Recherche Documentaire

Racine Mattieu - Jamet Jason - Grandjean Guillaume

Faculté des Sciences

23 Mars 2015

Recherche Documentaire I

Finefound

Le meilleur moteur de recherche Documentaire !

Sommaire

Introduction

- Cahier des charges
  - Développement d'un moteur de recherche
  - Recherches correspondant à un corpus de documents
  - Interface pour fournir sa requête
- Outils utilisés
  - Php - Mysql
  - Pecl : collection de bibliothèques PHP (pspell / stem)
  - QuickHash

Sommaire

Indexation

- Tableau en mémoire
- Base de données
- Table de Hachage
- Nouvelle indexation

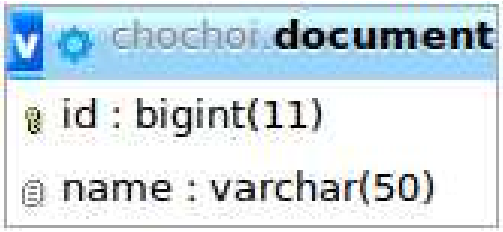
Indexation
Tableau en mémoire

Mots	Tableau Document Position
North	
Korea	...
Judg	

Document	Tableau Position
AP891231-0039	
AP891231-214	
AP891231-0325	

Position
152
254
317

Indexation
Base de données



Indexation
Table de Hachage

Indice	Liste
0	North
1	Korea
2	Judg
3	Democracy
4	Caesar
5	Glower
6	Amid
...	...
N	Cheer

Exemple avec le mot « Squat »

==> Présence dans la table de Hachage ?

==> Si oui Ajout dans le String de la requête Word.

==> Ajout de la position dans le String de la requête Position.

==> Ajout de la position dans le String de la requête Document.

Indexation
Nouvelle indexation

1. Ajout d'un nouveau document au répertoire
2. Lancer #new_index#
   - Vide les tables
   - Traite les nouveaux fichiers

Traitement des requêtes

- StopWords (a - this ...)
- KeyWords (and - or - not)
- Stemmer (judges => judg)
- Pspell (democrcy => democracy)
- Regex

Requête

And his dictatorship and makes democracy

Tableau de requête

array(2) [0]=> array(2) [0]=> array(2) [0]=> string(3) "his" [1]=> string(12) "dictatorship" [1]=> array(2) [0]=> string(5) "makes" [1]=> string(9) "democracy" [1]=> array(1) [0]=> string(3) "and"

Résultat

AP891216-0001&word=dictatorship,make,democraci&stopWord=his

Pertinence des documents

- Parcours de l'index
- Calcul du taux de pertinence

Pertinence des documents

Parcours de l'index

	
Complexitée de la requête	Requête et traitement simple
Execution lente	Execution et traitement rapide
Pagination possible	Pagination impossible

Calcul du taux de pertinence

Calcul du poids

TF : fréquence du terme

$$\frac{\text{“ nombre d'occurences du terme ”}}{\text{“ nombre de mots dans le document ”}}$$

IDF : Importance des termes moins fréquents

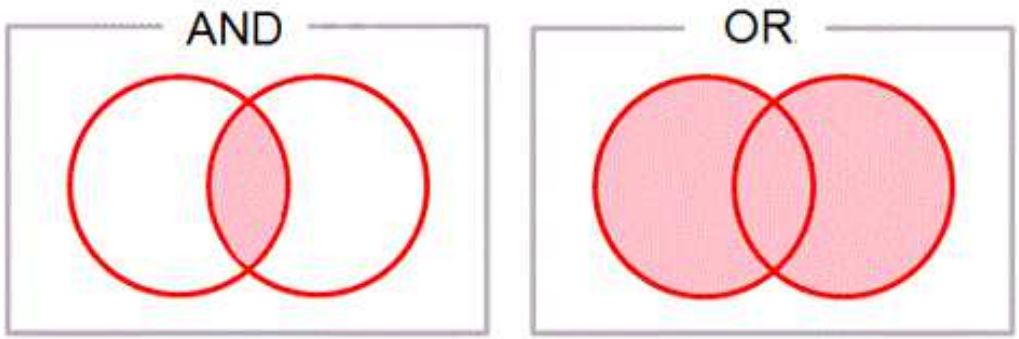
$$\log\left(\frac{\text{“ nombre total de documents ”}}{\text{“ nombre de documents où apparaît le terme ”}}\right)$$

Prise en compte de la position des mots

Calcul du taux de pertinence

Keywords

Inutilité des Keywords



Interface de recherche

- Conviviale
- Simple d'utilisation