'élément `<a>` permet de créer les liens entre deux objets (ancres).
  - Ancre source (`<a> ... </a>`)
  - Ancre cible : URI

Structure générale d'un document HTML

Ingrédients principaux

- Trois composantes [Raggett et al., 1999] :
  1. Information de version
  2. En-tête (balise `<head>`)
  3. Corps du document (balise `<body>`)
- Dans le corps, l'élément `<a>` permet de créer les liens entre deux objets (ancres).
  - Ancre source (`<a> ... </a>`)
  - Ancre cible : URI

Ancre simple

```
<a href="http ://www.pourpre.com">Le site des couleurs</a>
```

Plan

URI, URL et URN

Syntaxe et sémantique formelles

- URI (Uniform Resource Identifier) : ensemble général de tous les identificateurs
  - Correspondance conceptuelle entre entités
  - Interprétation non ambigüe, extensibilité et complétude
  - Pages web, index dans document, fax, téléphone, GSM, broadcast, etc.

URI, URL et URN

Syntaxe et sémantique formelles

- URI (Uniform Resource Identifier) : ensemble général de tous les identificateurs
  - Correspondance conceptuelle entre entités
  - Interprétation non ambigüe, extensibilité et complétude
  - Pages web, index dans document, fax, téléphone, GSM, broadcast, etc.
- URL (Uniform Ressource Locator) : codage de l'algorithme utilisé pour accéder à une ressource

URI, URL et URN

Syntaxe et sémantique formelles

- URI (Uniform Resource Identifier) : ensemble général de tous les identificateurs
  - Correspondance conceptuelle entre entités
  - Interprétation non ambigüe, extensibilité et complétude
  - Pages web, index dans document, fax, téléphone, GSM, broadcast, etc.
- URL (Uniform Ressource Locator) : codage de l'algorithme utilisé pour accéder à une ressource
- URN (Uniform Ressource Name) : URI qui doit persister même si la ressource n'est plus disponible

URI, URL et URN

Syntaxe et sémantique formelles

- URI (Uniform Resource Identifier) : ensemble général de tous les identificateurs
  - Correspondance conceptuelle entre entités
  - Interprétation non ambigüe, extensibilité et complétude
  - Pages web, index dans document, fax, téléphone, GSM, broadcast, etc.
- URL (Uniform Ressource Locator) : codage de l'algorithme utilisé pour accéder à une ressource
- URN (Uniform Ressource Name) : URI qui doit persister même si la ressource n'est plus disponible

Quelques exemples

- URI : `sms :+33688450409`, `tv.france2.fr`, `www.lahulotte.fr`
- URL : `ftp ://ftp.tout.org/rrk/rrk261006.swf`

Plan

Les couches de protocoles

Codage des messages, règles d'échange

- Modèles de référence et TCP/IP (ISO/OSI, 1983)
  - Couche application (HTTP, FTP, SMTP, POP, Telnet, etc.)
  - Couche transport (TCP, UDP) : routage des paquets
  - Couche réseau (IP) – internet
  - Couches physiques et division des données
- DNS (adresse de tout pilote physique sur le réseau), résolution des noms via une base de données distribuée hiérarchiquement
- HTTP (protocole requête / réponse) : fonctions `GET` et `POST`

Plan

Traces

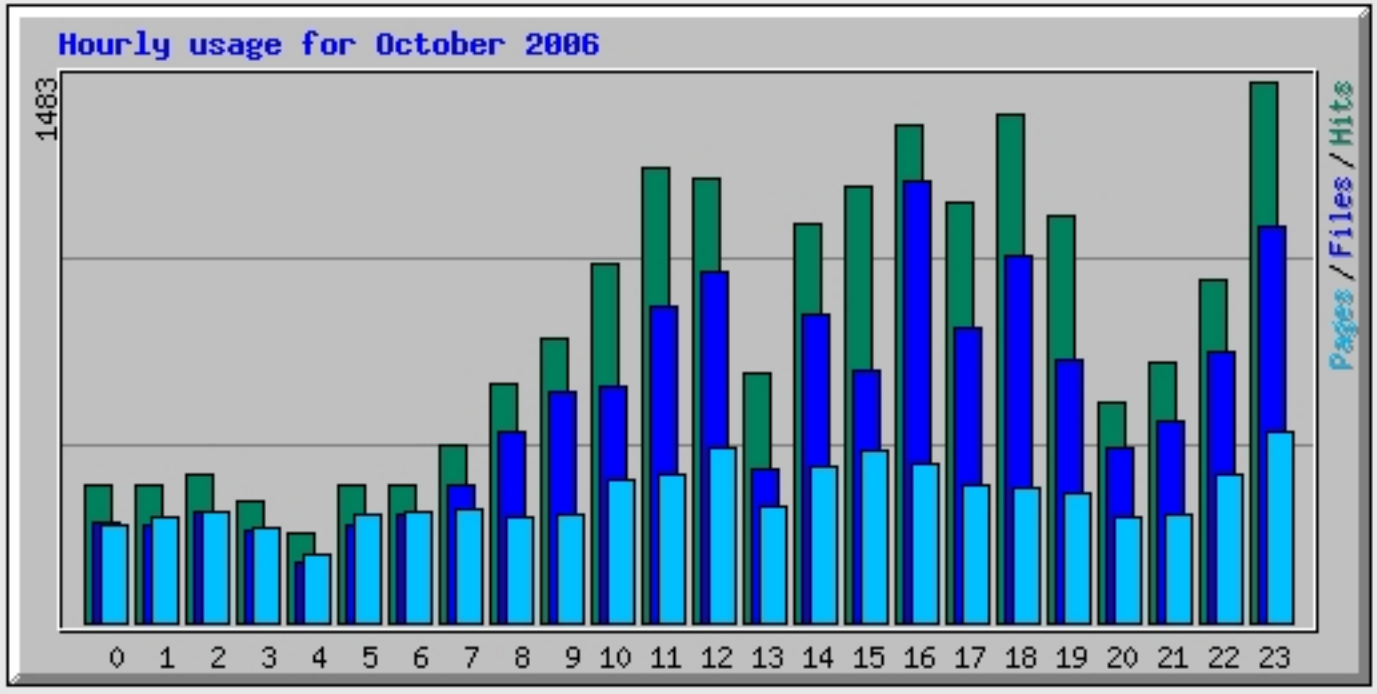
Enregistrement de l'activité d'un serveur

- Un serveur web enregistre toutes les transactions qu'il a traitées : fichier (access ou transfer) log
- Informations relatives à chaque échange
  - URL accédée
  - Adresse IP de l'utilisateur
  - Données de temps
  - Statut de l'échange, sa taille
  - Provenance de la requête (si web)
- Parsable, important pour le web usage mining.

Exemple d'entrée d'un fichier log

Field Name	Value
Requested URL	/pub/ietf/http/hypermail/1996q1/0264.html
Remote IP	65.65.193.164
Remote login name	—
Remote user	—
Time	[16/Nov/2002:23:51:23 -0800]
Method	GET /pub/ietf/http/hypermail/ 1996q1/0264.html HTTP 1.0
Status code	302
Bytes sent	315
Referrer	http://www.google.com/ search?q=%22maximum+length+of+a+url %22&hl=en&lr=&ie=UTF-8&oe=UTF-8 &start=10&sa=N
User agent	Mozilla/5.0 ... Windows NT 5.0 ...

Statististiques sur serveur



Hourly Statistics for October 2006

Hour	Hits			Files			Pages			KBytes		
	Avg	Total		Avg	Total		Avg	Total		Avg	Total	
0	18	374	1.91%	13	274	1.91%	13	263	3.16%	1122	22435	2.57%
1	18	375	1.91%	13	267	1.86%	14	289	3.47%	441	8827	1.01%
2	20	405	2.07%	15	303	2.11%	15	300	3.61%	1247	24936	2.85%

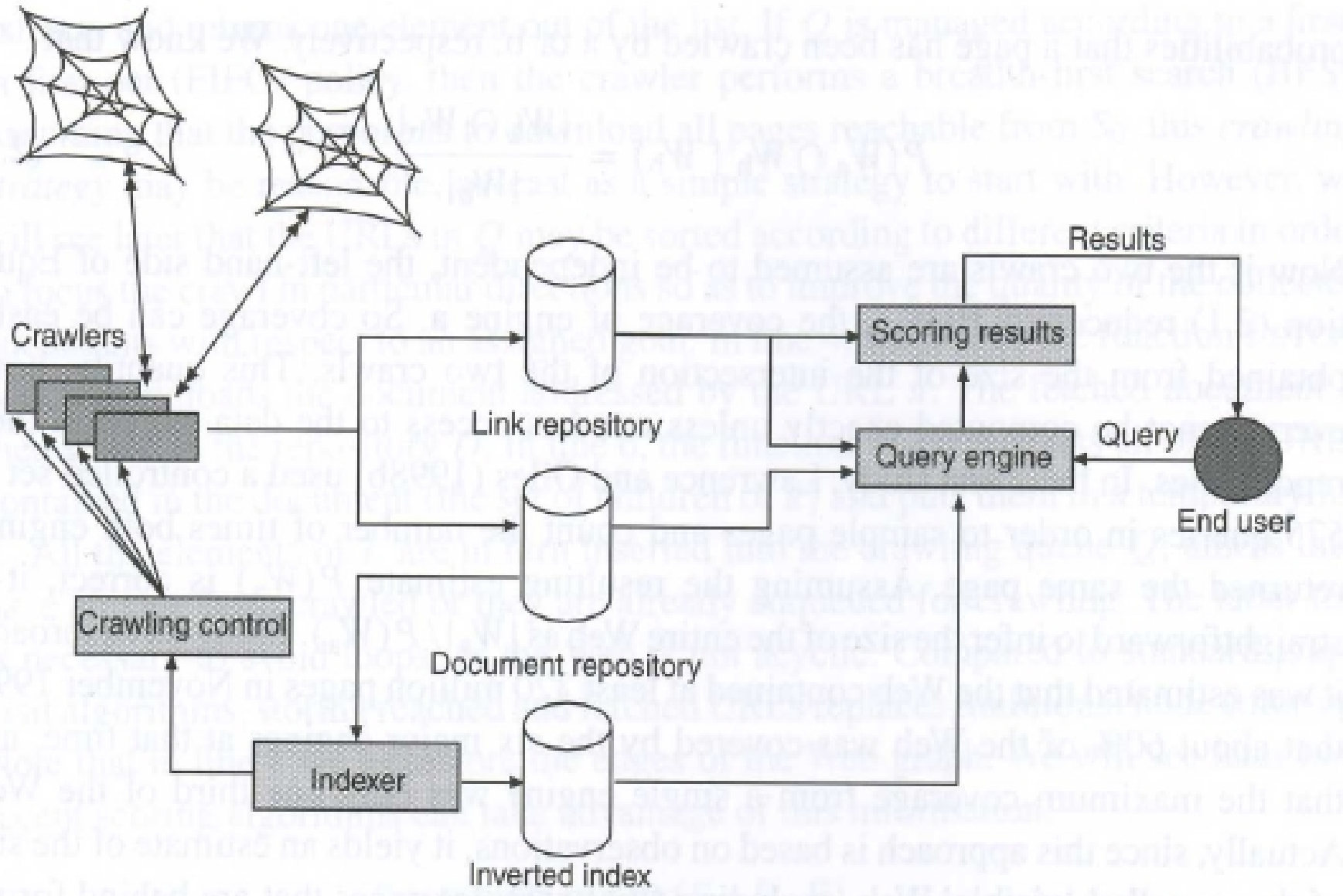
Plan

Architecture générale

Plusieurs acteurs

- En entrée, une requête (expression booléenne ou langage naturel)
- En sortie, une liste d'URLs ordonnée selon leur pertinence face à la requête initiale
- Les acteurs d'un moteur de recherche
  - Les crawlers (spiders, robots web) : programmes autonomes qui naviguent sur le web et tékchargent des documents
  - Répertoires internes indexés
  - Répertoires de liens : structure en graphe du web (actuellement, essentiel pour le *ranking*)
  - Moteurs de requêtes : traitement de la requête
  - Module de score : critères et calcul de pertinence

Composantes d'un moteur de recherche



Fonctions des moteurs de recherche

Caractéristiques

- Variétés de moteurs de recherche
  - Les moteurs généralistes (tous documents, tous utilisateurs)
  - Les moteurs spécifiques (types de documents, thématique, etc.)
- Propriétés essentielles
  - Couverture : fraction du web couverte (en pages)
  - Fraîcheur : fraction couverte (en pages mises à jour)
- Maintenir une bonne couverture et une bonne fraîcheur nécessite un crawling régulier
  - Mise à jour des répertoires
  - Défi majeur des moteurs généralistes

Couverture

Approche simple pour le calcul de couverture

- Soit W l'ensemble des pages web
- Soient $W_a \subset W$ et $W_b \subset W$ les pages crawlées par deux moteurs indépendants a et b
- Quelles sont les tailles de W_a et W_b relativement à W ?
- Soit $P(W_m)$ la probabilité qu'une page prise au hasard sur le Web soit crawlée par le moteur m.

Couverture

Approche simple pour le calcul de couverture

- Soit W l'ensemble des pages web
- Soient $W_a \subset W$ et $W_b \subset W$ les pages crawlées par deux moteurs indépendants a et b
- Quelles sont les tailles de W_a et W_b relativement à W ?
- Soit $P(W_m)$ la probabilité qu'une page prise au hasard sur le Web soit crawlée par le moteur m.

Simples probabilités

$$\begin{aligned} P(W_a \cap W_b | W_b) &= P(W_a \cap W_b)/P(W_b) \\ &= |W_a \cap W_b|/|W_b| \end{aligned}$$

Couverture

Approche simple pour le calcul de couverture

- Soit W l'ensemble des pages web
- Soient $W_a \subset W$ et $W_b \subset W$ les pages crawlées par deux moteurs indépendants a et b
- Quelles sont les tailles de W_a et W_b relativement à W ?
- Soit $P(W_m)$ la probabilité qu'une page prise au hasard sur le Web soit crawlée par le moteur m.

Simples probabilités

$$\begin{aligned} P(W_a \cap W_b | W_b) &= P(W_a \cap W_b)/P(W_b) \\ &= |W_a \cap W_b|/|W_b| \end{aligned}$$

Indépendance des moteurs a et b

$$\begin{aligned} P(W_a \cap W_b | W_b) &= P(W_a)P(W_b)/P(W_b) \\ &= P(W_a) = |W_a|/|W| \end{aligned}$$

Couverture

Approche simple pour le calcul de couverture

- Soit W l'ensemble des pages web
- Soient $W_a \subset W$ et $W_b \subset W$ les pages crawlées par deux moteurs indépendants a et b
- Quelles sont les tailles de W_a et W_b relativement à W ?
- Soit $P(W_m)$ la probabilité qu'une page prise au hasard sur le Web soit crawlée par le moteur m.

Estimation de la taille du Web

$$|W| \approx \frac{|W_a||W_b|}{|W_a \cap W_b|}$$

Couverture

Approche simple pour le calcul de couverture

- Soit W l'ensemble des pages web
- Soient $W_a \subset W$ et $W_b \subset W$ les pages crawlées par deux moteurs indépendants a et b
- Quelles sont les tailles de W_a et W_b relativement à W ?
- Soit $P(W_m)$ la probabilité qu'une page prise au hasard sur le Web soit crawlée par le moteur m.

Estimation de la taille du Web

$$|W| \approx \frac{|W_a||W_b|}{|W_a \cap W_b|}$$

Limites du calcul de la couverture

Tailles disponibles seulement dans les répertoires des moteurs

Estimation de la couverture Lawrence and Giles [1998]

Expérimentation

- Ensemble contrôlé de 575 requêtes
- échantillonnage des pages et calcul du nombre de fois que deux moteurs retournent la même page
- Si $P(W_a)$ est correcte, la taille du web $W = \frac{|W_a|}{P(W_a)}$

Estimation de la couverture Lawrence and Giles [1998]

Expérimentation

- Ensemble contrôlé de 575 requêtes
- échantillonnage des pages et calcul du nombre de fois que deux moteurs retournent la même page
- Si $P(W_a)$ est correcte, la taille du web $W = \frac{|W_a|}{P(W_a)}$

Résultats (web visible)

- 1997 : $|W| = 320.10^6$, couverture collective (6 moteurs) de 60% de W, couverture max individuelle = 30% de W
- 1999 : $|W| = 800.10^6$, couverture collective (11 moteurs) de 40% de W, couverture max individuelle = 16% de W
- 2002 : $|W| = O(10^9)$

Estimation de la couverture Lawrence and Giles [1998]

Expérimentation

- Ensemble contrôlé de 575 requêtes
- échantillonnage des pages et calcul du nombre de fois que deux moteurs retournent la même page
- Si $P(W_a)$ est correcte, la taille du web $W = \frac{|W_a|}{P(W_a)}$

Résultats (web visible)

- 1997 : $|W| = 320.10^6$, couverture collective (6 moteurs) de 60% de W, couverture max individuelle = 30% de W
- 1999 : $|W| = 800.10^6$, couverture collective (11 moteurs) de 40% de W, couverture max individuelle = 16% de W
- 2002 : $|W| = O(10^9)$

Autres études (bayesiennes) : Fienberg et al. [1999], Aslam and Montague [2001] (meta-recherche)

Processus de crawling : Simple-Crawler

Initialisation

- Les graines : collection d'URLs S_0
- Les répertoires D (documents) et E (liens)
- Liste Q des nœuds déjà visités (FIFO)

Objectif

Récupérer toutes les pages accessibles depuis S_0

Algorithme

Parcours en largeur d'abord avec détection de cycles

Limitations (extraits)

Inhérentes aux caractéristiques du Web

- Hypothèse de web statique : Q et D caractérisent l'état du crawler (sur disque).
- En pratique, arrêt forcé. Donc URLs rapatriées et URLs découvertes.
- Problème du temps de rapatriement inconnu : fetching threads et parallélisation
- Respect du serveur (Robots Exclusion Protocol : meta-tags et/ou fichiers d'exclusion)
- *Alias* : quand un site possède plusieurs URLs (DNS). Solution : la canonicalisation
- *Pièges* : quand un script CGI génère de faux documents pointant sur d'autres faux documents
- Coût en temps d'un parcours DNS (alias, ou transformation IP/nom) : multi-threading
- Le web est dynamique (contenu, topologie)

Javed A. Aslam and Mark Montague. Models for metasearch. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001.

Stephen E. Fienberg, Matthew S. Johnson, and Brian W. Junker. Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*,, 162(3) :383–405, 1999.

Steve Lawrence and C. Lee Giles. Searching the World Wide Web. *Science*, 280 (5360) :98–100, 1998. URL `citeseer.ist.psu.edu/lawrence98searching.html`.

Dave Raggett, Arnaud Le Hors, and Ian Jacobs. *HTML 4.01 Specification*. World Wide Web Consortium, Recommendation REC-html401-19991224, 1999.