EDSP

Rechercher et publier sur le Web

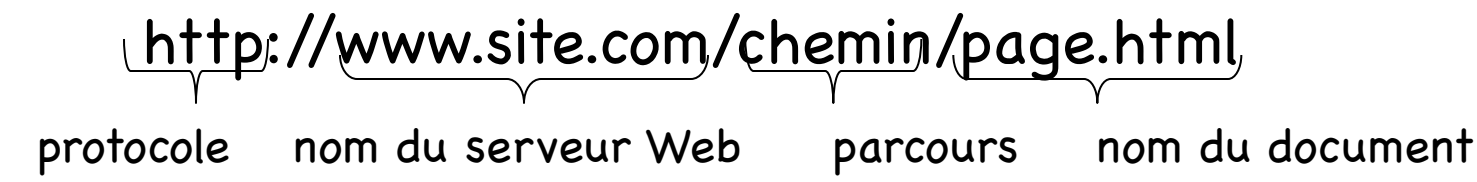
Intervenants:

Cristina Sirangelo et Françoise Tort

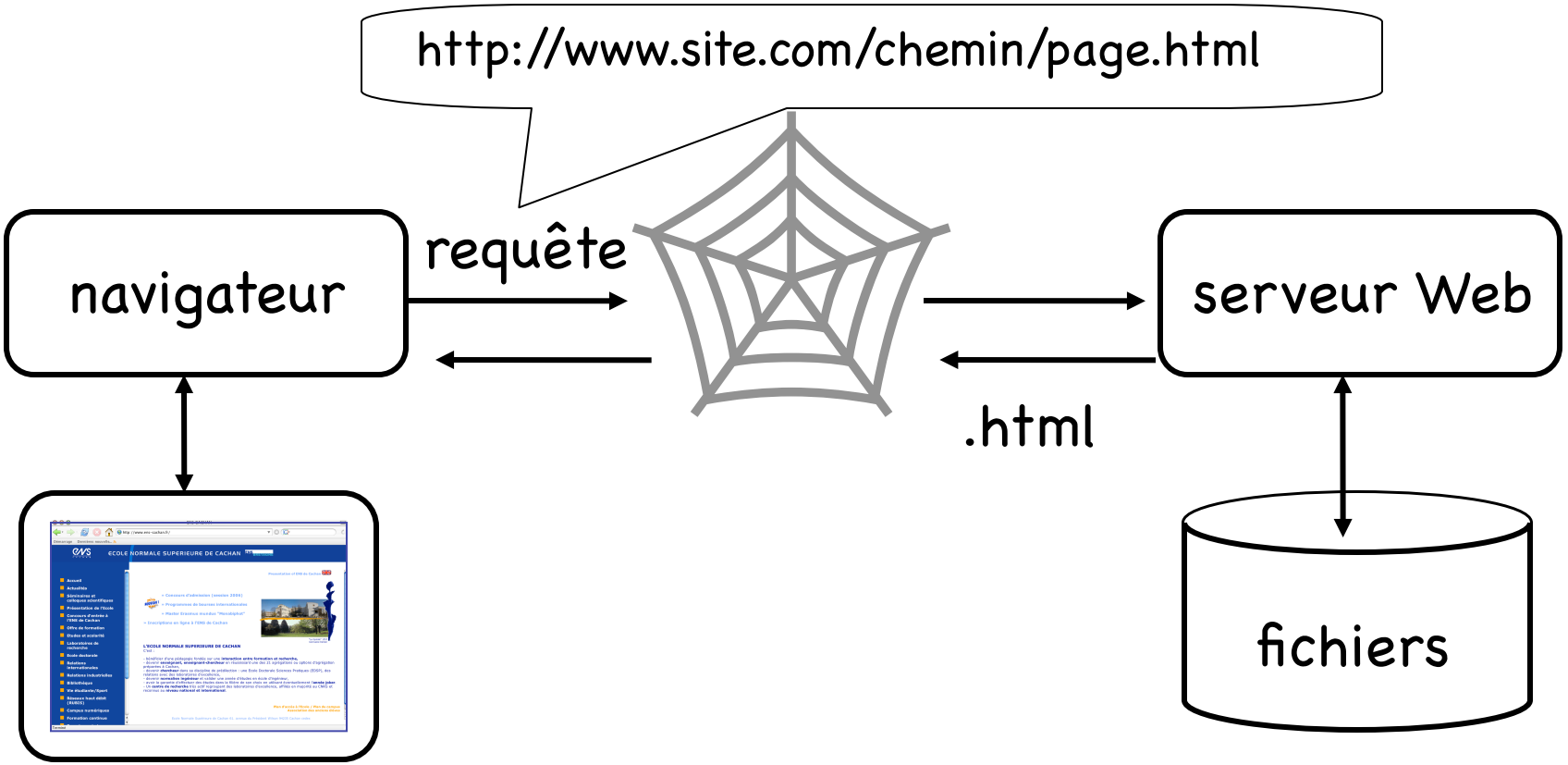
La recherche sur le Web

Le WWW

- World Wide Web (WWW): une collection de documents hypertextuels corrélés (appelés pages Web)
- HTML: langage de définition d'hypertextes
- HTTP: protocole d'échange de documents sur le Web
  - Serveur Web: serveur qui gère un ensemble de pages Web
  - Client (Navigateur) : application qui demande et affiche le contenu des pages Web
- URL (Uniform Resource Locator): identificateur d'un document sur le Web



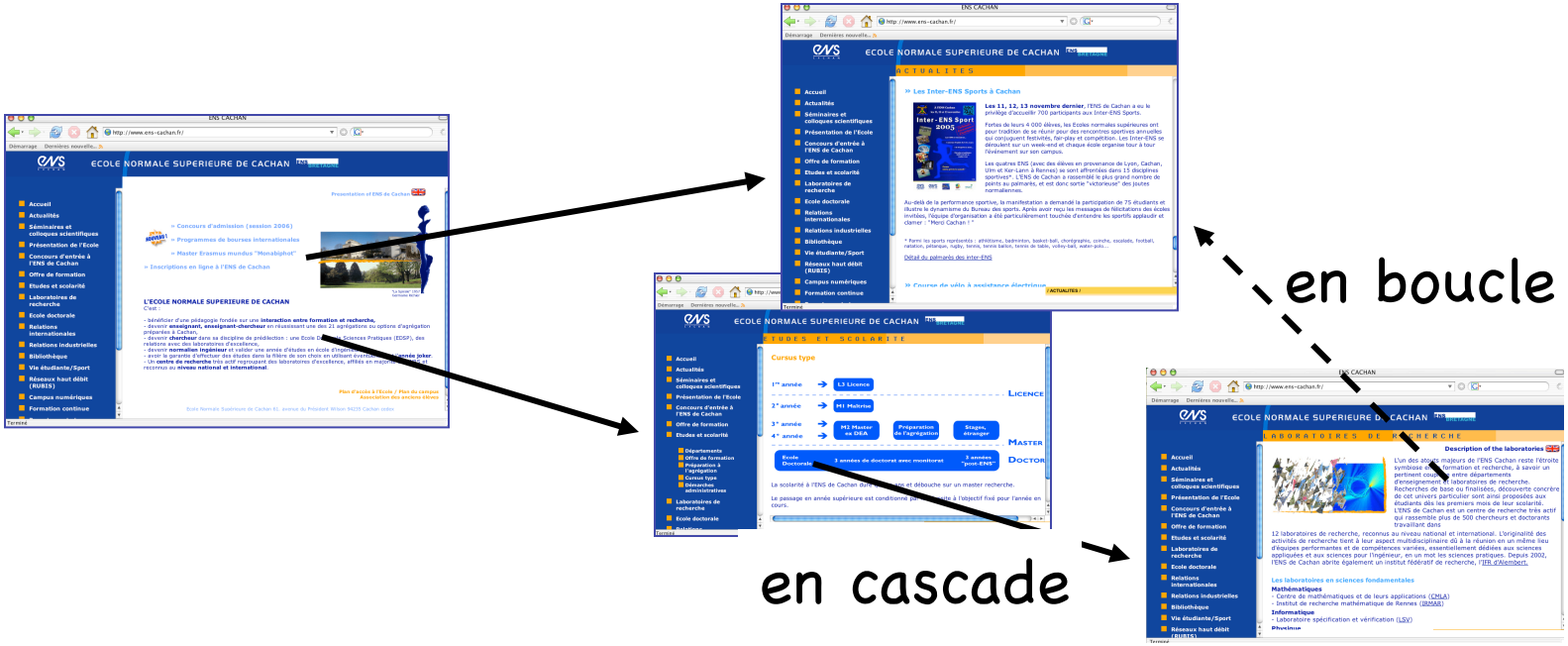
HTTP


http://www.site.com/chemin/page.html
navigateur
requête
serveur Web
.html
fichiers

Navigation dans le Web

Une page Web a une **adresse** (URL)

http://www.ens-cachan.fr/



Chaque page comporte un ensemble de pages accessibles via les **liens hypertextes**

Recherche sur le Web

- Le Web présente un volume énorme d'information
- Cette information serait inutile sans des instruments pour la rechercher
- Deux approches principales à la recherche:
  - Annuaires:

    présentent le contenu du Web à travers une hiérarchie de catégories et sous-catégories
  - Moteurs de recherche:

    effectuent une recherche sur le contenu des pages indexées, sur la base d'une requête posée par l'utilisateur

Les annuaires

- Les plus représentatifs:
  Yahoo! Directory (http://dir.yahoo.com/)
  Open Directory (http://www.dmoz.org/)

Les annuaires

- Les annuaires gardent pour chaque page référencée:
  - Le titre, l'url, un descriptif , la catégorie
- Référencement: inscription de la part du responsable du site
- Modalité de recherche:

1. (Obsolète) Descendre l'arborescence des catégories jusqu'au domaine spécifique d'intérêt présentant une liste de sites représentatifs

2. Spécifier des mots clefs: ils sont recherchés dans les noms des catégories, dans les titres et descriptives des sites référencés

Pas de recherche dans le contenu des page Web!

Limites des annuaires

- Maintenus par des êtres humains:
  - Catégories et fiches descriptives des sites définies et mises à jour par des documentalistes
  - Passage a l'échelle limité
  - Subjectivité de la catégorisation
  - Ensemble de sites indexés choisi par des humains

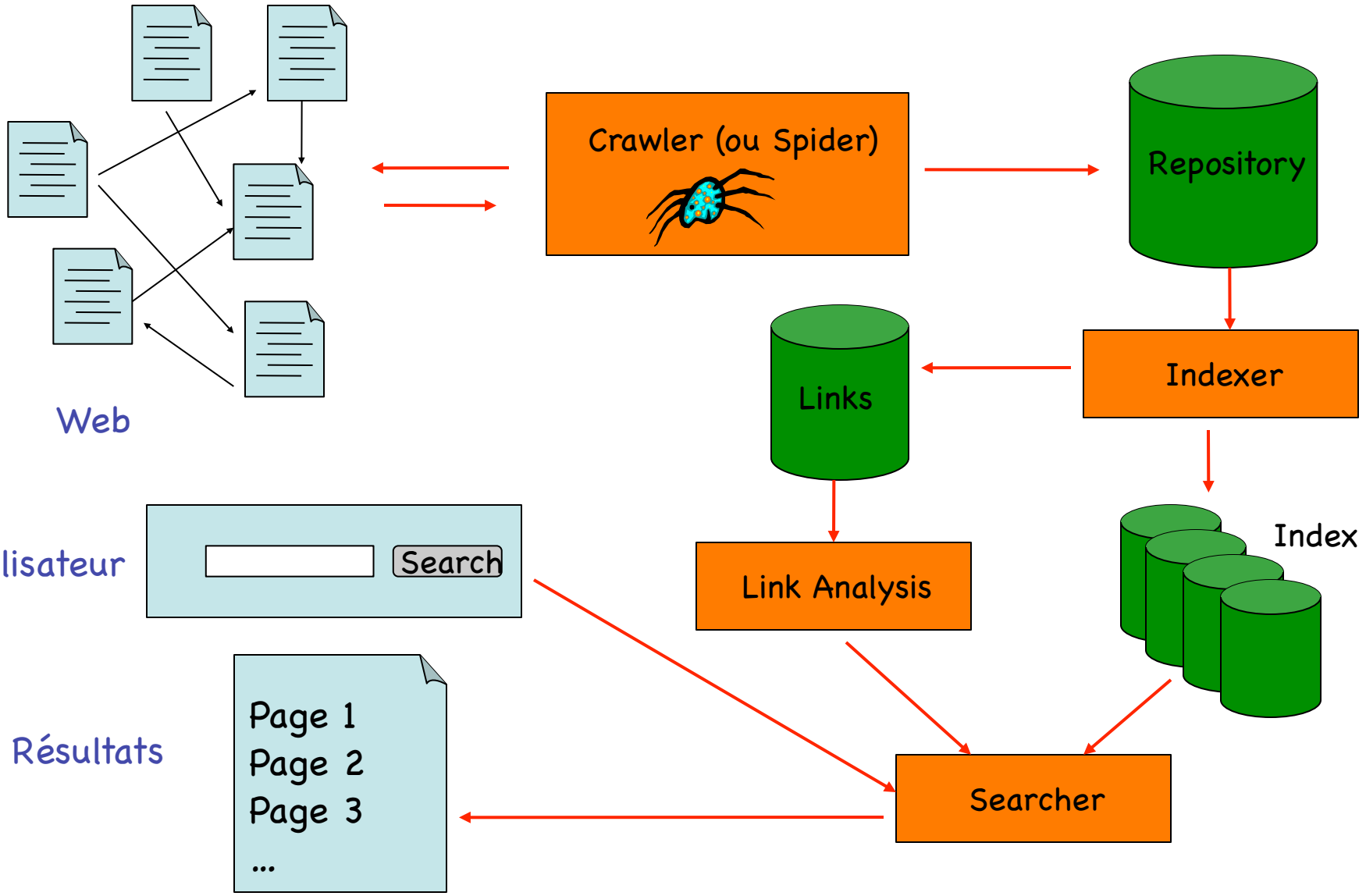
- Le contenu des pages Web n'est pas indexé

Les moteurs de recherche

Moteurs de recherche

- Parcourent le Web de façon automatique et collectent les pages Web visitées (Crawling)
- Analysent le contenu des pages collectées et construisent des structure d'accès efficaces, basées sur les mots contenus dans les pages (Indexing)
- Traitent les requêtes des utilisateurs (saisie de mots clefs) (Search):
  - À travers l'index, récupèrent les pages contenant les mots clefs spécifiées
  - Classent les résultats de la recherche sur la base de leur pertinence à la requête et de leur importance (Ranking)
- Les plus connus: Google, Bing, Yahoo! Search

Architecture d'un moteur de recherche

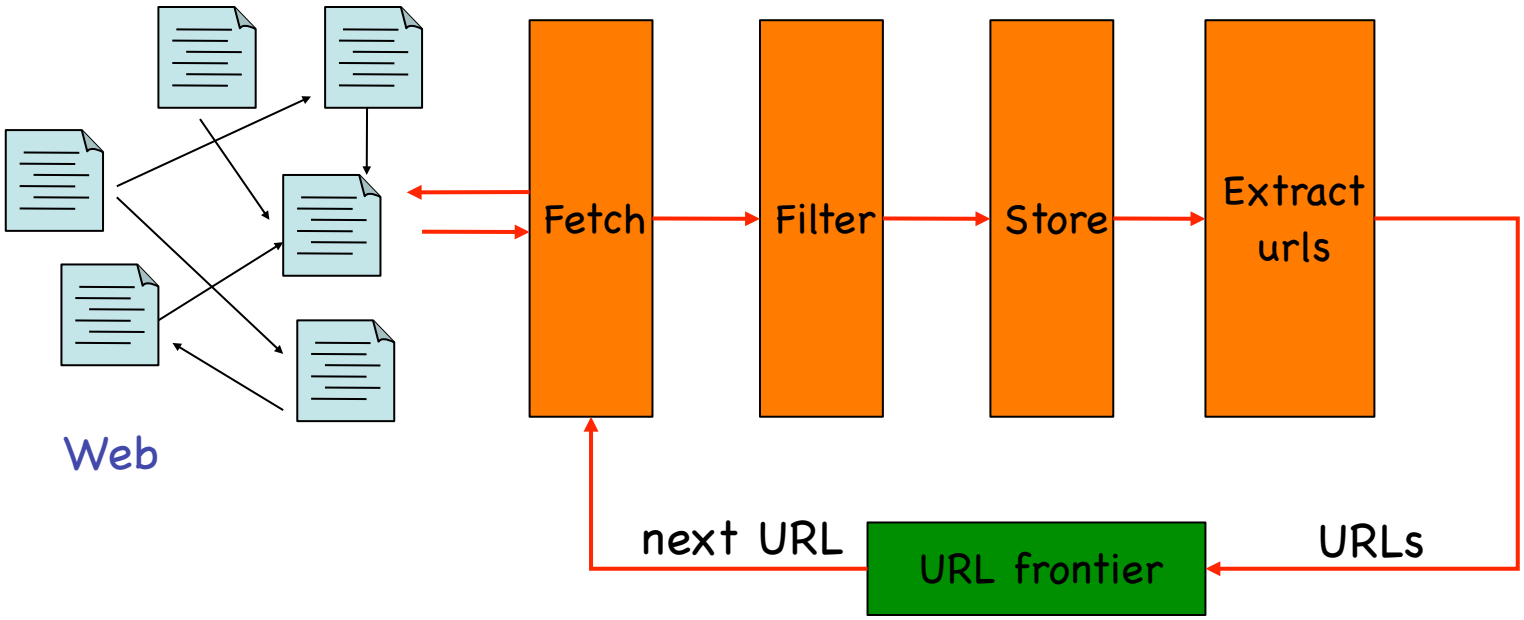


Crawler

- Aussi appelé: spider, agent, robot, bot
- Part d'une url donnée (ou d'un ensemble d'urls)
- Lit la page correspondante et stocke son code html dans le "repository"
- Détecte les liens contenus dans la page
- Pour chaque lien, continue de la même façon:

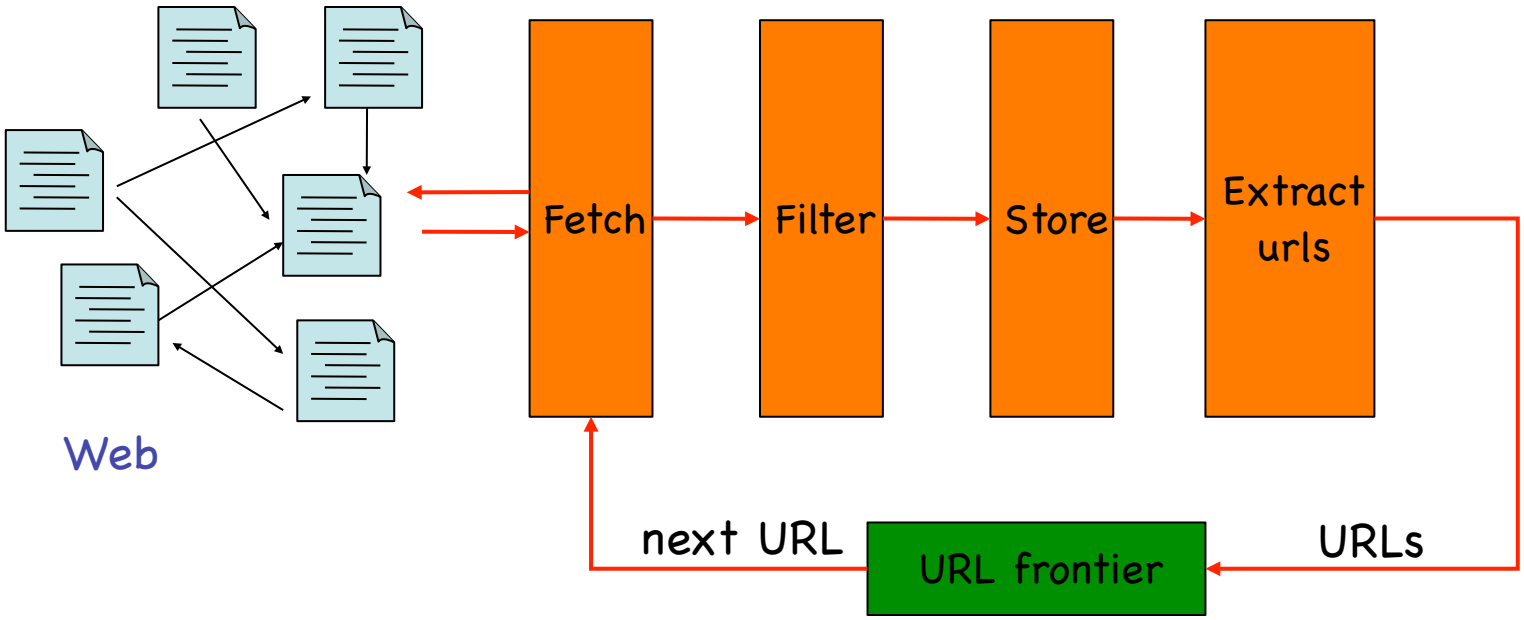
  se rend à la page correspondante, stocke son contenu, détecte ses liens etc.
- Processus de visite sans terminaison
- Les plus connus: GoogleBot, Yahoo! Slurp, MSNBot

Crawler



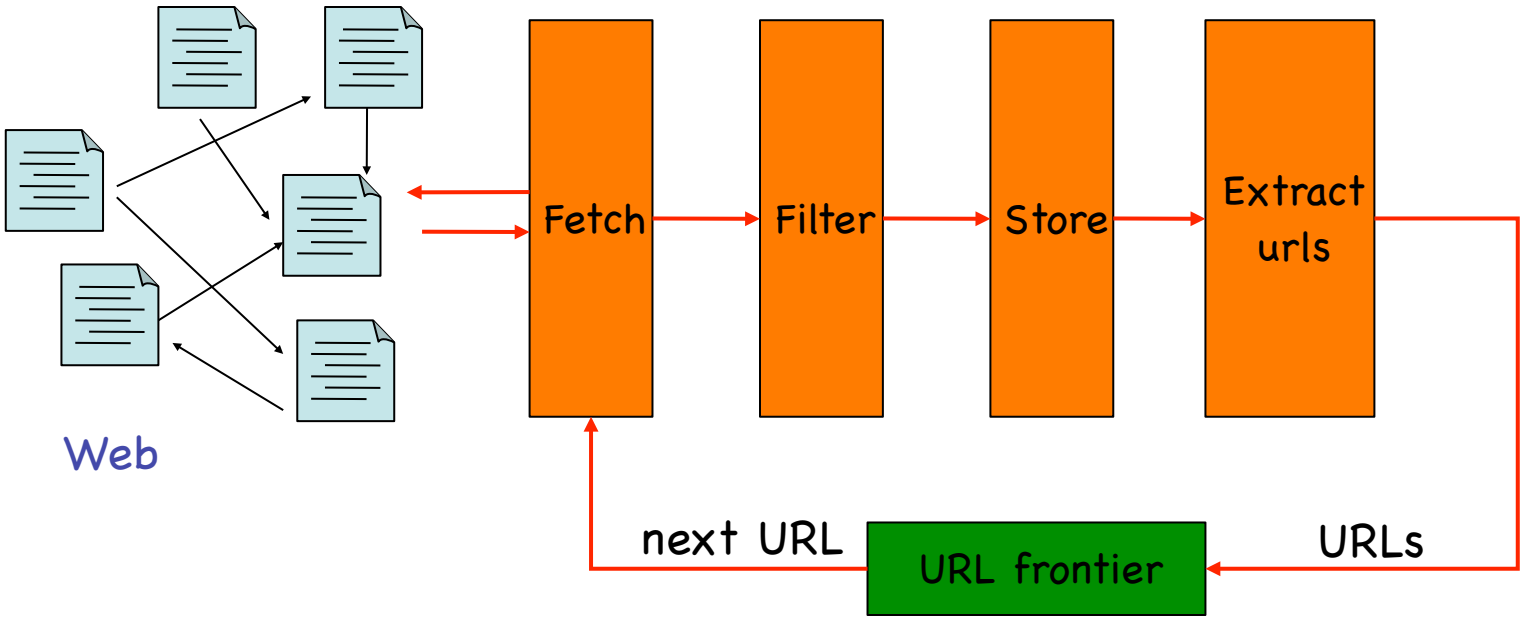
- Il est nécessaire de filtrer les pages extraites
  - Il peut y avoir des boucles dans la structure des liens du Web (réels ou "spider traps")
  - Si une page a déjà été visitée ET sont contenu n'a pas changé (suffisamment) elle est filtrée (pas visitée de nouveau)

Crawler



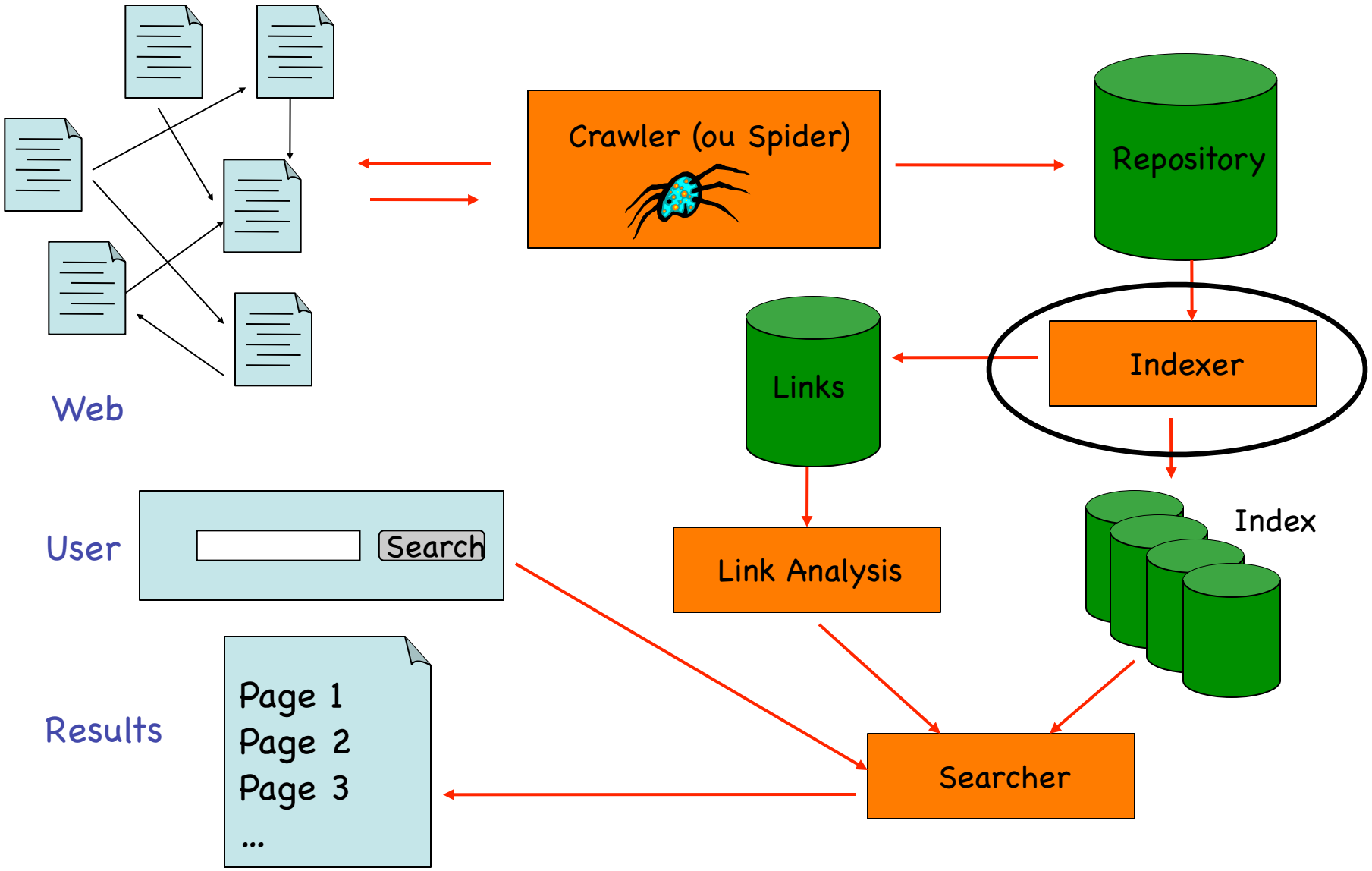
- Avant le 'fetch', d'autres tests sur l'url extraite peuvent être nécessaires:
  - Si l'url appartient à un domaine que le crawler veut exclure, il est filtré
  - Le site à la racine de l'url extrait peut avoir adopté un protocole d'exclusion des robots (fichier robots.txt)

Crawler



- Un crawl complet du Web peut durer des jours
- La fréquence de mise à jour des pages déjà visitées est variables:
  - Le pages à fort taux de renouvellement des contenus sont visitées plus fréquemment
- Dans tous les cas, un délai minimum entre deux requêtes successives au même serveur Web est à respecter

Architecture d'un moteur de recherche

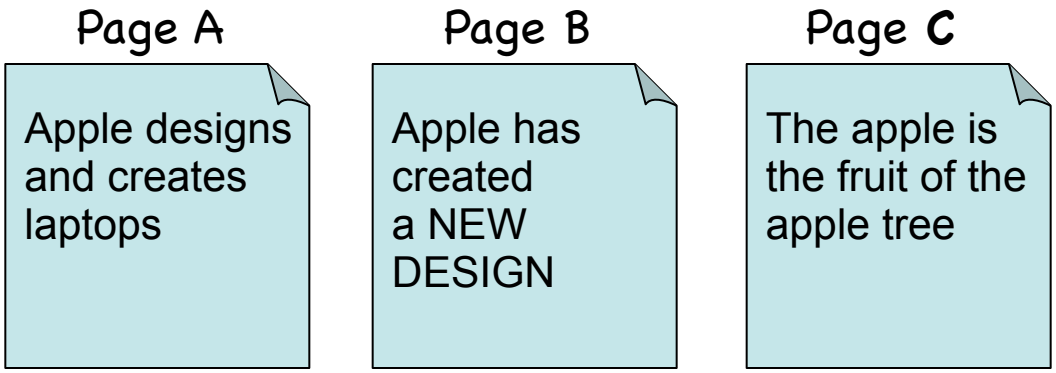


Indexer

- Après le crawling, le contenu des pages collectées est analysé par le moteur d'indexation
- L'Indexer construit trois structures:
  - Un index des pages, qui associe un id à chaque page et stocke avec l'id, l'url et des informations statistiques sur le document
  - Un index inverse, qui permet la recherche des document par mot clef
  - Une bases de données des liens, qui mémorise quelle page (id) pointe vers quelle page (id)

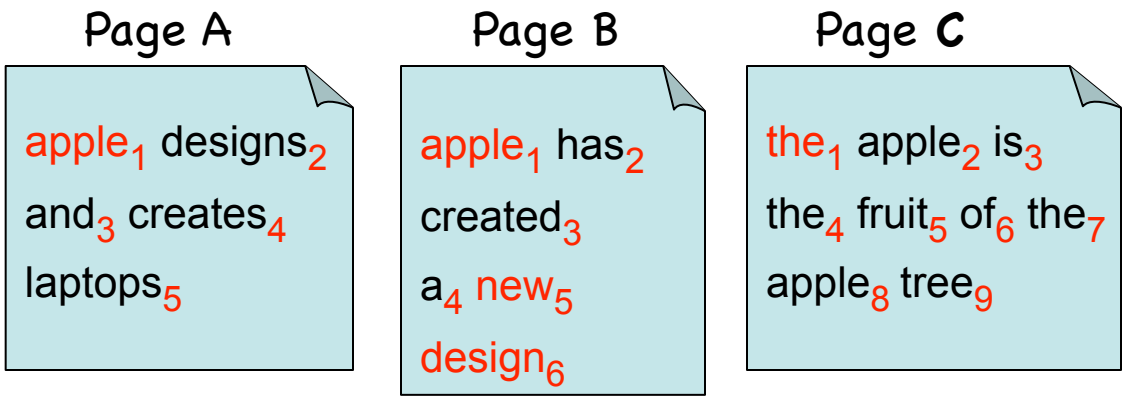
Index Inverse

- Objectif: permettre la recherche efficace des pages dans lesquelles se trouve un mot clef donné
- Index inverse: une liste ordonnée de tous le mots rencontrés dans toutes les page collectées, chaque mot étant associé à toutes le pages où il figure



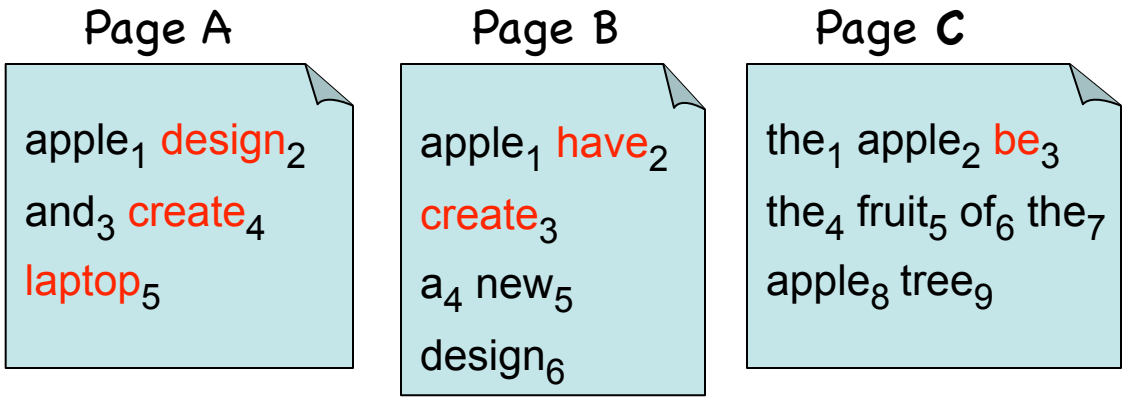
Index Inverse

- Avant de créer l'index, l'Indexer traite les pages (Preprocessing):
  1. (Tokenization) Les termes sont individualisés et associés à leur position dans le document. Dans cette phase la ponctuation est supprimée et la casse uniformisée



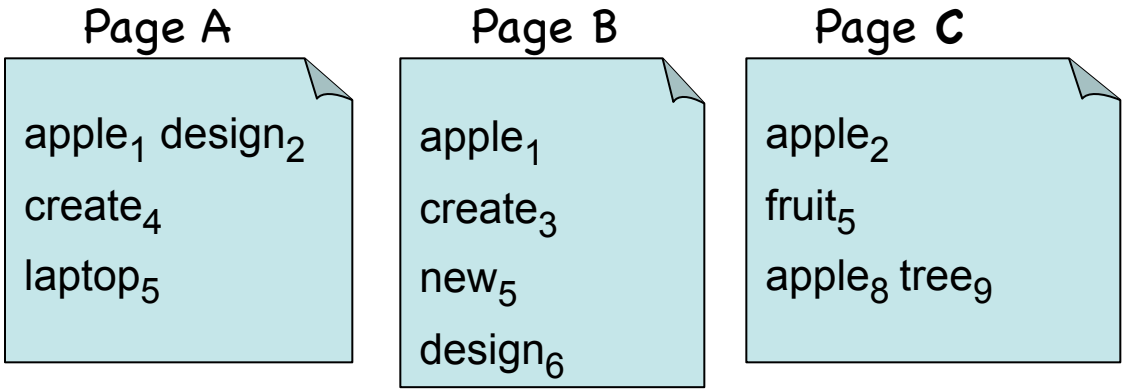
Index Inverse

- Avant de créer l'index, l'Indexer traite les pages (Preprocessing):
  1. (Tokenization) Les termes sont individualisés et associés à leur position dans le document. Dans cette phase la ponctuation est supprimée et la casse uniformisée
  2. (Stemming) les inflexions des mots sont supprimées (pluriels, genres, temps des verbes, etc.)

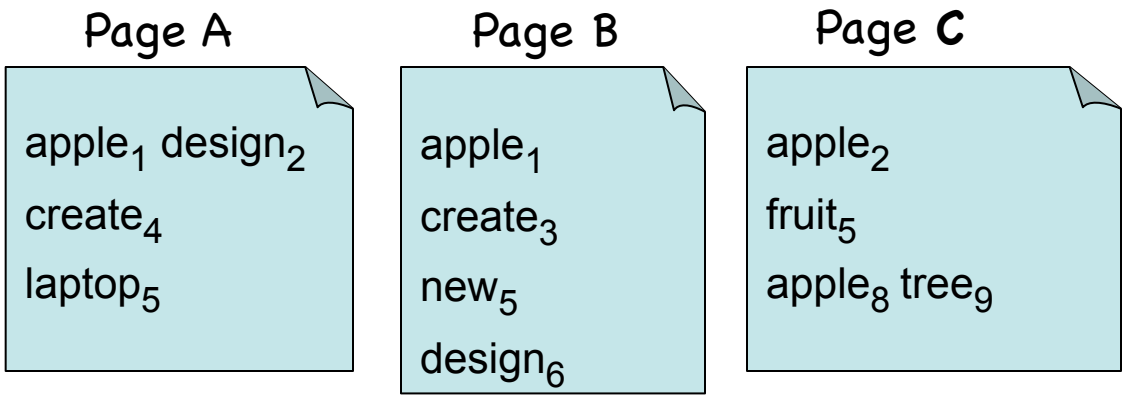


Index Inverse

- Avant de créer l'index, l'Indexer traite les pages (Preprocessing):
  1. (Tokenization) Les termes sont individualisés et associés à leur position dans le document. Dans cette phase la ponctuation est supprimée et la casse uniformisée
  2. (Stemming) les inflexions des mots sont supprimées (pluriels, genres, temps des verbes, etc.)
  3. (Suppression des « mots vides ») Les mots vides sont supprimés des documents (mots vides ou stop words: mots pas informatifs comme « the », « a » , « of », « be » en anglais)



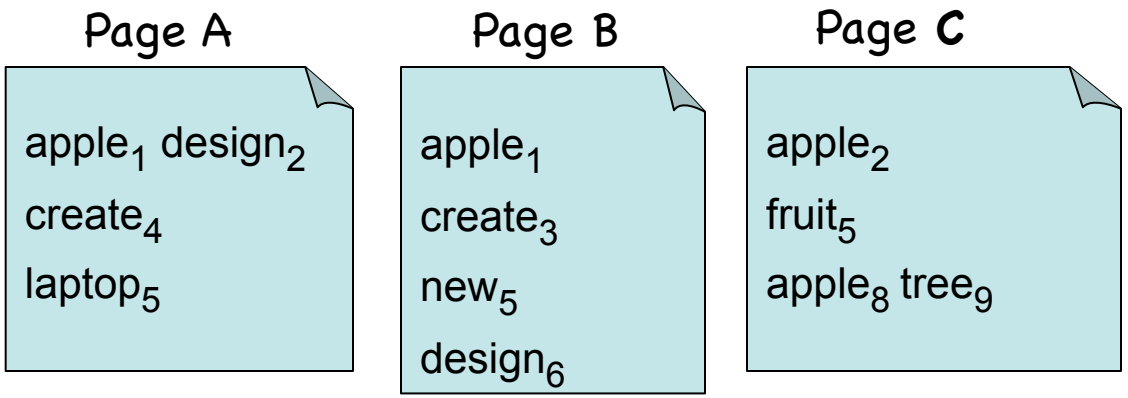
Index Inverse



Appelé 'hitlist' du mot clef 'apple'

apple	A B C
create	A B
design	A B
fruit	C
laptop	A
new	B
tree	C

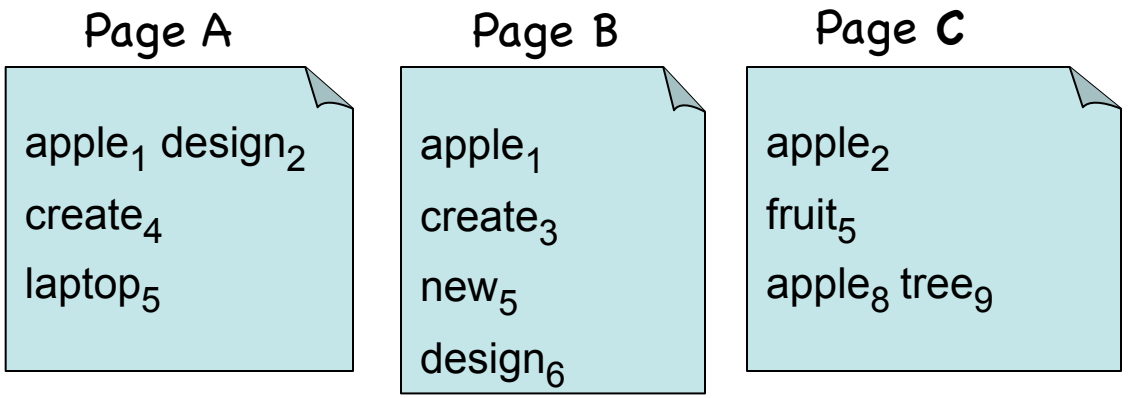
Index Inverse



Les occurrences d'un mot dans un document sont aussi stockées

apple	A/1 B/1 C/2-8
create	A/4 B/3
design	A/2 B/6
fruit	B/5
laptop	A/5
new	B/5
tree	C/9

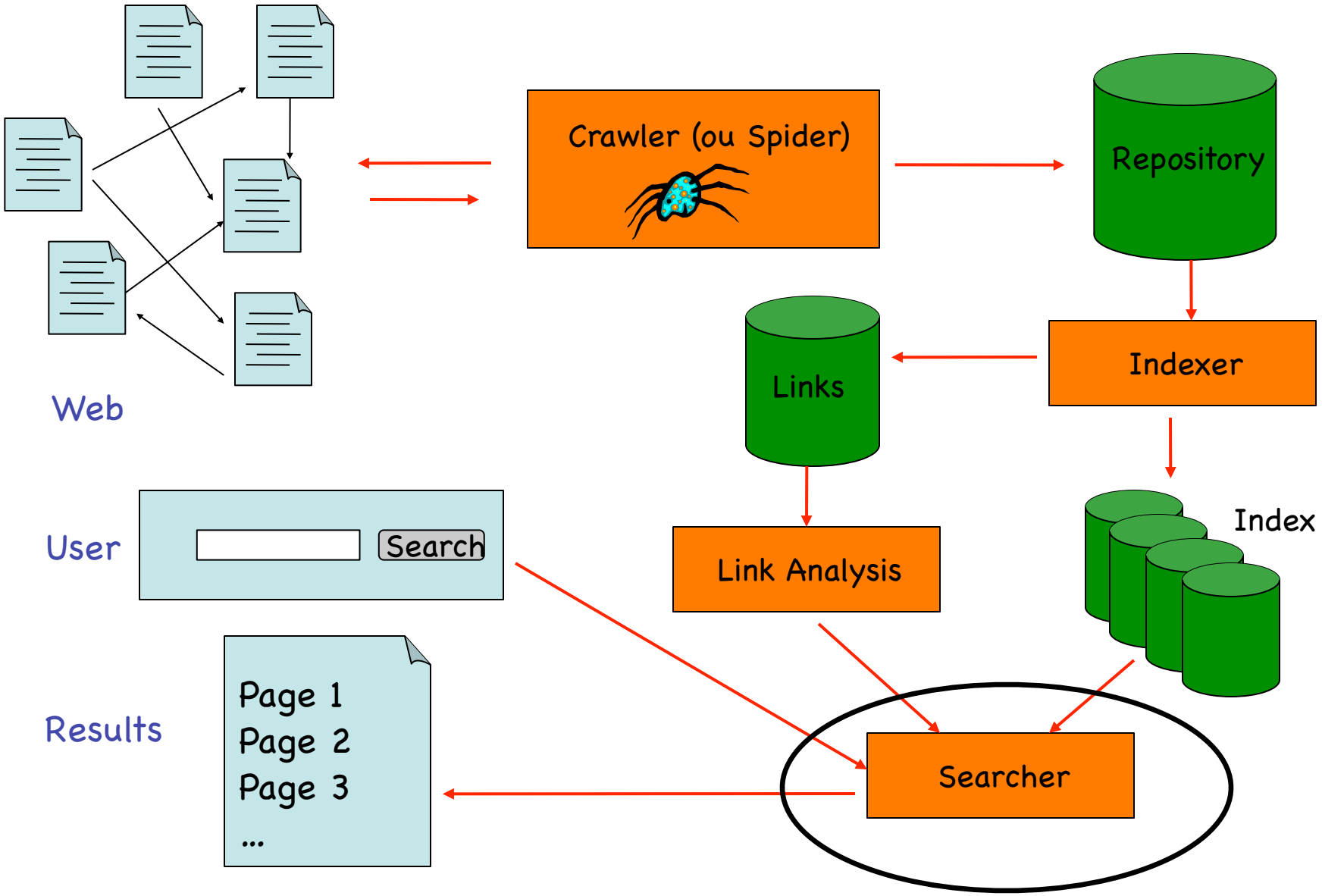
Index Inverse



D'autres informations sont stockées dans l'index inverse, pour chaque document et chaque mot :

- si le mot figure dans le titre du document, dans une url, un lien, une meta balise, ...
- si le mot est en capitale, en gras, italique, ...
- la police du mot et sa taille, ..
- etc.

Architecture d'un moteur de recherche



Searcher

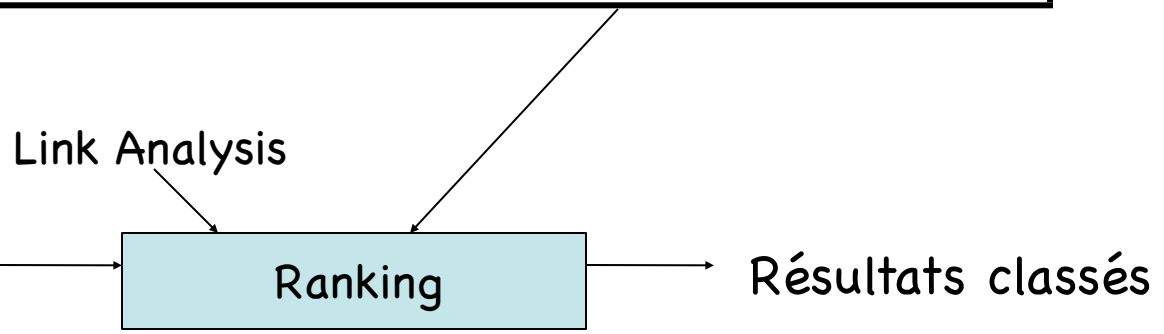
- Objectif: rechercher et classer les pages pertinentes pour une requête comportant un ensemble de mots clefs
- Technique (pour une requête à plusieurs mot clefs)
  - Cherche dans l'index inverse tous les mots clefs
  - Prend les pages qui figurent dans la hitlist de chaque mot clef
  - Classe les pages résultat (ranking):
    - Calcule un indice de pertinence de chaque page pour la requête
    - Calcule un indice d'autorité de chaque page (Link analysis: PageRank pour Google)
    - Combine les deux dans un seul indice (rank), et classe les pages résultat par rank

Searcher

- Requête: [apple, create, design]

Index Inverse

apple	A/1 B/1 C/2-8
create	A/4 B/3
design	A/2 B/6
fruit	B/5
laptop	A/5
new	B/5
tree	C/9



- Résultats: {A, B}

Résultats classés

Classement des résultats (Ranking)

Fait sur la base de:

- un indice de pertinence de chaque page résultat pour la requête
- un indice d'autorité de chaque page (indépendant de la requête)

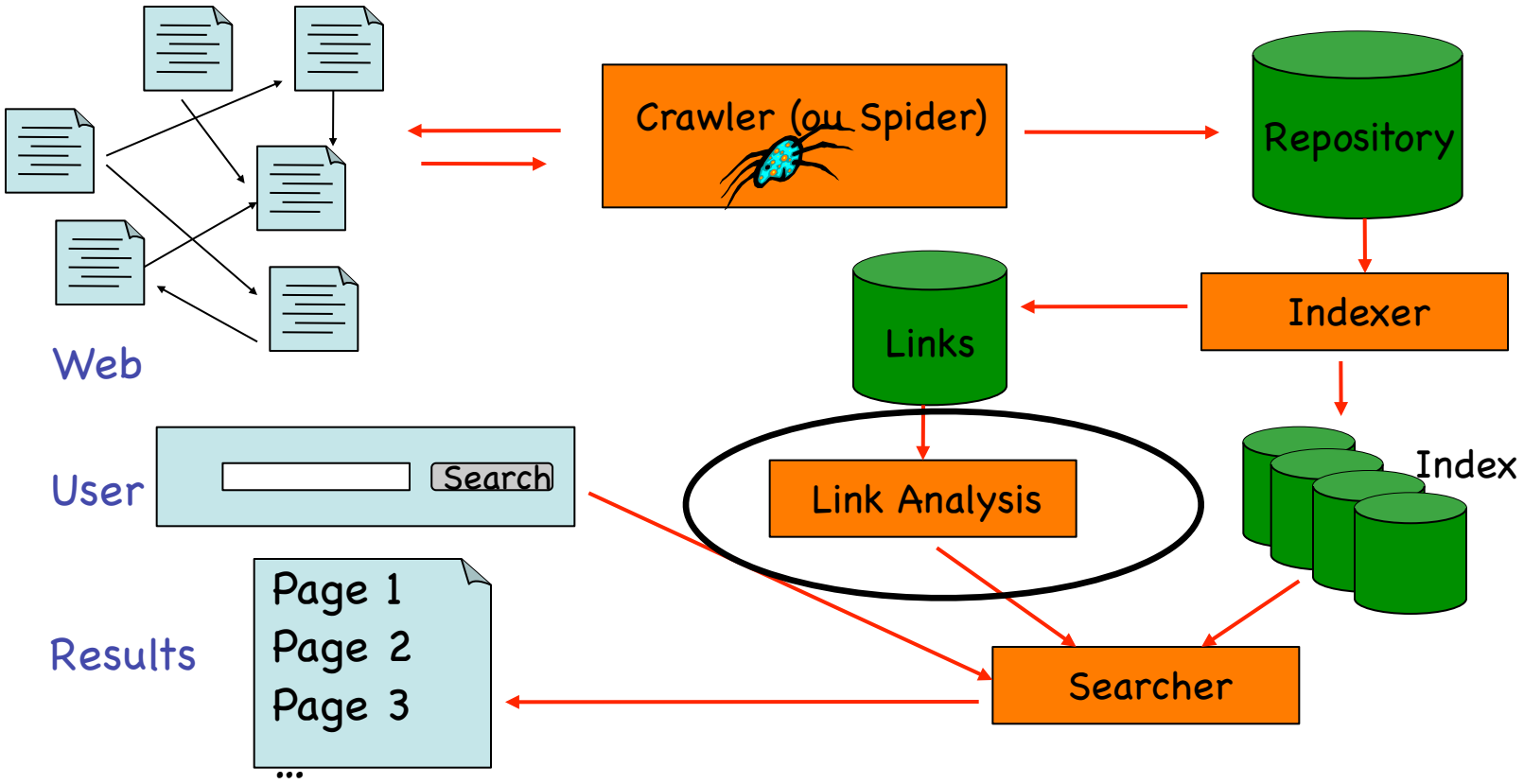
La combinaison des deux indices dans le rank final est souvent considérée comme un secret industriel!

Ranking : Indice de pertinence

- L'indice de pertinence d'une page pour un ensemble de mots clefs est calculé en fonctions de plusieurs facteurs:
  - Le nombre d'occurrences de chaque mot clef dans la page, par rapport au nombre total de mots dans la page
    (une fréquence élevée augmente la pertinence)
  - Le nombre de pages dans lesquelles les mots clefs figurent par rapport au nombre total de pages
    (un rapport élevé réduit la pertinence)
  - La proximité des mots clefs de la requête dans la page
    (les mots proches l'un de l'autre sont favorisés)
  - L'ordre des mots clefs de la requête dans la page
    (le même ordre que la requête est favorisé)
  - La position des mot clefs dans la page
    (début, titre, url etc. ont un poids de pertinence supérieur)
  - La mise en forme des mots clefs dans la page
    (gras, capitale, taille élevée de la police, etc. sont favorisés)

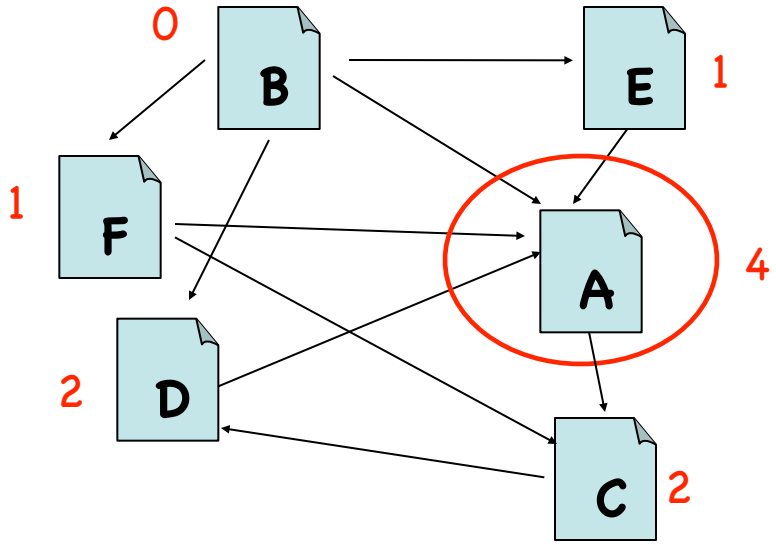
Ranking : Indice d'autorité

- Pas considéré dans les premiers moteurs de recherche
- Introduit par Google avec l'algorithme de PageRank
- Calculé par le module de Link Analysis:



PageRank

- Dans le Web l'autorité (ou importance) est conférée par les liens
- On pourrait considérer le nombre de liens vers la page comme mesure d'importance…



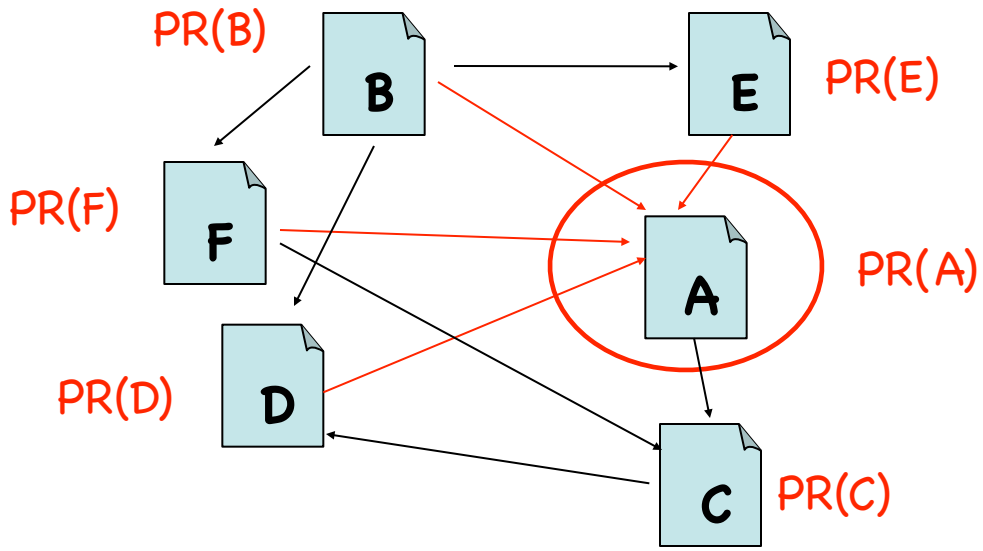
- …mais la qualité des liens doit être prise en compte

PageRank

- Qualité des liens:
  - Un lien depuis une page elle-même importante vaut plus cher
  - Un lien depuis une page contenant beaucoup d'autres liens a moins de valeur

PageRank

- Si on appelle PR(A) l'indice d'autorité (PageRank) de la page A



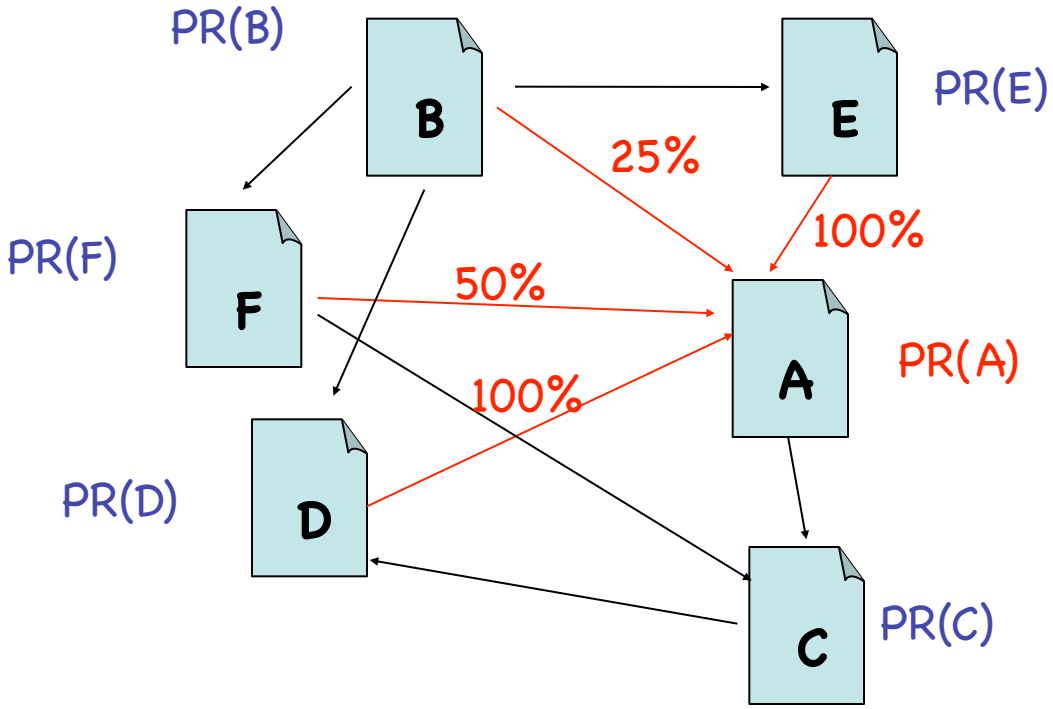
$$PR(A) = \frac{PR(E)}{C(E)} + \frac{PR(B)}{C(B)} + \frac{PR(F)}{C(F)} + \frac{PR(D)}{C(D)}$$

C(P) : nombre de liens présents dans la page P

PageRank

- Chaque page qui pointe vers A contribue au PageRank de A
- La contribution d'une page P ($\frac{PR(P)}{C(P)}$) est
  - directement proportionnelle au PageRank de P et
  - inversement proportionnelle au nombre de liens présents dans P

PageRank

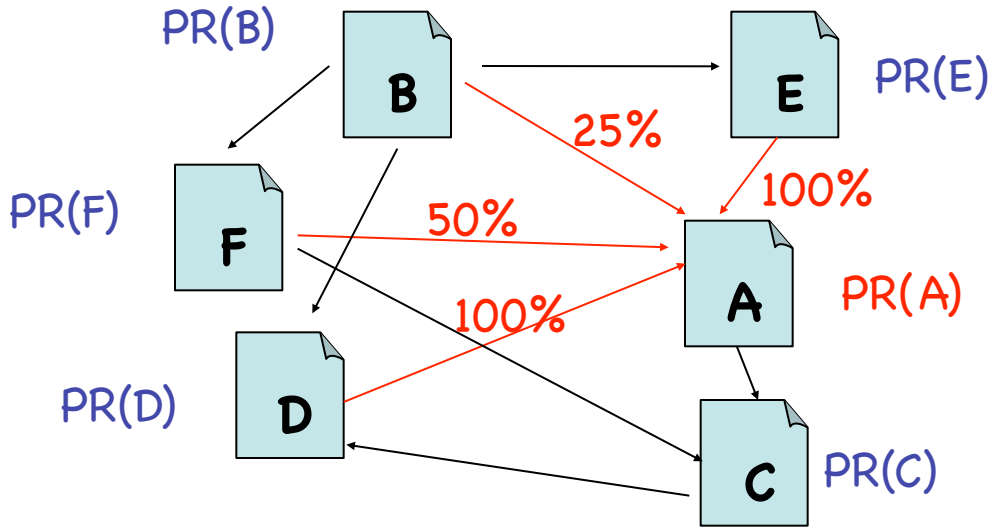


$$PR(A) = \frac{PR(E)}{1} + \frac{PR(B)}{4} + \frac{PR(F)}{2} + \frac{PR(D)}{1}$$

PageRank

- Interprétation alternative:
  - Imaginer un internaute qui visite les pages du Web en suivant les liens hypertextes
  - sur chaque page il choisit et clique au hasard sur un des liens présents (random walk)
  - PageRank d'une page P: la probabilité que cet internaute visite P pendant le random walk

PageRank



$$PR(A) = \frac{PR(E)}{1} + \frac{PR(B)}{4} + \frac{PR(F)}{2} + \frac{PR(D)}{1}$$

- La probabilité que l'internaute visite A est la somme de:
  - la probabilité qu'il visite E (PR(E)) et qu'il choisisse le lien E→A (100%)
  - la probabilité qu'il visite B (PR(B)) et qu'il choisisse le lien B→A (25%)
  - la probabilité qu'il visite F (PR(F)) et qu'il choisisse le lien F→A (50%)
  - la probabilité qu'il visite D (PR(D)) et qu'il choisisse le lien D→A (100%)

PageRank

- La formule qui définit le PageRank est en réalité un peu plus complexe
- La formule:

$$PR(A) = \frac{PR(E)}{C(E)} + \frac{PR(B)}{C(B)} + \frac{PR(F)}{C(F)} + \frac{PR(D)}{C(D)}$$

n'est pas un modèle correct du random walk :

  - Il peut y avoir des pages sans lien sortant ou sans lien entrant
  - L'internaute peut, à tout moment, décider de repartir d'une nouvelle page de départ

PageRank

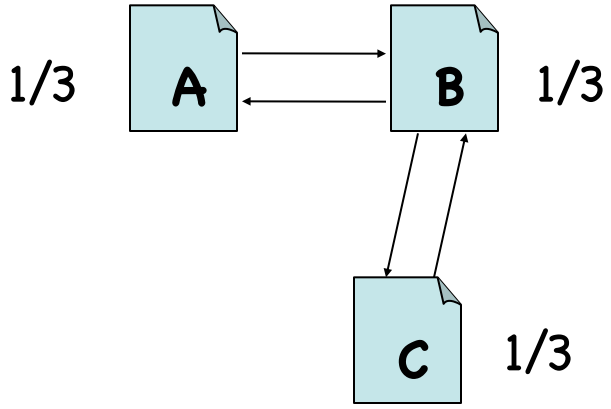
- La formule du PageRank est corrigée comme suit:

$$PR(A) = (1-d) \times \frac{1}{N} + d \times \left(\frac{PR(E)}{C(E)} + \frac{PR(B)}{C(B)} + \frac{PR(F)}{C(F)} + \frac{PR(D)}{C(D)} \right)$$

- N est le nombre total de pages
- d est la probabilité que l'internaute suive les liens
- (1-d) est la probabilité qu'il visite une page directement
- d était autour de 0.85 au lancement de Google

Calcul du PageRank

- Définition récursive: pour calculer le PageRank d'une page il faut connaître le PageRank des autres pages
- Calcul itératif:
  - Au début toutes les pages ont le même PR (l'internaute a la même probabilité d'être sur chaque page)
  - Ensuite la formule est appliquée pour calculer les nouveaux PR:



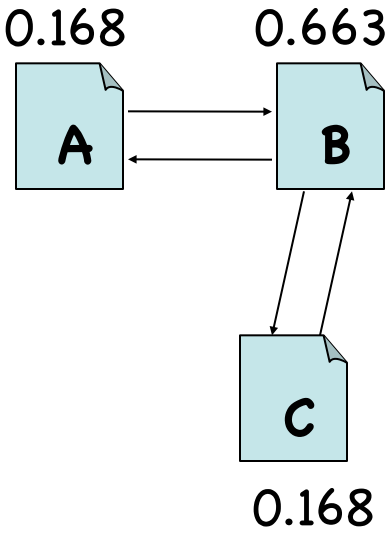
PR(A)=0.15/3 + 0.85 (1/3 × 1/2) = 0.168

PR(B)=0.15/3 + 0.85 (1/3 + 1/3) = 0.663

PR(C)=0.15/3 + 0.85 (1/3 × 1/2) = 0.168

Calcul du PageRank

- Calcul itératif:
  - Les PR des pages sont mis-à-jour
  - La formule du PageRank est appliquée sur les nouvelles valeurs



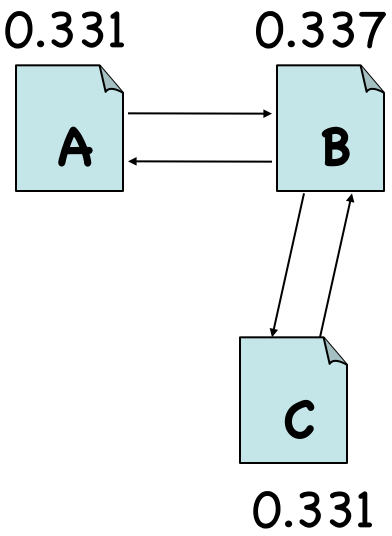
PR(A)=0.15/3 + 0.85 (0.663 × 1/2) = 0.331

PR(B)=0.15/3 + 0.85 (0.168 + 0.168) = 0.337

PR(C)=0.15/3 + 0.85 (0.663 × 1/2) = 0.331

Calcul du PageRank

- Calcul itératif:
  - Les PR des pages sont mis-à-jour
  - La formule du PageRank est appliquée sur les nouvelles valeurs



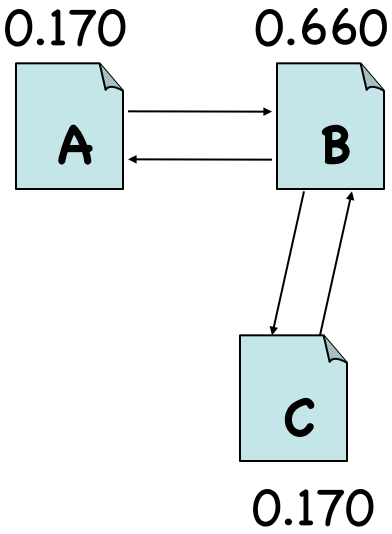
PR(A)=0.15/3 + 0.85 (0.337 × 1/2) = 0.170

PR(B)=0.15/3 + 0.85 (0.331 + 0.331) = 0.660

PR(C)=0.15/3 + 0.85 (0.337 × 1/2) = 0.170

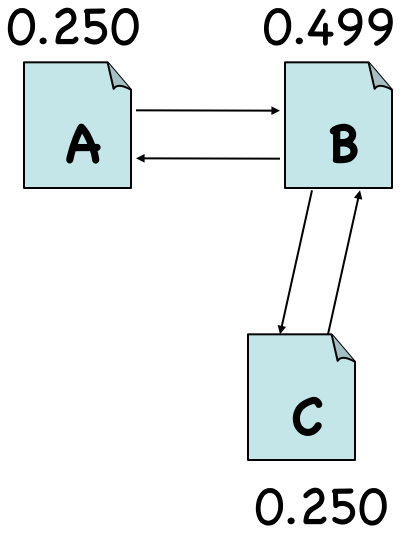
Calcul du PageRank

- Calcul itératif:
  - Les PR des pages sont mis-à-jour
  - Etc…



Calcul du PageRank

- Calcul itératif:
  - Le calcul s'arrête quand les valeurs « convergent »
    (les nouveaux PR ne diffèrent pas significativement des valeurs courantes)



PR(A)=0.15/3 + 0.85 (0.499 × 1/2) = 0.250

PR(B)=0.15/3 + 0.85 (0.250 + 0.250) = 0.499

PR(C)=0.15/3 + 0.85 (0.499 × 1/2) = 0.250

Calcul du PageRank

- Le calcul converge toujours (pour des valeurs de d < 1)

- Les valeurs stables de PR ne dépendent pas des valeurs initiales

Spamdexing

- Techniques pour augmenter de façon artificielle le PR d'un site
  - Link farms: pages contenant des nombreux liens où on peut inclure gratuitement son site
  - Systèmes d'échange de liens: ensemble de milliers de pages qui pointent vers une page demandée (parfois de façon payante)
- Aujourd'hui les moteurs de recherche ont de bons algorithmes de détection du spamdexing:
  - Les Link Farms ont peu d'influence grâce à l'algorithme de PageRank
  - Les moteurs de recherche gardent une blacklist (« liste noire ») de pages à ôter dans le calcul du ranking.

Recherche avancée

- Les moteurs des recherche offrent des fonctionnalités de recherche avancée:
  - Requêtes booléennes
  - Recherche de phrases
  - Recherche de mots exacts
  - Recherche par synonymes
  - Recherche par troncature
  - Recherche dans plusieurs langues
  - Recherches restreintes à un site, à un type de fichier, ...
  - Recherche sur domaines personnalisés
  - ...

Requêtes booléennes (Google)

- AND implicite

  [apple tree new]

  - retourne une page seulement si elle contient TOUS les termes de la requête
- Opérateur OR (en majuscule!)

  [apple OR tree OR new]

  - retourne une page si elle contient au moins un terme de la requête
- Termes négatifs

  [apple -tree]

  - retourne une page si elle contient le terme "apple" mais elle ne contient pas le terme "tree"
  - à utiliser pour exclure des significations d'un mot
  - au moins un terme positif est obligatoire
  - pas d'espace entre "-" et le terme négatif
  - Toujours un espace avant "-"

Requêtes booléennes (Google)

- Combinaison d'opérateurs:
  - Le "-" est appliqué en premier
  - En suite le OR
  - Seulement à la fin le AND (implicite)

[vacances grêce OR italie -france] ↔

[vacances (grêce OR italie) -france]

Retourne une page si:

- elle contient "vacances" et
- elle ne contient pas "france" et
- elle contient soit "grêce" soit "italie"

Requêtes booléennes (Google)

- Combinaison d'opérateurs:
  - Le "-" est appliqué en premier
  - En suite le OR
  - Seulement à la fin le AND (implicite)

[apple OR tree -apple OR -tree] ↔

[(apple OR tree) (-(apple) OR -(tree))]

Retourne une page si:

- elle contient "apple" ou "tree" et
- elle ne contient pas "apple" et "tree" ensemble

Recherche de phrases (Google)

- Entourer une expression par des guillemets pour rechercher une phrase exacte
  - ["le plus grand bâtiment de paris"]

    retourne une page seulement si elle contient l'expression exacte "le plus grand bâtiment de paris"
  - [le plus grand bâtiment de paris]

    peut retourner une page

    "...**le plus grand** illustrateur du XXème siècle travaille sur le magnifique **bâtiment** des Arts décoratifs **de Paris**..."

Recherche de mots exacts (Google)

- Les moteurs de recherche normalement ignorent la ponctuation et le mots vides (stop words) comme les articles, les conjonctions etc.
- Précéder un mot par "+" pour forcer l'inclusion du mot dans la recherche

  [star wars +I]

  retourne des informations sur le premier épisode de star wars

Recherche par synonymes (Google)

- L'opérateur "~" devant un mot clef, force la recherche soit du mot soit de ses synonymes

  [~inexpensive restaurant]

  retourne aussi des pages qui contiennent "cheap restaurant" "affordable restaurant" etc.

Recherche par troncature (Google)

- Le symbole "*" peut correspondre à un ou plusieurs mots dans la recherche:
  - Surtout utiles dans la recherche de phrase
  - ["to * or not to *"]

    retourne des pages qui contiennent la phrase
    - "to be or not to be" ou
    - "to read or not to read" ou
    - "to go away or not to think"
- L'opérateur "*" ne s'applique pas à des portions de mots:
  - ["la Pinaco* de Paris"]

Recherches restreintes (Google)

Plusieurs restrictions de recherche dans Google peuvent être spécifiées dans le formulaire de recherche avancée (Paramètres / Recherche avancée)

Recherches restreintes (Google)

Recherche avancée - restriction de la recherche :

- aux pages écrites dans une langue particulière
- aux seules pages des sites ".edu" (ou "youtube.com", etc)
- aux fichiers d'un format particulier
- aux pages qui ont été mises à jour dans un certain laps de temps
- aux pages où les mots clefs figurent dans une position particulière (par exemple le titre)
- préférablement aux pages provenant d'une zone géographique particulière

Recherche sur domaines personnalisés

- Certains moteurs des recherche, comme Google, permettent de définir des moteurs de recherche personnalisés. Ils comportent :
  - Un ensemble de sites et pages auxquelles la recherche est restreinte
  - Une interface de recherche personnalisée
  - La possibilité d'exporter le moteur sur sa propre page Web
- Exemples de moteurs de recherche personnalisés :
  - moteur spécialisé dans la recherche des offres d'emploi
  - Moteurs orientés pour la recherche de solutions techniques pour mac
  - ...

Création d'un moteur de recherche

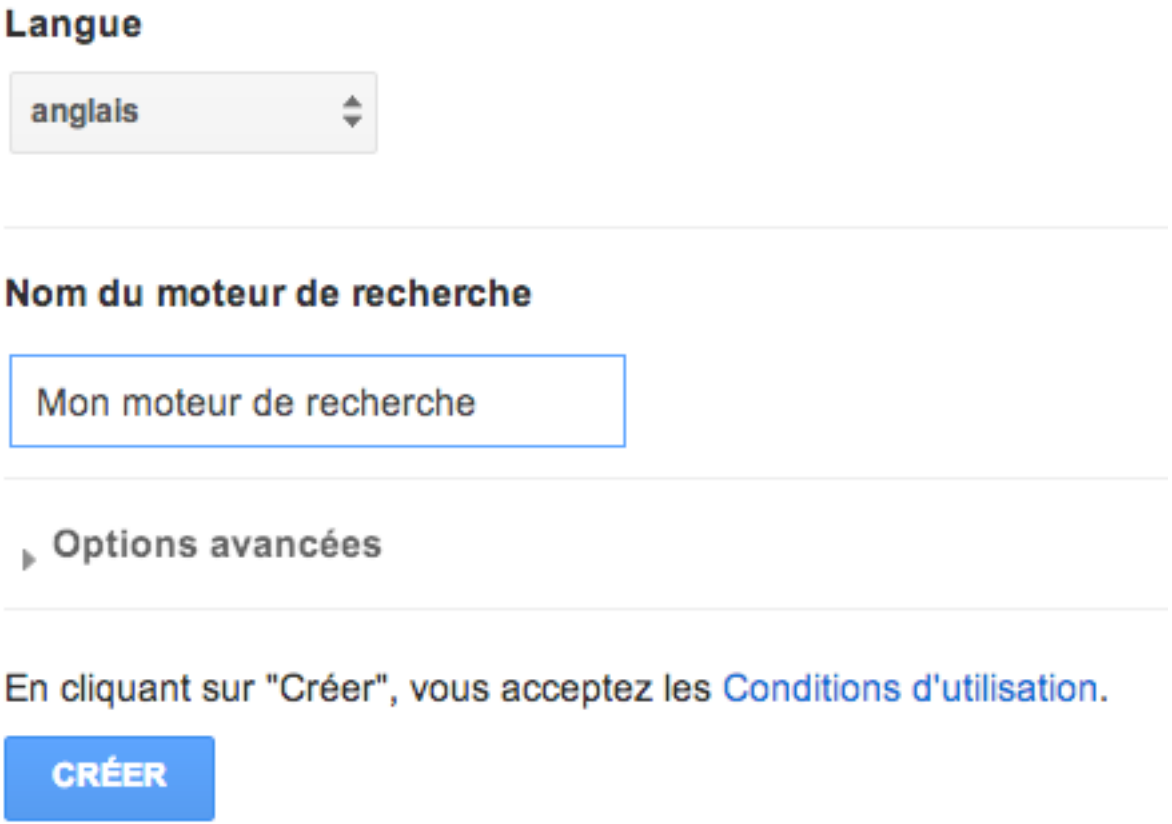
- Pré-requis: avoir cherché et sélectionné un ensemble initial de sites sur le sujet d'intérêt
- Avoir un compte Google
- Adresse du service: http://www.google.com/cse/

Création d'un moteur de recherche

- Lister les sites à inclure

Création d'un moteur de recherche

- Donner un nom au moteur de recherche et choisir la langue d'interface
- ensuite créer le moteur



Création d'un moteur de recherche

- Le moteur de recherche créé est prêt à l'emploi sur sa propre homepage
- Choisir "Modifier le moteur de recherche " pour continuer la personnalisation

Création d'un moteur de recherche

- Autres personnalisations:
  - Ajout de nouveaux sites à tout moment (rubrique "Configuration"- "Sites sur lesquels effectuer des recherches")
  - Création de filtres (rubrique "Fonctionnalités de recherche"-"Filtres")
    - Des filtres (étiquettes) peuvent être créés et affectés aux sites inclus dans le moteur de recherche (affecter le "libellé" des sites dans la rubrique "Configuration")
    - les étiquettes figureront dans chaque fenêtre de recherche
    - Après la recherche, l'utilisateur peut sélectionner une étiquette "etiq" pour filtrer le résultats obtenus: seulement ceux obtenus par les sites étiquetés par "etiq" sont sélectionnés
  - Personnalisation de l'aspect de l'interface (rubrique "Apparence")