

Antoine WALTER (GI02)

Développement d'un agrégateur web d'actualités via RSS

Université de Technologie de Compiègne – UV IC05
Février à juin 2006

Table des matières

I. Introduction, choix du sujet.....	3
I.A. Moteurs de recherche « classique » : un marché inaccessible ?.....	3
I.B. Moteurs de recherche d'actualités (news) : perfectibles.....	4
I.C. Définition de mon projet dans le cadre de l'UV IC05.....	6
II. Explications techniques.....	7
II.A. Choix des technologies.....	7
II.B. Structure de l'application.....	7
II.C. Moteur de l'application.....	8
II.C.1. Module « Finder ».....	8
II.C.1.aUtilisation des flux RSS.....	8
II.C.1.bExtraire le corps d'un article d'une page HTML.....	9
II.C.2. Module « Analyzer ».....	10
II.C.2.aAnalyse sémantique de l'article.....	10
II.C.2.bRegroupement d'articles connexes par thèmes.....	10
II.D. Etat d'avancement du projet (au 25/06/2006).....	11
III. Utiliser l'application.....	12
III.A. Tester la démo en ligne.....	12
III.B. Installer l'application chez soi.....	12
III.C. Captures d'écran.....	13
III.C.1. Derniers thèmes développés dans l'actualité.....	13
III.C.1.aNavigation par source.....	13
III.C.1.bStatus du moteur d'indexation et statistiques.....	14

I. Introduction, choix du sujet

Les moteurs de recherche sur le web m'ont toujours impressionné, par la mission qu'ils se sont donnée : indexer un maximum de pages web et offrir aux utilisateurs la possibilité d'y effectuer des recherches en un minimum de temps.

Cette mission est loin d'être évidente :

- Du fait de la quantité colossale de pages web à indexer et à maintenir à jour
- Du fait de la diversité des contenus (images, vidéos, sons...) et des technologies qu'elles utilisent (css, javascript, flash, applets java...)
- Du fait du temps de recherche qui ne doit pas excéder quelques centièmes de secondes

Le développement d'un moteur de recherche demande de bonnes compétences en programmation, mais surtout une bonne compréhension de l'organisation du web dans sa globalité. Il est en effet nécessaire d'établir une stratégie d'exploration pour obtenir un moteur de recherche efficace.

I.A. Moteurs de recherche « classique » : un marché inaccessible ?

Ayant de solides connaissances en développement web (5 ans d'expérience personnelle), j'ai souvent songé à créer mon propre moteur de recherche.

Cependant, le développement d'un moteur de recherche « classique » (indexation full-text), qui puisse être réellement utile à un grand nombre d'utilisateurs, demande un espace de stockage et une puissance de calcul importants. Pour cette raison, il est aujourd'hui impossible pour un étudiant d'atteindre le niveau des leaders tels que Google, Yahoo, Exalead... qui disposent chacun de milliards de pages web référencées dans des clusters de milliers de serveurs.

Bien qu'il eût été possible de développer un moteur de recherche miniature, j'en voyais peu l'intérêt. Qui voudrait utiliser un moteur de recherche qui n'indexe qu'une partie infime de l'Internet ? Sauf pour effectuer des recherches très particulières avec focus, qui prendraient de toute manière un temps énorme pour découvrir des pages web intéressantes sur un thème donné, un moteur de recherche miniature me paraissait sans avenir.

Cela ne veut pas dire que les moteurs de recherche « classiques » leaders resteront leaders de la recherche sur internet, mais plutôt qu'ils ont acquis plusieurs années d'avance dans l'indexation du web et qu'il faudrait investir des moyens considérables pour les battre sur leur propre terrain. Il faut par conséquent trouver de nouvelles méthodes de recherche ou de nouveaux types de recherche (images, vidéos, blogs, news...) pour les concurrencer.

I.B. Moteurs de recherche d'actualités (news) : perfectibles

Le secteur des moteurs de recherche « classiques » étant pour moi inaccessible, je me suis intéressé à d'autres domaines de recherche sur internet, qui nécessitent moins d'espace de stockage mais probablement plus d'analyse lors de la conception du moteur.

J'utilise depuis plusieurs années la rubrique « Actualités » de Google, qui donne un aperçu rapide des principaux articles publiés dans la semaine. *Google Actualités* me permet de me tenir informé de l'actualité lorsque je n'ai pas le temps de lire la presse, et de découvrir les dernières avancées technologiques dans la rubrique « Science/Tech ».

Remarque : il existe différents sites proposant des services similaires, mais je suis plus familier à celui de Google, qui est par ailleurs l'un des rares portails d'actualités à ne nécessiter aucune intervention humaine pour le tri et la hiérarchisation des articles. C'est pourquoi je n'analyserai que celui-ci dans le paragraphe ci-dessous.

The screenshot shows the Google Actualités (News) page for France. The top navigation bar includes links for Web, Images, Groupes, Annuaire, and Actualités plus. A search bar is present with the text 'Recherche Actualités' and 'Rechercher sur le Web'. Below the search bar, it says '500 sources d'information mises à jour en continu'. The main content area is titled 'Science/Tech' and features three article snippets, each enclosed in a dashed purple box. The first article is 'DADVSI : un compromis, sans aucune opposition' from PC Impact, dated 22 juin 2006. The second is 'E-administration: les Français de plus en plus exigeants' from ZDNet, dated 23 juin 2006. The third is 'Avec son geoportail, l'IGN rebat les cartes' from Libération, dated 22 juin 2006. A sidebar on the left contains a menu with categories like 'À la une', 'International', 'France', 'Économie', 'Science/Tech', 'Sports', 'Culture', and 'Santé'. There are also links for 'Alertes Actualités', 'RSS | Atom', and 'À propos des lux'. Annotations in green and purple highlight specific elements: a green arrow points to the search bar, a green circle highlights the 'Autres articles' link, and purple arrows point to the article titles and descriptions.

Comme montré ci-dessus, les articles provenant de diverses sources sont regroupés (par un procédé apparemment automatique) par thèmes, et pour chaque thème, un article est mis en avant (titre, photo, description).

Cependant, je trouve que *Google Actualités* propose une interface rudimentaire peu claire et qui n'exploite qu'une infime partie des possibilités offertes par ce concept :

Quelques critiques concernant Google Actualités :

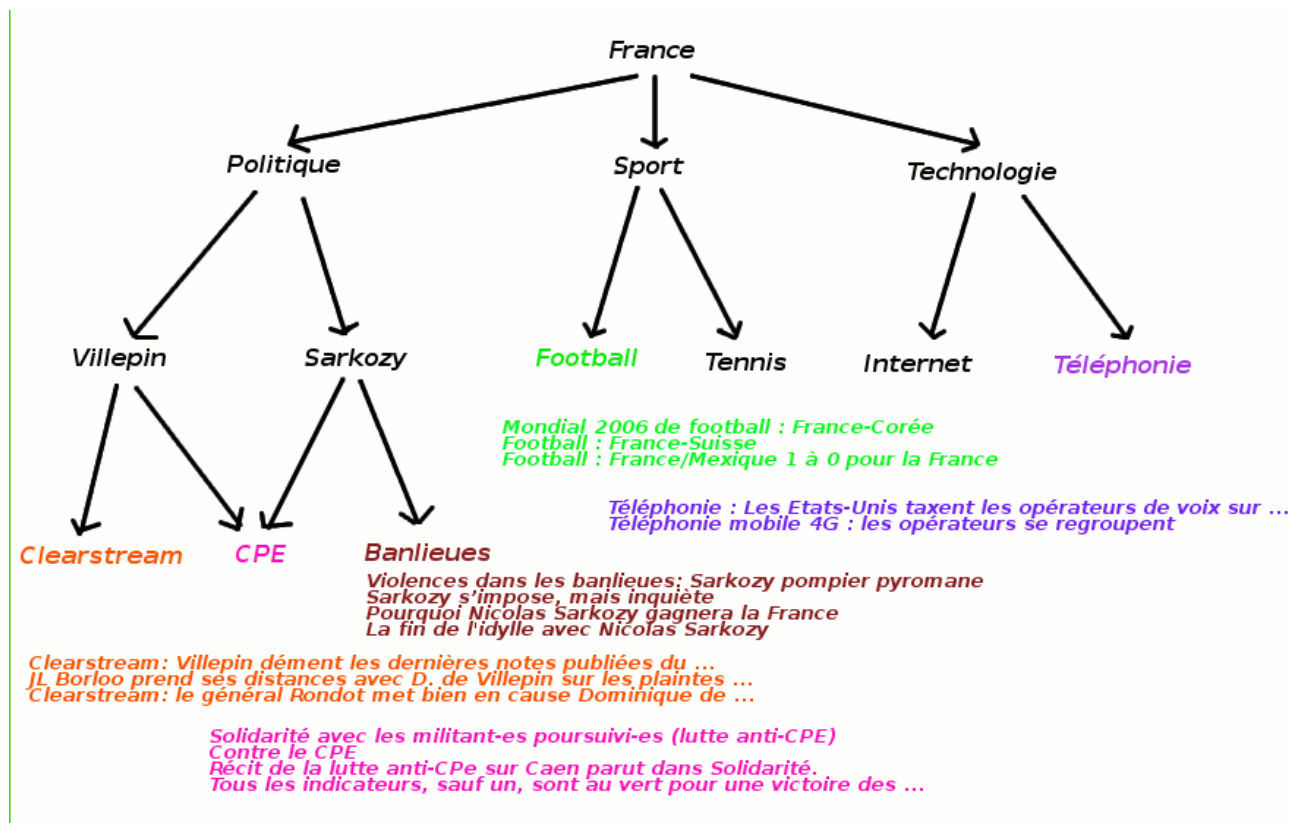
- La première limitation vient du fait que *Google Actualités* n'indique pas clairement par quels critères les thèmes d'actualités sont ordonnés et sélectionnés. Il arrive ainsi que des articles parus il y a plusieurs jours apparaissent au dessus d'autres articles plus récents (comme c'est le cas sur la capture d'écran).
La méconnaissance de cette information pourtant primordiale me gêne beaucoup : pourquoi tel thème apparaît-il en tête de la page, et pourquoi tel autre thème n'apparaît-il nulle part ?
- La seconde limitation vient du fait que *Google Actualités* n'indique pas comment le système

détermine si un article doit être mis en évidence ou simplement ajouté dans la longue liste d'articles connexes. Le pire est qu'il suffit d'actualiser la page pour changer totalement le classement des articles en évidence / connexes !

- Enfin, la fonction de recherche est très rudimentaire et ne permet pas de hiérarchiser les résultats par thèmes.

Quelques idées de nouvelles fonctionnalités :

- Permettre de filtrer les articles à afficher selon des critères précis : date de parution, nombre d'articles connexes, nombre de sources ayant traité le même thème, longueur de l'article...
- Créer une nouvelle vue qui permette de suivre l'actualité par ordre chronologique et à différentes échelles : affichage des principaux thèmes d'actualités du jour, de la semaine, du mois, de l'année, des 10 dernières années...
Cela permettrait d'avoir une vue globale des principaux événements s'étant déroulés sur une période donnée.
- Proposer des outils de statistiques permettant de se faire une idée plus précise de l'évolution des publications (nombre de parutions dans le temps, nombre de thèmes traités, temps moyen entre la publication d'articles connexes, sources étant les premières à publier de nouvelles thématiques...)
- Créer des agrégats d'articles et de thèmes à différents niveaux, pour parcourir l'actualité en surfant d'agrégat à agrégat, à l'aide de « tags ». Pour chaque agrégat, on pourrait soit consulter la liste des articles s'y rapportant, soit cliquer sur un sous-agrégat.



Remarque : Certaines fonctionnalités décrites ci-dessus ne sont pas proposées par les portails existants car ils n'en ont pas l'autorisation. En effet, il n'est pas légal de reprendre dans leur intégralité des contenus de sites web à leur insu pour les republier.

I.C. Définition de mon projet dans le cadre de l'UV IC05

J'ai décidé de créer mon propre portail d'actualités, que l'on pourrait également appeler « agrégateur de news » étant donné que sa tâche sera (dans l'ordre) de :

- Surveiller une liste prédéfinie de sources (« feeds »)
- Récupérer et analyser les nouveaux articles publiés par ces sources
- Agréger ces articles entre eux par thème
- Proposer une interface pratique pour consulter ces articles, et éventuellement proposer de nouvelles vues détaillées (recherche par critères, par échelle de temps, par tags/agrégats, statistiques...)

Ce projet s'inscrit directement dans le cadre de l'UV IC05 car il fait appel à de nombreux concepts étudiés en cours : qu'est-ce qu'une page web, comment l'analyser sémantiquement, quelles sont les interactions entre les différents sites web, comment fonctionne un crawler (moteur de recherche), quelles sont les techniques pour analyser un contenu web et le classer (pagerank, mots clés...).

Une fois ce programme réalisé, je compte l'utiliser à titre personnel, et éventuellement y donner accès à des proches. Je ne pourrai probablement pas l'ouvrir au public dans son intégralité pour des questions de droit d'auteur, mais peut-être mettre en ligne une version à fonctionnalités limitées (principalement : ne pas permettre aux internautes de lire un article dans son intégralité).

II. Explications techniques

II.A. Choix des technologies

PHP5 : Ce projet est résolument orienté web. PHP est la référence des langages web open-source. Il est capable d'être exécuté au sein de pages web, mais aussi comme un langage de programmation plus classique, interprété en mode console. Il est gratuit et open-source, et est très souple à l'utilisation. La version 5 permet une programmation orientée objet plus propre.

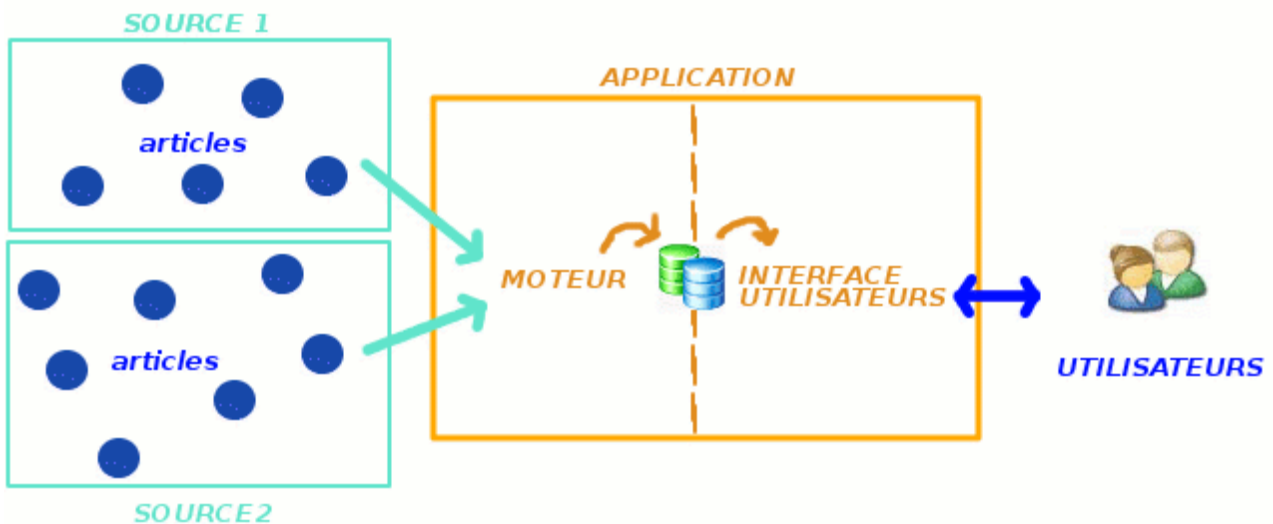
PostgreSQL8 : Ce système de gestion de bases de données est robuste et offre toutes les fonctionnalités que l'on peut attendre d'une base de données (transactions, sous-requêtes complexes, contraintes d'intégrité, triggers...), alors que d'autres SGBD comme MySQL ne les proposent pas. Il est gratuit et open-source.

Durant son développement, le projet est hébergé sur une machine personnelle Linux et desservi par un serveur web Apache.

II.B. Structure de l'application

L'application se décompose en deux grandes parties :

- Le moteur, qui se charge d'approvisionner la base de données à partir des différentes sources d'informations
- L'interface utilisateur, qui propose différentes fonctionnalités aux utilisateurs à partir de la base de données

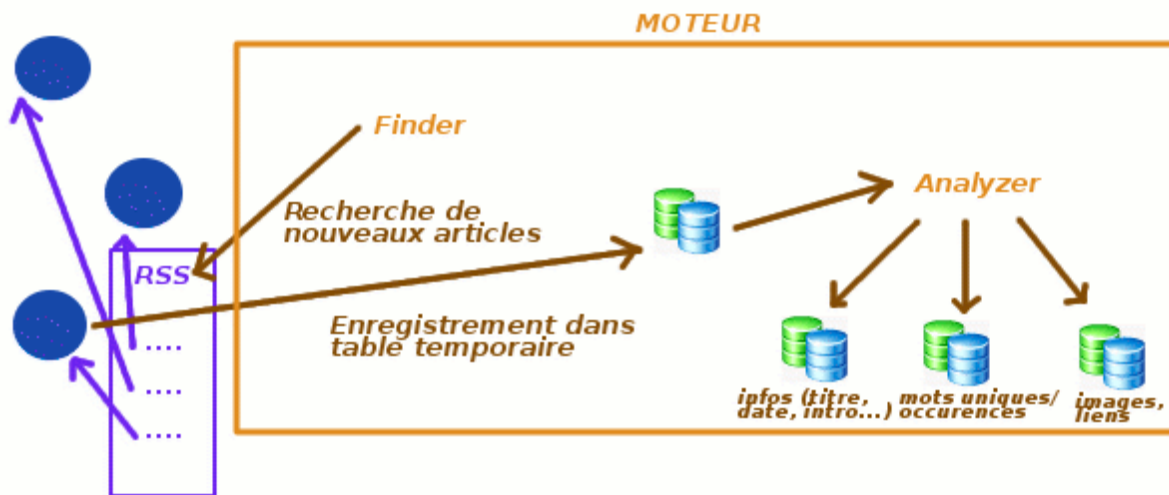


II.C. Moteur de l'application

Afin d'optimiser les performances de l'application, il a paru nécessaire assez rapidement de séparer l'application en plusieurs modules autonomes qui puissent s'exécuter de façon parallèle.

Ainsi, le moteur de l'application se compose des deux modules suivants :

- Le module « Finder » : son rôle est de parcourir régulièrement la liste des sources (« feeds »), de repérer les nouveaux articles parus, d'extraire leur contenu du reste de la page et de l'enregistrer dans une table temporaire de la BD.
- Le module « Analyzer » : son rôle est d'analyser au niveau sémantique chaque nouvel article trouvé par le module « Finder » (extraction de mots clés, images, liens), et de l'enregistrer de manière définitive dans la BD.
Il s'occupe également de regrouper des articles similaires en thèmes d'actualité.



II.C.1. Module « Finder »

II.C.1.a Utilisation des flux RSS

Une liste de flux RSS à surveiller est enregistrée dans la base de données.

Les flux RSS sont des fichiers XML permettant de lister les contenus d'un site web. La plupart des sites qui publient des articles de manière régulière tiennent à jour des fils RSS.

Voici un exemple de flux RSS qui référence deux articles 01net :

```
<rss version="2.0">
  <channel>
    <title>01net. Actualités</title>
    <link>http://rss.01net.com/www.01net.com/actus/?r=rss/actus.xml</link>
    <description>
      Toute l'actualité quotidienne des nouvelles technologies avec 01net. (01Informatique, 01DSI, Décision Informatique, Décision Distribution, 01 Réseaux, Micro Hebdo, L'Ordinateur Individuel, Telecharger.com, Micro Photo Video, UniversMac, Micro Achat)
    </description>
    <language>fr</language>
    <item>
      <title>service - Les FAI envoient leurs dépanneurs à domicile</title>
      <link>http://rss.01net.com/www.01net.com/article/320258.html?r=rss/actus.xml</link>
      <description>
        Le nombre de foyers connectés à Internet explose... les problèmes techniques des clients aussi. Les FAI proposent aux abonnés de se rendre chez eux pour l'installation et les dépannages.
      </description>
      <pubDate>Fri, 23 Jun 06 17:50:00 GMT</pubDate>
    </item>
    <item>
      <title>cartographie - Une carte de France à faire pâlir Google Earth</title>
      <link>http://rss.01net.com/www.01net.com/article/320238.html?r=rss/actus.xml</link>
      <description>
```

Geoportail.fr permet de se déplacer au-dessus de la France, à travers des clichés ou des cartes 2D, en attendant un service de navigation en 3D à la rentrée.

```
</description>
<pubDate>Fri, 23 Jun 06 17:40:00 GMT</pubDate>
</item>
</channel>
</rss>
```

Comme le montre l'exemple ci-dessus, les flux RSS permettent de connaître pour chaque article :

- le titre de l'article
- l'adresse de la page contenant cet article
- parfois, une introduction à cet article

Par contre, les flux RSS ne permettent pas d'obtenir le corps de l'article, ceci le plus souvent pour des raisons commerciales afin d'obliger les internautes à se connecter au site web pour le lire.

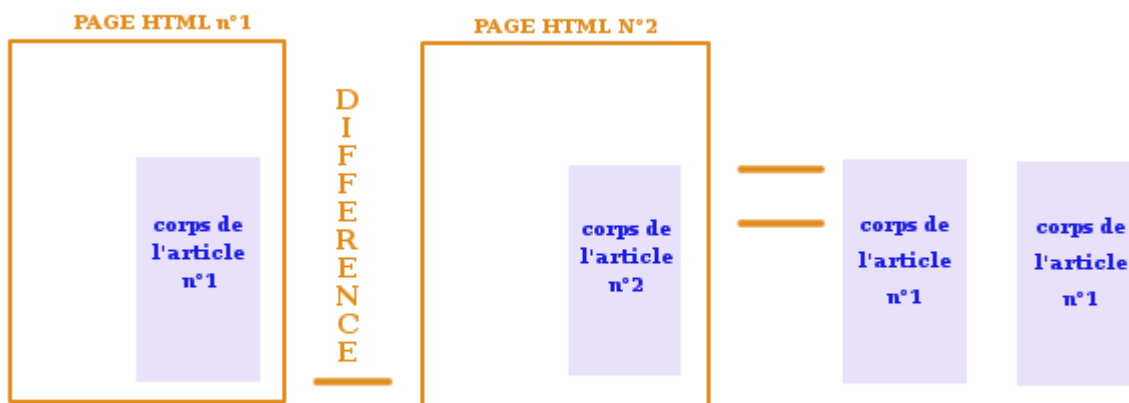
II.C.1.b Extraire le corps d'un article d'une page HTML

Une des parties les plus sensibles de ce module est celle qui se charge d'extraire un article de sa page. Les flux RSS ne proposant la plupart du temps pas le contenu de l'article en accès direct, la seule solution est de l'extraire de la page HTML originale.

Ceci est loin d'être évident car chaque site est organisé de façon différente. Le langage HTML étant très peu structuré et ne donnant aucune information sur le type de contenu qu'il renferme, il est très compliqué de différencier le corps d'un article du reste de la page (menus, bannières, liens vers des articles contextuels, commentaires d'internautes au bas de la page...).

Pour extraire le corps d'un article de sa page HTML, j'ai mis en application l'idée suivante : le plus souvent, les pages d'un même site web changent très peu d'un article à l'autre.

Il est par conséquent possible d'isoler le corps de l'article du reste de la page, en comparant deux pages HTML contenant deux articles différents publiés par une même source, et en éliminant toutes les parties communes (qui correspondent entre autres aux menus, headers, footers...).



Remarque : pour effectuer la différence des deux codes sources HTML, j'utilise la classe PHP « Diff » extraite de la librairie libre PEAR.

Une fois extrait de la page HTML, le contenu de l'article est inséré dans une table temporaire de la BD tel quel. Ce contenu sera ensuite analysé par le module « Analyser ».

II.C.2. Module « Analyser »

II.C.2.a Analyse sémantique de l'article

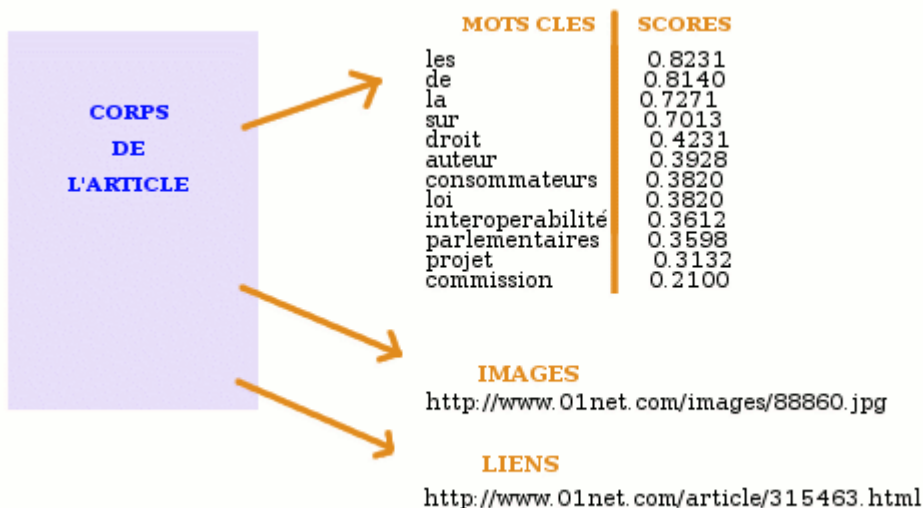
Le module « Analyser » analyse le corps de l'article récupéré du web par le module « Finder ». Le corps de cet article nécessite tout d'abord d'être nettoyé de toutes sortes de codes invisibles dans le navigateur, mais présents dans le code source (HTML, Javascript, CSS...).

Pour cela, j'ai défini mes propres méthodes dans la classe « Words » (sources/classes/class_words.php). Cette classe règle également les divers problèmes de jeux de caractères en convertissant si besoin les codes sources en UTF8.

Le module extrait ensuite les mots clés de l'article, en dressant un tableau associatif de tous les mots présents dans l'article associés à un score, calculé selon le nombre d'occurrences dans l'article et selon leur place (le score est plus élevé si le mot se trouve dans le titre ou l'introduction). Le score est calculé selon la formule : $\text{nb_occurrences} * \text{coeff} / \text{nb_total_mots}$

Enfin, le module extrait différentes informations telles que les images et les liens présents dans l'article.

Toutes ces informations sont enregistrées dans plusieurs tables de la BD de façon à être rapidement retrouvées.



II.C.2.b Regroupement d'articles connexes par thèmes

Cette étape a été la plus complexe à concevoir et a nécessité plusieurs expérimentations.

Pour regrouper des articles similaires, il faut comparer les articles entre eux deux à deux et chercher des similitudes.

J'ai essayé différentes méthodes de comparaison basées sur les mots clés des articles, avec différents seuils, et j'ai trouvé que la méthode suivante donnait les meilleurs résultats :

La méthode pour rechercher des articles connexes à un article donné, consiste à :

- Rechercher les X mots ayant le plus haut score pour l'article donné
- Rechercher les autres articles ayant dans leurs X meilleurs mots au moins Y mots en commun avec les X meilleurs mots de l'article donné

Ainsi, des articles connexes sont des articles ayant Y mots en communs dans leurs X meilleurs mots.

Actuellement, les seuils sont fixés à X=20 et Y=4. Ils permettent de regrouper correctement les articles dans environ 95% des cas !

Pour information, voici la requête SQL (extraite de la classe Article_agregator) qui permet d'effectuer cette recherche :

```
SELECT article_id, article_agregat, article_title, count(wr.wordrel_word) as nbcommonwords
FROM article INNER JOIN wordrel wr ON article_id=wr.wordrel_article
WHERE wr.wordrel_word IN (
    SELECT wordrel_word FROM wordrel wr2 WHERE wr2.wordrel_article=wr.wordrel_article ORDER BY
    wr2.wordrel_count DESC LIMIT ".Article_agregator::$COMMON_WORDS_RANGE." )
AND wr.wordrel_word IN($topwords)
AND article_id!='{ $this->getId()}'
GROUP BY article_id, article_agregat, article_title
HAVING count(wr.wordrel_word)>=".Article_agregator::$COMMON_WORDS_MIN." ORDER BY nbcommonwords
DESC");
```

Remarque importante : pour que cette méthode fonctionne, il est nécessaire d'éliminer les mots communs tels que « de », « il », « les », « pour »... qui fausseraient les résultats.

Pour cela, j'active un drapeau « word_common » à tous les mots qui sont présents dans au moins 1 article sur 6 parmi ceux déjà analysés.

II.D. Etat d'avancement du projet (au 25/06/2006)

Ce projet est ambitieux et il a nécessité une longue phase d'analyse et déjà plusieurs dizaines d'heures de programmation. Actuellement, le projet est toujours en développement, bien que la base du moteur d'indexation et d'analyse des contenus soit déjà opérationnelle.

Le module « Finder » extrait correctement les nouveaux articles des sources en un temps raisonnable (environ 2 articles par seconde, temps de téléchargement inclus).

Le module « Analyzer » analyse et enregistre les articles en base de données à une bonne vitesse (environ 4 articles par seconde). Il parvient également à reconnaître dans environ 95% des cas les articles qui traitent du même thème, et ce dans un temps correct (environ 2 articles par seconde). L'application a déjà pu enregistrer plusieurs milliers d'articles et plusieurs centaines de milliers de mots clés uniques, après quelques jours de fonctionnement.

Voici l'état du projet au 25/06/2006, date de ce rapport d'activité :

- Base de données : 90% (changements mineurs à prévoir pour optimisation)
- Module « Finder » : 95% (totalement opérationnel, je compte implémenter une répartition des téléchargements simultanés sur différentes sources, pour éviter de les saturer)
- Module « Analyzer » : 70% (temps de regroupement des articles à améliorer, implémenter l'aspiration des images)
- Interface utilisateur : 10% (tout est à faire ! L'interface actuelle est rudimentaire, elle permet juste d'avoir un aperçu des articles regroupés par thèmes)

III. Utiliser l'application

III.A. Tester la démo en ligne

Bien que l'application soit toujours en développement, elle peut être testée à l'adresse suivante :

<http://ic05.devs.fr>

Il est possible qu'elle soit par moments inaccessible, ou qu'elle affiche quelques erreurs. Vous trouverez également sur cette page des informations concernant l'évolution de son développement.

III.B. Installer l'application chez soi

Pour installer l'application, il faut avoir un serveur web (Apache/PHP/PostgreSQL) configuré, et dont la *locale* est « UTF8 ».

Ensuite, il suffit de créer la base de données à partir du fichier *dump.sql*, de copier le répertoire sources dans votre espace web et d'adapter le fichier de configuration *config.php*.

Enfin, il faut démarrer les processus du crawler en ligne de commande dans l'ordre suivant :

php moteur_finder.php

attendre ensuite quelques secondes que « Finder » trouve quelques articles... puis :

php moteur_analyzer.php

Les deux processus ne devraient jamais s'arrêter. Ils se mettent automatiquement en pause lorsqu'il n'y a pas de nouveaux articles à traiter.

Il est possible de suivre l'avancement des deux processus dans la rubrique « Stats » du site web.

III.C. Captures d'écran

Voici quelques captures d'écran de l'interface utilisateur actuelle. Celle-ci sera prochainement remaniée et proposera bientôt de nouvelles fonctionnalités.

III.C.1. Derniers thèmes développés dans l'actualité

Home | Latest headlines | Stats | Browse feeds

Irak : l'armée US a un plan de retrait [26/06/2006 16:51] 26/06/2006 16:54 [www] Le Nouvel Observateur
Washington a affirmé qu'elle a un plan pour un retrait massif de ses troupes d'Irak à la fin 2007. Mais il dépend "des conditions sur le terrain".
26/06/2006 16:42 **L'armée américaine a un plan de retrait [26/06/2006 16:37]** [www] NouvelObs Ebranger
26/06/2006 16:36 **Dépêche: Irak: la Maison Blanche confirme un plan militaire pour un retrait** [www] Le Monde

[Investissements] Alcatel installera la 3G en Malaisie 26/06/2006 16:48 [www] L'Atelier
Alcatel ouvre les portes du haut débit mobile à la Malaisie. L'équipementier français vient d'obtenir un contrat de 70 millions d'euros pour être le principal fournisseur de Celcom Berhad...
26/06/2006 16:33 **Alcatel installera la 3G en Malaisie** [www] Silicon.fr

Un homme de 24 ans avoue le meurtre du teknival de Carnoët 26/06/2006 16:36 [www] L'Express
empty
26/06/2006 16:29 **Meurtre de Mathilde: le suspect a avoué [26/06/2006 15:29]** [www] Le Nouvel Observateur

Perpétuité requise en appel contre Emile Louis 26/06/2006 16:36 [www] L'Express
empty
26/06/2006 16:36 **Emile Louis: Perpétuité requise en appel** [www] L'Express
26/06/2006 16:29 **Perpétuité requise contre Emile Louis [26/06/2006 16:18]** [www] Le Nouvel Observateur
26/06/2006 16:25 **La peine maximale requise contre Emile Louis** [www] Le Figaro à la Une

Le deuxième sommet France-Océanie s'ouvre à l'Elysée 26/06/2006 16:36 [www] L'Express
empty
26/06/2006 16:31 **Sommet France-Océanie à l'Elysée [26/06/2006 10:58]** [www] NouvelObs Ebranger

Menaces et pourparlers après l'enlèvement d'un soldat israélien 26/06/2006 16:36 [www] L'Express
empty
26/06/2006 16:31 **Soldat de Tshal enlevé : riposte en vue [26/06/2006 15:15]** [www] NouvelObs Ebranger
26/06/2006 16:29 **Soldat de Tshal enlevé : riposte en vue [26/06/2006 15:15]** [www] Le Nouvel Observateur

Warren Buffett promet 37 milliards de dollars à des fondations 26/06/2006 16:36 [www] L'Express
empty
26/06/2006 16:30 **Warren Buffett, recordman du don caritatif** [www] Libération
26/06/2006 16:25 **Le gourou de la finance Warren Buffett fait un don humanitaire record** [www] Le Figaro à la Une

Cette rubrique affiche en temps réel les articles groupés par thème, triés du plus récent au plus ancien. En cliquant sur le titre d'un article, il est possible d'en lire le contenu sans avoir à se rendre sur le site officiel. Il est également possible de visualiser la liste complète des articles ayant traité le thème par ordre chronologique.

III.C.1.a Navigation par source

Home | Latest headlines | Stats | Browse feeds

You are browsing crawled feeds. This page may take some time to be loaded.
[<< Previous feed](#) | [All feeds](#) | [Next feed >>](#)

Le Figaro à la Une (5)

Meurtre du teknival de Carnoët : le suspect fait des aveux complets (22) 26/06/06 16:25:25
Un jeune homme de 24 ans, un ancien de la Marine nationale interpellé à Marseille, a été «mis en examen du chef d'assassinat» dans le cadre de l'enquête sur le meurtre de Mathilde Croguennec...

La tension monte après l'enlèvement d'un soldat franco-israélien à Gaza (23) 26/06/06 16:25:25
Guilad Shalit a une vingtaine d'années. De père français, ce jeune artilleur dans un blindé israélien gardait dimanche un poste-frontière entre Israël et la Bande de Gaza lorsque des activistes palestiniens ont émergé d'un tunnel creusé sous la barrière de sécurité...

Le plus dur commence pour les Bleus (24) 26/06/06 16:25:25
Notre dossier spécial Coupe du monde Grâce à son succès sur le Togo (2-0), l'équipe de France a évité le pire : rentrer à la maison prématurément, comme en 2002, après un tout petit tour...

Cette rubrique permet de visualiser pour chaque source tous ses articles publiés.

III.C.1.b Status du moteur d'indexation et statistiques

[Home](#) | [Latest headlines](#) | [Stats](#) | [Browse feeds](#)

- Process Finder : **RUNNING**
- Process Analyzer : **RUNNING**
- Période en mémoire : du 26/06/06 16:25 au 26/06/06 17:42
- 19 feeds surveillés
- 338 articles analysés
- 0 nouveaux articles en attente d'analyse
- 134 articles ignorés (trop courts)
- 16941 mots clés uniques

Cette rubrique affiche l'état du moteur d'indexation : il est possible de vérifier l'état de chaque processus (Finder, Analyzer...).

Cette rubrique donne aussi des statistiques concernant la base de données : nombre d'articles analysés, en attente d'analyse, etc.

Pour toutes questions, remarques ou suggestions, vous pouvez me contacter via :

[antoine.utc \[at\] gmail.com](mailto:antoine.utc[at]gmail.com)