

Introduction à l'indexation *fulltext*

Robert VISEUR

Assistant (FPMs) – Guideur technologique (CETIC)

robert.viseur@fpms.ac.be

Solutions Linux 2008

31 janvier 2008

Objectifs de l'exposé

- Proposer une introduction à la recherche *fulltext*.
- Exemples d'applications pour lesquelles l'exposé est utile:
 - moteur de recherche pour boutique en ligne,
 - moteur de recherche spécialisé (veille),
 - moteur de recherche de fichiers d'entreprises,
 - etc.
- Couvrir les différentes étapes de création d'un outil de recherche *fulltext*.

Pour quoi faire?

Exemples (1/4)

[Rt] Pour la 1ère fois, une majorité d'opinions défavorables à Sarkozy

Pour la première fois depuis qu'il est à l'Élysée le président de la République réunit plus d'opinions défavorables (48%) que favorables (45%, -6 points), selon un sondage BVA-Orange-Express publié mardi. En deux mois sa cote aura baissé de 10 (...)

[Newsisfree] Sarkozy: pour la 1ère fois, une majorité d'opinions défavorables

[Newsisfree] Sondage BVA : Sarkozy recueille plus d'opinions défavorables que favorables

Investissement en Bourse

Bourse & Placements Moins Chers 0
Abonnement - 0 Frais de Compte

Téléphone Chine

Appeler à l'international? Gratuit vers 15
pays!

Annonces Google

Page 1 sur 10 - 100 résultat(s) jugé(s) pertinent(s) sur 100 pour la requête : **sarkozy**.

En direct : les vœux de Sarkozy aux forces vives

En partenariat avec Public Sénat , suivez le discours du président de la République qui présente ses vœux aux représentants des syndicats, des entreprises et des associations.

<http://www.newsisfree.com/iclick/i,255758061,3297,f/>

Date : 17/01/2008 - Pertinence : 100%

Sauver :           

Nicolas Sarkozy reçoit cinq organisations syndicales pour évoquer sa "politique de la jeunesse"

(...) syndicales ((Unef, Uni, Fage, PDE, Confédération étudiante), pour faire le tour des dossiers en cours.

Nicolas Sarkozy a annoncé qu'une « politique de la jeunesse » serait prochainement mise en route, traitant des questions de l'emploi (...)

http://www.studyrama.com/article.php3?id_article=31149

Date : 17/01/2008 - Pertinence : 100%

Sauver :           

Moteur de recherche d'actualités (SQLite, algorithmes « maison »).

Pour quoi faire?

Exemples (2/4)

Requête : sarkozy (fr)

[Tous les audios de la-Croix.com](#)

Ecoutez les analyses des journalistes et des témoins de l'actualité sur la-Croix.com / Tous les audios de la-Croix.com / www.la-croix.com




- ◆ [MP3] [Certains mesures de Sarkozy sont incompatibles avec l'UE](#)
Marie Verdier, journaliste au service France de La Croix, revient sur les annonces de Nicolas Sarkozy en conclusion du Grenelle de l'environnement, et sur les suites à attendre de cette grande conférence
- ◆ [MP3] [Au QG de Royal : Jamais une personnalité ne nous avait fait rêver comme Royal](#)
Denis Peiron, de retour du QG de campagne de Ségolène Royal, raconte la déception des militants socialistes et leur crainte à l'égard de Nicolas Sarkozy
- ◆ [MP3] [Jean Véronis : L'immoralité politique est paradoxalement un thème de Bayrou](#)
Linguiste et blogueur émérite, Jean Véronis analyse à chaud les mots employés lors du débat qui a opposé Ségolène Royal à Nicolas Sarkozy, mercredi 2 mai

   - www.la-croix.com - Pertinence : ★ ★ ★ ★ ○

[UD Force Ouvrière Isère](#)

Podcast de l'UD Isere / UD Force Ouvrière Isère / udfo.isere.free.fr

- ◆ [MP3] [APRES LA RECONTRE MAILLY SARKOZY 16/05/2007](#)
MAILLY SARKOZY
- ◆ [MP3] [ACTION SAISONNIERS ETE 25/06/2007](#)
MAILLY SARKOZY

   - udfo.isere.free.fr - Pertinence : ★ ★ ★ ★ ○

Recherche de podcasts
(FeedReader, MySQL fulltext)

Pour quoi faire?

Exemples (3/4)

☰ Recherche sur RobertViseur.Be

Page 1 sur 9 - 89 résultat(s) jugé(s) pertinent(s) sur pour la requête : echonimo.

Echonimo propose des actualités belges - 22/04/2006 - RobertViseur.Be - Journal personnel

(...) billets Vous êtes ici: Accueil > Billets > Nouvelle du 22/04/2006 Nouvelle du 22/04/2006 [Mes travaux] [22-04-2006] **Echonimo** propose des actualités belges Un agrégateur externe a été mis en place pour **Echonimo** . Il permet la création de (...)

<http://www.robertviseur.be/news-20060422.php>

Popularité: 1 - Pertinence: 1

Sauver:           

Newsengine est lancé sous le nom d'Echonimo - 13/10/2005 - RobertViseur.Be - Journal personnel

(...) Billets > Nouvelle du 13/10/2005 Nouvelle du 13/10/2005 [Mes travaux] [13-10-2005] Newsengine est lancé sous le nom d'**Echonimo** Les premiers tests de Newsengine ayant été concluants sur la durée, Newsengine vient d'être lancé en version (...)

<http://www.robertviseur.be/news-20051013.php>

Popularité: 1 - Pertinence: 0.803369946594

Sauver:           

Moteur de recherche pour blog
(Lucene / Zend Search)

Pour quoi faire?

Exemples (4/4)

Page 1 sur 2 - 15 résultat(s) jugé(s) pertinent(s) sur pour la requête : **philips**.

[74.3€] Philips PHILIPS VOIP 3211S



(...) Caméscope Prêt à porter Maison / Literie Le sport Vin Newsletter : LA MOME Ed. Collector 2 DVD 17?99 > Voir mon panier : aucun article Vous êtes ici : Accueil > Telephonie > Les Téléphones Fixes > Les téléphones sans fil > Philips > PHILIPS VOIP (...)

<http://www.cdiscount.com/telephonie/telephone-fixe/philips-voip-3211s/f-144210109-PHILIPSVOIP3211S.html?search=voip&trilist=0&numpage=1>

Popularité: 1 - Pertinence: 1

Sauver:            

[144.9€] Philips VOIP4332S - Téléphone sans fil / téléphone



(...) Casque téléphonique Téléphone IP Téléphone filaire GSM Rechercher Téléphone IP > Telephone IP Philips VOIP4332S - Téléphone sans fil / téléphone Philips VOIP4332S - Téléphone sans fil / (...)

http://www.microchoix.com/cat/fiche252218-99-31-5-Philips_voip4332s_telephone_sans_fil_telephone_F

Popularité: 1 - Pertinence: 0.442293368537

Sauver:            

[69.89€] Philips VOIP321 (VOIP3211S)



Mon espace perso Votre panier est vide Tout le site Composants Connectique Destockage HiFi et Home-Cinéma Logiciels et livres Moniteurs LCD MP3 PC de bureau PC Portables Périphériques Photo et Caméscopes (...)

http://www.materiel.net/ctl/Telephonie_IP/23451-VOIP321_VOIP3211S_.html

Popularité: 1 - Pertinence: 0.4409727887

Sauver:            

Comparateur de prix (Lucene / Zend Search, extraction par regex)

Quatre phases

- Quatre phases importantes dans la réalisation d'un moteur de recherche:
 - L'extraction des données
 - L'indexation des données
 - La recherche d'informations
 - La présentation des résultats

Phase I: l'extraction (1/2)

- Conversion du fichier à indexer en texte brut.
- Cas simples:
 - fichiers « texte » structurés
 - Exemples:
 - XML (via PHP::SimpleXML),
 - RSS (via PHP::SimplePie),
 - HTML (via PHP::strip_tags ou analyseur HTML).
 - formats complexes documentés
 - Exemple: ODF = fichiers XML compressés (ZIP).

Phase I: l'extraction (2/2)

- Cas complexes: formats binaires non documentés.
 - Exemple: formats Office (97, 2000, XP,...)
 - Utilisation de projets Open Source:
 - Jakarta POI (documents MS Office), XLS2CSV (MS Excel), CatDoc (MS Word), PDFinfo (PDF),...
 - Extraction souvent imparfaite (~20% d'erreur avec POI).
 - Utilisation des iFilters (sous MS Windows).
 - Extensions proposées par les éditeurs eux-mêmes pour extraire le contenu de fichiers (MS Office, Autocad, etc).

Phase II: l'indexation (1/3)

- Langage SQL: recherche dans des chaînes de caractère.
 - `SELECT news.title, news.url FROM news WHERE news.title LIKE '%linux%'`
 - Pas adapté aux recherches dans de gros volumes de données pour des raisons de performance et de pertinence.
 - La pertinence peut être améliorée par un filtrage des résultats SQL (par exemple, via une expression régulière).
 - En pratique, il est fortement recommandé de recourir à l'utilisation d'un index inversé (dictionnaire).

Phase II: l'indexation (2/3)

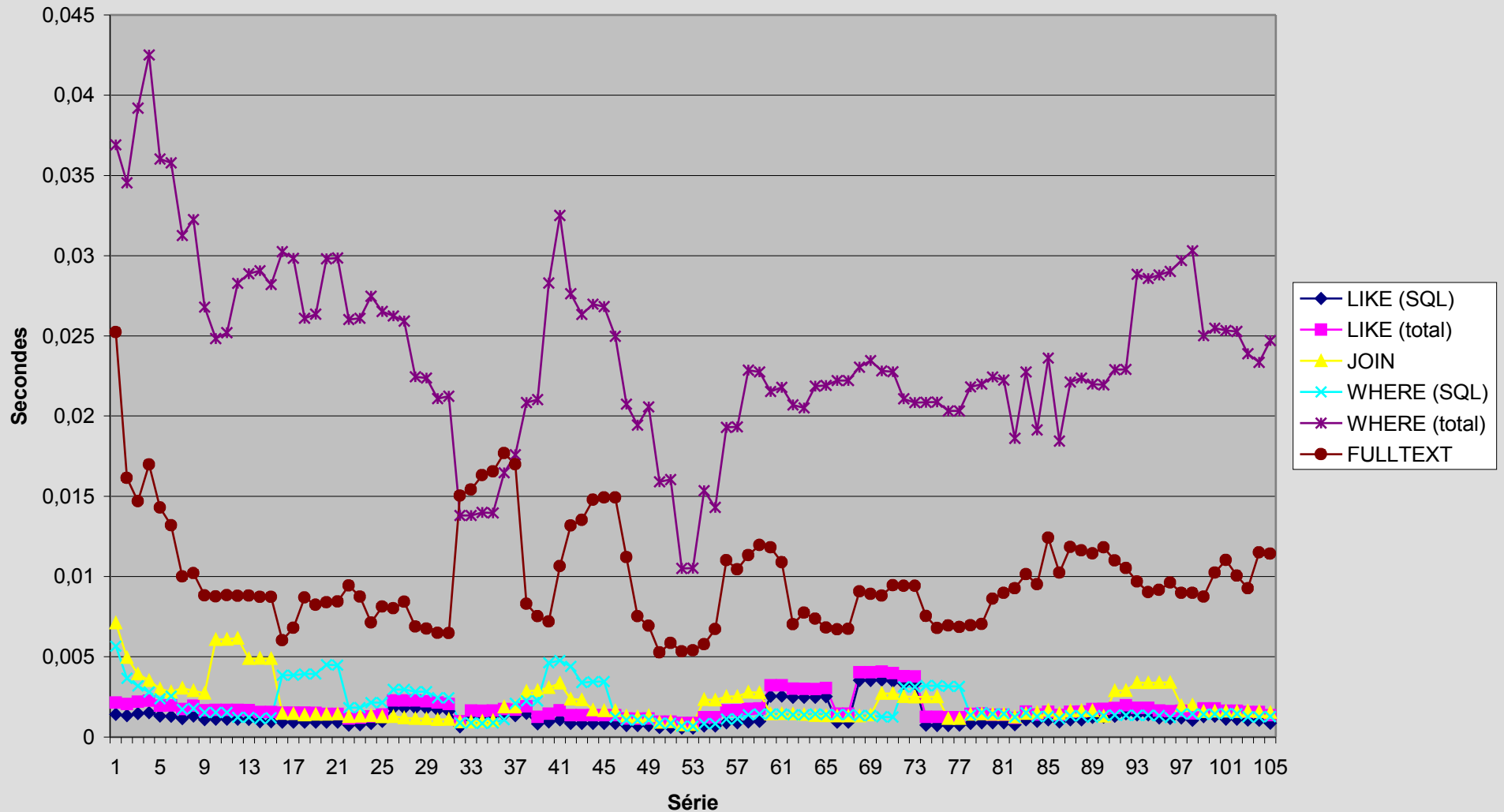
- Création du dictionnaire:
 - Filtrage du texte par la suppression des caractères non alphanumériques.
 - Décomposition du texte filtré en « termes » (*tokens*).
 - Suppression des mots noirs (le, la, les, mon, ton, son, notre, votre, leur,...)
 - Mise en place d'une table de correspondance (identifiant du document, terme)
 - Chaque terme est associé à une liste de documents comprenant ce terme.

Phase II: l'indexation (3/3)

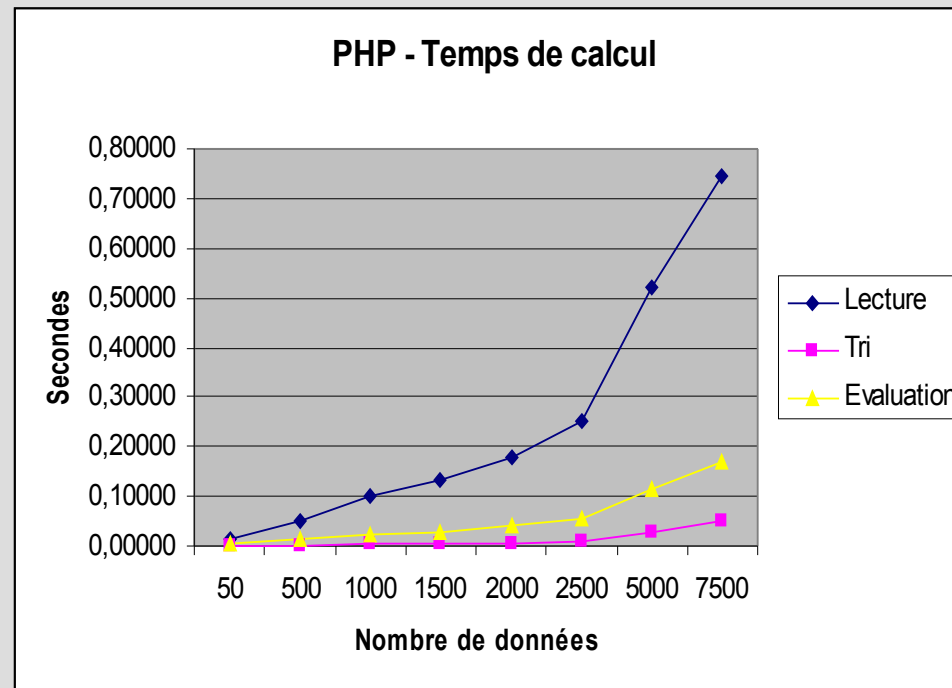
- Améliorations possibles:
 - Lemmatisation des termes.
 - Chaque terme est remplacé par sa forme canonique.
 - Exemple: étudier, étudiants, étudiant, étudiantes,... => « étud ».
 - Plusieurs implémentations Open Source de l'algorithme de Porter (ex.: Snowball).
 - Associations aux termes ou aux lemmes d'une forme phonétique (soundex, metaphone, etc).
 - Attention: lemmatisation et phonétisation sont sensibles à la langue.
- Index inversé: difficile à mettre en oeuvre.

Index inversé: difficulté de mise au point (1/3)

q=windows vista - n=16500 - MyISAM



Index inversé: difficulté de mise au point (2/3)



Index inversé: difficulté de mise au point (3/3)

- Pour des tests réalisés sous PHP5 (16500 enregistrements):
 - La manière d'écrire sa requête SQL a une influence importante sur le temps de calcul.
 - Sur ce volume de données, une requête LIKE est exécutée rapidement.
 - Les traitements en aval du SQL sont peu consommateurs de temps, tant qu'ils recourent essentiellement à des fonctions prédéfinies (ex.: array_multisort) et portent sur un nombre de données limité.
- D'où, intérêt de solutions standards.

Phase II et III: indexation *fulltext* sous MySQL (1/2)

- MySQL propose un mode automatique d'indexation *fulltext*.
 - Création: CREATE TABLE news (id INT UNSIGNED AUTO_INCREMENT NOT NULL PRIMARY KEY, title VARCHAR(256), body TEXT, **FULLTEXT (title, body)**)
 - Sélection: SELECT id, title, body, **MATCH (title,body) AGAINST ('linux') AS score** FROM news WHERE **MATCH (title,body) AGAINST ('linux')** ORDER BY **score**
- Avantages: prise en charge de la création du dictionnaire, de l'analyse de la requête, disponibilité d'opérateurs de recherche, évaluation d'un score de pertinence, mécanisme d'extension de requêtes,...

Phase II et III: indexation *fulltext* sous MySQL (2/2)

- Inconvénients (limitations):
 - pas de contrôle sur l'analyse du texte, (*tokenisation* mais pas lemmatisation),
 - Taille minimum des *tokens* (termes) fixée par défaut à 4 caractères (pas modifiable sur un serveur mutualisé).
- Notes importantes:
 - Il existe des solutions équivalentes sur d'autres SGBD (SQL Server, Oracle, etc).
 - Il existe des extensions pour SGBD (Tsearch2 pour PostgreSQL, etc).

Phase II et III: indexation par Lucene (1/3)

- Lucene n'est pas un système de base de données, Lucene est un indexeur. Cependant, Lucene intègre une notion de champs (pas de typage fort, pas de contraintes d'intégrité, pas de SQL, etc).
- Lucene permet la recherche par mots-clefs, ainsi que l'ajout, la modification et la suppression de documents, propose des opérateurs évolués de recherche et permet un tri par champs (sous Java).

Phase II et III: indexation par Lucene (2/3)

- Lucene est devenu une sorte de standard d'indexation.
 - Lucene en Java, PyLucene en Python, Lucene.Net en Dot Net, Zend Search en PHP, etc.
 - Compatibilité entre les index.
 - Plus qu'un logiciel, Lucene est aussi un format standard de fichier d'indexation.
- Lucene permet le contrôle du mécanisme d'analyse de texte.

Phase II et III: indexation par Lucene (3/3)

- Exemple (recherche):

```
require_once('Zend/Search/Lucene.php');
$index = new Zend_Search_Lucene('/index/news');
$hits = $index->find($query);
foreach ($hits as $hit)
{
    echo $hit->id;
    echo $hit->score;
    echo $hit->title;
    echo $hit->url;
}
```

Architecture de solutions de recherche

- Intérêt de proposer les résultats de recherche en XML.
 - Faciliter d'intégration de résultats en provenance de plusieurs outils de recherche.
 - Conversion de format aisée avec XSL.
 - Exemple: Exportation en RSS pour permettre la diffusion de contenu.
 - Facilité pour en décliner une API.
 - Travail avec des partenaires.
 - Intérêt pour une boutique en ligne, par exemple (affiliation).

Phase IV: présentation des résultats de recherche (1/2)

- Disponibilité d'outils Open Source permettant d'améliorer la présentation des résultats de recherche.
 - Classification (regroupement) des résultats.
 - Exemple: Python::Reverend.

Exemple:

1 regroupement(s) jugé(s) pertinent(s) pour la requête : carlos.

[Lemonde] [Le chanteur Carlos est mort](#)

Figure de la chanson populaire française, **Carlos** est mort, jeudi, d'un cancer "foudroyant", à l'âge de 64 ans, a-t-on appris auprès de sa soeur Catherine Dolto-Tolitch. Barbe fleurie, silhouette de bon vivant enveloppée dans des chemises à fleurs (...)

[\[Lemonde\]](#) [Le chanteur Carlos est mort \(Reuters\)](#)

[\[Newsifree\]](#) [Le chanteur Carlos est mort d'un cancer](#)

[\[Rt\]](#) [Le chanteur Carlos est mort](#)

[\[Nouvelobs\]](#) [Le chanteur Carlos est mort d'un cancer](#)

[\[Lesoir\]](#) [MUSIQUE: Le chanteur Carlos est mort](#)

Phase IV: présentation des résultats de recherche (2/2)

- Disponibilité d'outils Open Source permettant d'enrichir les résultats de recherche.
 - Correction orthographique.
 - Exemples:
 - Aspell.
 - Suggestion de mots similaires sur base de l'index (utilisation d'algorithmes de type Soundex ou Levenshtein).

Exemple:

0 résultat(s) jugé(s) pertinent(s) sur 0 pour la requête : **amhadinedjad**.

Il n'y a pas de résultat.

Voulez-vous dire ?

ahmadinejad

11693 news dans la base de données.

Conclusion

- Une bonne indexation est difficile à réaliser. Les solutions standards (ex.: MySQL *fulltext*), malgré quelques limitations, sont donc les bienvenues.
- Un peu d'ingéniosité donne parfois de bons résultats, sans mettre en oeuvre des solutions très compliquées.
- Réaliser un bon moteur de recherche ne se limite pas à l'indexation ou à la recherche, la présentation ou l'architecture peuvent aussi faire la différence.

Merci pour votre attention.

Des questions ?