MASTER TAL 2007-2008

Cours « Projet Encadré »

Navigations dans Les Fils du Monde

Fleury Serge, CLA²T/SYLED Serge.fleury@univ-paris3.fr

Sommaire

1 Préambule

Les Lieux du projets :

Sur le blog pluriTAL[1], la page « Navigations dans Les Fils du Monde[2] » sera le lieu d'une synthèse "*au fil de l'eau*" de l'évolution de projet.

Contenu de ce document :

Ce document présente tout d'abord le cadre général dans lequel s'inscrit le travail à réaliser par les participants à ce projet.

On trouvera ensuite les différentes tâches à réaliser.

Ce document sera accompagné des différents corpus décrits *infra* (sur CD joint).

[1] http://tal-p3.wordpress.com/

[2] http://tal-p3.wordpress.com/navigation-dans-les-fils-du-monde-0708/

2 Contexte de la recherche

Le travail présenté dans ce document s'appuie sur 2 projets en cours [Fleury 2005] :

2.1 Phase 1

« *Le Monde en Surface* »

Ce projet a commencé en Octobre 2005. Il est composé de 2 modules.

1. Le premier (« **Fil(s) de presse** ») correspond au module permettant de traiter un fil de presse donné (au format RSS) et de construire des traitements sur le contenu de ce fil (au départ, un « nuage de mots[3] »).
2. Le second (« **Archivage des Fils de Presse** ») correspond au module permettant d'archiver les fils de manière continue et automatique afin de constituer la mémoire de ces fils.

On dispose à ce jour d'un corpus de fils RSS archivés toutes les heures et d'une série d'outils de traitement de ces fils (en développement).

URL du projet : http://tal.univ-paris3.fr/filspresse/

« *Le Monde Profond* »

Chaque version quotidienne du journal *Le Monde* a été régulièrement récupérée sur le site web du journal[4] : dans sa version HTML et dans sa version PDF. La version HTML[5] du journal a été traitée pour produire différents états :

1. un état quotidien des contenus textuels du journal sous la forme d'une version normalisée au format XML et une version compatible avec le logiciel ***Lexico3***
2. des états statistiques quotidiens

Les états quotidiens des contenus textuels ont ensuite été nettoyés et concaténés pour produire des corpus chronologiques couvrant l'ensemble des dates de récupération.

Le démarrage de ce processus a commencé le 12 avril 2003 et s'est arrêté[6] le 19 septembre 2006 *i.e.* on dispose à ce jour d'un corpus regroupant l'ensemble des versions électroniques de chaque journée couvrant cette période.

URL du projet : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm

---

[3] Le nuage de mots-clés (*tag cloud* en anglais) est une représentation visuelle des mots-clés (tag) les plus utilisés sur un site web ou utilisés pour classer des objets numériques (Source : http://fr.wikipedia.org/wiki/Nuage_de_mots). Dans le projet « Fils de Presse », les nuages de mots construits donnent à voir l'ensemble des mots présents dans les descriptions des articles des fils d'un journal en ligne à un moment donné. Dans la représentation de ces nuages, la taille de la police de caractères utilisée pour afficher le mot dans le nuage est déterminée par la fréquence du mot dans l'ensemble des articles scrutés pour un fil donné, et chaque mot est associé à un comportement donné.

[4] http://www.lemonde.fr/

[5] La version HTML traitée ici est celle dite "simplifiée (sans image de la une et sans menu déroulant)"

[6] A cette date, seule la version PDF du journal est disponible aux abonnés.

2.2 Phase 2

Ce projet est une extension du projet précédent dont le volet « Monde Profond » a été interrompu le 19/09/2006.

URL du projet : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm

« *Le Monde en Surface* »

Projet similaire à celui présenté dans la Phase 1.

« *Le Monde Profond* »

Ce projet ne prend plus appui sur la version électronique du journal mise à la disposition des abonnés. Il s'articule autour de l'archivage en parallèle des fils RSS et des articles complets associés aux items décrits dans les fils.

Dans la phase précédente, une première étape expérimentale a été mise en œuvre pour construire la version « enrichie » appelée *le Monde semi-Profond*. Le processus initialement mis en place a été « optimisé » et permet désormais d'archiver complètement les articles longs associés aux fils.

Le démarrage de ce processus a commencé le 20 novembre 2006.

3 Présentation des corpus fournis et à construire

3.1 Les données disponibles

Le journal Le Monde propose sur son site des fils RSS : *Les fils RSS sont des flux de contenus gratuits en provenance de sites Internet. Ils incluent les titres des articles, des résumés et des liens vers les articles intégraux à consulter en ligne. Les dernières informations publiées sur Le Monde.fr peuvent ainsi venir enrichir automatiquement votre site Internet ou compléter vos sources d'informations déjà agrégées via un logiciel de lecture des flux RSS.*

Ces fils donnent accès à 2 niveaux textuels différents :

- les contenus textuels des fils RSS, *i.e.* les contenus textuels des éléments *description* dans les fils RSS (le *Monde en Surface*)

- les contenus textuels de tous les articles complets décrits dans les fils RSS, *i.e.* les contenus textuels pointés par les éléments *link* dans les fils RSS (le *Monde Profond – Phase 2*)

Un processus a été mis en place pour archiver ces fils RSS. Le processus mis en place aspire régulièrement les fils visés et créé des pages de navigation pour donner à voir les données archivées[7] et les nuages de mots créés sur chacun des fils.

Un fil RSS archivé est un fichier XML dont la figure suivante donne à voir la structure et le contenu :

Contenu d'un fil RSS

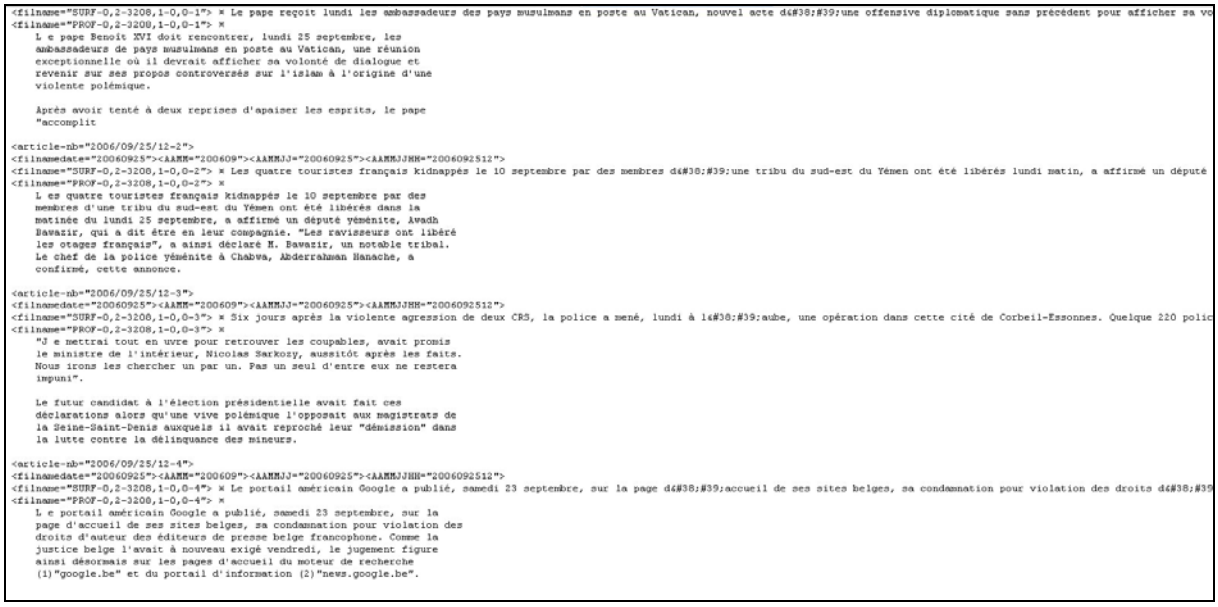
Ce fichier contient un cartouche de présentation et un ensemble d'éléments appelés *item* (les articles), chacun d'eux est structuré de la manière suivante :

```
<item>
<title>Titre de l'article</title>
<link>Lien vers l'article complet</link>
<description>Contenu textuel donnant une présentation courte de l'article</description>
<pubDate>Date de publication</pubDate>
```

[7] Les données de l'ensemble du projet « Fils de Presse » sont visibles ici : http://sfmac.no-ip.com/fils-presse-arch/index.xml (accès restreint).

```
…
</item>
```

Le processus mis en place archive donc ces fichiers, il déclenche aussi en parallèle un processus complémentaire permettant de récupérer, pour chacun des *items* contenu dans un fil RSS, l'article complet associé à cet *item* (correspondant au contenu de l'élément *link*), puis de stocker cet article complet ; le processus construit donc un archivage de *textes alignés* : une description courte d'un article et sa version longue. Dans la figure qui suit, on trouve alternativement le texte « court » puis le texte « long » pour un fil donné.



```
<filename="SURF-0,2-3208,1-0,0-1"> x Le pape reçoit lundi les ambassadeurs des pays musulmans en poste au Vatican, nouvel acte d&#38;#39;une offensive diplomatique sans précédent pour afficher sa vo
<filename="PROF-0,2-3208,1-0,0-1"> x
    L e pape Benoit XVI doit rencontrer, lundi 25 septembre, les
    ambassadeurs de pays musulmans en poste au Vatican, une réunion
    exceptionnelle où il devrait afficher sa volonté de dialogue et
    revenir sur ses propos controversés sur l'islam à l'origine d'une
    violente polémique.

    Après avoir tenté à deux reprises d'apaiser les esprits, le pape
    "accomplit

<article-nb="2006/09/25/12-2">
<filenamedate="20060925"><AAMM="200609"><AAMMJJ="20060925"><AAMMJJHH="2006092512">
<filename="SURF-0,2-3208,1-0,0-2"> x Les quatre touristes français kidnappés le 10 septembre par des membres d&#38;#39;une tribu du sud-est du Yémen ont été libérés lundi matin, a affirmé un député
<filename="PROF-0,2-3208,1-0,0-2"> x
    L es quatre touristes français kidnappés le 10 septembre par des
    membres d'une tribu du sud-est du Yémen ont été libérés dans la
    matinée du lundi 25 septembre, a affirmé un député yéménite, Awadh
    Bawazir, qui a dit être en leur compagnie. "Les ravisseurs ont libéré
    les otages français", a ainsi déclaré M. Bawazir, un notable tribal.
    Le chef de la police yéménite à Chabwa, Abderrahman Hanache, a
    confirmé, cette annonce.

<article-nb="2006/09/25/12-3">
<filenamedate="20060925"><AAMM="200609"><AAMMJJ="20060925"><AAMMJJHH="2006092512">
<filename="SURF-0,2-3208,1-0,0-3"> x Six jours après la violente agression de deux CRS, la police a mené, lundi à 1&#38;#39;aube, une opération dans cette cité de Corbeil-Essonnes. Quelque 220 polic
<filename="PROF-0,2-3208,1-0,0-3"> x
    "J e mettrai tout en uvre pour retrouver les coupables, avait promis
    le ministre de l'intérieur, Nicolas Sarkozy, aussitôt après les faits.
    Nous irons les chercher un par un. Pas un seul d'entre eux ne restera
    impuni".

    Le futur candidat à l'élection présidentielle avait fait ces
    déclarations alors qu'une vive polémique l'opposait aux magistrats de
    la Seine-Saint-Denis auxquels il avait reproché leur "démission" dans
    la lutte contre la délinquance des mineurs.

<article-nb="2006/09/25/12-4">
<filenamedate="20060925"><AAMM="200609"><AAMMJJ="20060925"><AAMMJJHH="2006092512">
<filename="SURF-0,2-3208,1-0,0-4"> x Le portail américain Google a publié, samedi 23 septembre, sur la page d&#38;#39;accueil de ses sites belges, sa condamnation pour violation des droits d&#38;#39
<filename="PROF-0,2-3208,1-0,0-4"> x
    L e portail américain Google a publié, samedi 23 septembre, sur la
    page d'accueil de ses sites belges, sa condamnation pour violation des
    droits d'auteur des éditeurs de presse belge francophone. Comme la
    justice belge l'avait à nouveau exigé vendredi, le jugement figure
    ainsi désormais sur les pages d'accueil du moteur de recherche
    (1)"google.be" et du portail d'information (2)"news.google.be".
```

Alignements dans le Monde archivé

La chaîne de traitements mise en œuvre permet donc de construire deux niveaux de représentation des contenus textuels mis en ligne sur le site *via* les fils RSS :

- les contenus textuels des fils RSS, *i.e.* les contenus textuels des éléments *description* dans les fils RSS (le *Monde en Surface*)
- les contenus textuels de tous les articles complets décrits dans les fils RSS, *i.e.* les contenus textuels pointés par les éléments *link* dans les fils RSS (le *Monde semi-Profond*)

Les processus d'archivage des fils RSS du Monde ont démarré progressivement et on dispose à ce jour des données suivantes :

- Le *Monde en surface* : depuis 2005
- Le *Monde semi-Profond* : depuis 2006

3.2 Les données traitées

Au cours de ce projet vous travaillerez sur une arborescence de fils « récoltés » tous les jours à 19h sur le site du Monde depuis novembre 2006 (période traitée : 11/2006, 03/2008).

3.3 Les données à construire et à traiter pour le projet

A partir des données fournies vous devrez dans un premier temps construire des corpus regroupant tous les fils **d'une même rubrique**, puis vous devrez formater ce corpus thématique au format Lexico3.

Pour mener à bien cette tâche on utilisera le programme fourni sur le CD qu'il conviendra de mettre à jour (programme : `parcours-lesfilsdumonde-et-makerubrique.pl`).

On pourra ensuite choisir de travailler :

- **soit** sur des états du corpus associés à la *surface du Monde* (le contenu textuel des fils RSS)

- **soit** sur des états associés à *la profondeur du Monde* (le contenu textuel des articles associés aux fils)

en réalisant une extraction dans les fichiers de rubrique créés précédemment des zones textuelles correspondantes. Les 2 types de zones de texte étant facilement repérable dans les fichiers disponibles (un filtrage avec egrep suffit…).

IMPORTANT : avant tout traitement sur les fichiers de rubrique construits (par le programme `parcours-lesfilsdumonde-et-makerubrique.pl`) , il faudra probablement utiliser la commande unix2dos sur ces fichiers.

4 « Navigations dans Les Fils du Monde »

4.1 Préambule

Lecture initiatique :
Dégrouper les sens à partir d'attestations, B. Habert (LIMSI – CNRS & université Paris X – Nanterre)
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/PROJETSM1/COMMUNAUTE/Site/ProjetDegrouperSensWebLeMonde.pdf

4.2 Corpus de travail

Cf partie précédente : *les données à construire et à traiter pour le projet.*

4.3 Outils de travail

Lexico3 :
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

Lexico3 est l'édition 2001 du logiciel ***Lexico*** dont la première version remonte à 1990. Les fonctionnalités présentes dès la première version (segmentation, concordances, décomptes portant sur les formes graphiques, spécificités et analyses factorielles portant sur les formes et les segments répétés) ont été conservées et, la plupart du temps notablement améliorées. L'originalité principale de la série ***Lexico*** est qu'elle permet à l'utilisateur de garder la maîtrise sur l'ensemble des processus lexicométriques depuis la segmentation initiale jusqu'à l'édition des résultats finaux. Les unités qui seront ensuite automatiquement décomptées sont exclusivement constituées à partir de la liste des délimiteurs fournie par l'utilisateur, sans recours à des ressources dictionnairiques extérieures. Au-delà du repérage des seules formes graphiques, le logiciel permet d'étudier dans les textes la répartition d'unités plus complexes composées de séquences de forme : *segments répétés, couples de forme en cooccurrence, etc.* au contenu souvent moins ambigu que les formes graphiques dont elles sont composées.

Le Trameur :
http://tal.univ-paris3.fr/trameur/

Programme de génération puis de gestion de la trame et du cadre d'un texte (*le métier lexicométrique*) pour construire des opérations lexicométriques. *Le Trameur* intègre le programme *treetagger* : système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation.

4.4 Entrées lexicales choisies

1. On travaillera à partir des 60 entrées lexicales étudiées par Jean Véronis[8] dans le cadre de la campagne d'évaluation en désambiguïsation sémantique ***Romanseval***. Le fichier contenant les 60 entrées lexicales est disponible sur le CD fourni.
2. On pourra choisir d'autres entrées lexicales en s'inspirant d'initiative comme "Les mots de la rencontre" : une opération réalisée, en partenariat entre la D.G.L.F.L.F du ministère de la culture et de la communication, l'inspection générale (I.G.E.N) et la direction générale de l'enseignement scolaire (DGESCO) du ministère de l'éducation nationale dans le cadre de la 13e Semaine de la langue française qui se déroulera du 14 au 24 mars 2008. Les dix mots retenus sont : apprivoiser, boussole, jubilatoire, palabre, passerelle, rhizome, s'attabler, tact, toi, visage ; les liens précédents pointent vers les définitions des mots concernés sur le site de la semaine de la langue française[9].
3. On pourra aussi choisir des entrées lexicales particulières pour les étudier sur les corpus fournis en s'inspirant par exemple du projet **LexiMedia2007**[10]

> *LexiMédia2007 analyse en permanence les flux issus de Presse 2007 et extrait automatiquement les mots ("débat", "retraite", "délinquance") et les expressions ("logement sociaux", "débat interne", "régimes spéciaux de retraite", "carte scolaire", "prévention de la délinquance") utilisés dans les articles. LexiMédia2007 donne l'**évolution au fil des semaines** de la fréquence d'utilisation de ces expressions, globalement et par journal. Pour chaque semaine, il donne les expressions les plus utilisées, les expressions en forte hausse, les expressions en forte baisse et celles dont la variation (hausse et baisse confondues) est la plus importante. Pour chaque expression, LexiMédia2007 donne le détail de son évolution (courbe de fréquence sur l'ensemble des semaines) et les liens vers les articles dans lesquels il apparaît.*
>
> *(source : http://aixtal.blogspot.com/2007/01/outil-leximedia2007.html)*

4.5 Etape 1 « Connaissance du corpus »

Avant de « partir », il faut prendre connaissance du corpus. En particulier, à partir de la présentation du projet, spécifiez :

- les documents qui constituent le *corpus* (nature et limites dans le temps) ;
- la manière dont ces documents sont organisés/traités pour qu'on puisse s'en servir comme corpus ;

Donnez, en vous appuyant sur un exemple, votre définition des termes techniques suivants :

1. fil RSS ;
2. occurrence ;
3. forme ;
4. corpus ;
5. partie de corpus.

4.6 Activités Etape 2 « Constellations »

Etape n°1 Choisir un mot et un des corpus construits

Etape n°2 Avec ***Lexico3***, construire la carte des sections contenant ce mot. A partir des « *Mots spécifiques dans les sections contenant le mot choisi i.e. les co-occurrents de ce mot* », commencez à organiser les mots qui cooccurrent avec *le mot choisi.*

Une des manières de faire est de créer un tableau de la forme :

---

[8] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/PROJETSM1/COMMUNAUTE/Site/JVeronis1998senseval.pdf

[9] http://www.semainelf.culture.fr/?idD=7

[10] http://erss.irit.fr/LexiMedia2007/

	noms	adjectifs	verbes	autres
thématique 1				
...				
thématique n				

Trois conseils pratiques :

1. ce peut être une bonne chose d'avoir une catégorie "balai" dans laquelle on mettra tout ce qu'on n'arrive pas à classer dans les thématiques ;
2. évitez d'avoir trop de thématiques (un petit ensemble de catégories est plus facilement utilisable d'un éventail trop large) ;
3. donnez-vous une limite (par exemple les 100 premières lignes) et classez tranquillement ces premiers mots. Ensuite, parcourez rapidement le reste pour voir s'il y a des thématiques qui ont été manquées en se concentrant sur la tête de liste.

4.7 Etape 2 « associations dominantes dans le Monde »

4.7.1 Associations dominantes des noms

Décrire les emplois (et éventuellement les sens) de noms de ***Romanseval***. Il y a 20 noms. On commencera par ceux pour lesquels les contextes sont les moins nombreux.

4.7.2 Angles d'attaque

4.7.2.1 Repérer les contraintes d'emploi

Un certain nombre d'approches permettent de mettre en évidence les micro-constructions dans lesquelles figure un mot, les attirances qu'il entretient :

- segments répétés (*Lexico3*) ;
- information mutuelle (couple de mots s'attirant) ;
- constructions syntaxiques privilégiées ;
- ...

4.7.2.2 Organiser les emplois

Les attirances et les micro-constructions sont structurées par les emplois du mot. Par exemple, *barrage* rentre dans la construction *faire barrage* quand il s'agit d'un emploi "politique", dans *tir de barrage* ou *match de barrage* pour l'emploi sportif, et dans *franchir un barrage* pour l'emploi militaire ou policier. Les zones denses du graphe des attirances entre mots fournissent un repérage grossier des emplois, avec des intersections et des limites quand le graphe est trop fourni et trop enchevêtré. D'autres méthodes peuvent aider à organiser les emplois : les classifications des phrases contenant le mot.

4.7.3 Démarche n°1 *via* *Lexico3* ou avec *Le Trameur*

Etape n°1 Choisir un mot à décrire ;

Etape n°2 Construire avec ***Lexico3*** ou avec *Le Trameur* les segments répétés.

Etape n°3 Examiner les segments contenant ce mot.

Etape n°4 Construire avec *Le Trameur* des séquences (correspondant au patron syntaxique potentiel) contenant ce mot ; construire aussi les graphes donnant à voir les associations induites

Etape n°5 Classer les constructions et attirances sur le plan syntaxique manifestées par le mot choisi et faire des hypothèses sur les emplois correspondants ; on utilisera éventuellement les concordances pour examiner en contexte certains emplois (en jouant sur les tris sur le contexte gauche / sur le contexte droit). On gardera dans le rapport les diverses concordances.

Etape n°6 Calculer l'*information mutuelle*[11] et des graphes basées sur l'information mutuelle à partir des données disponibles (on commencera ici par lire le document, visé par la note précédente, qui présente la mesure « information mutuelle » et la chaîne de traitement à mettre en œuvre pour la calculer) ;

Etape n°7 Intégrer les renseignements fournis par l'information mutuelle dans la description des constructions et attirances.

4.7.4 Démarche n°2 *via Lexico3*

Etape n°1 Choisir un mot à décrire ;

Etape n°2 Construire avec ***Lexico3*** la carte des sections contenant le mot choisi

Etape n°3 Déterminer les cooccurrents de ce mot.

- A partir de la carte des sections, recherchez les mots spécifiques contenus dans l'ensemble des carrés colorés obtenus (*i .e* recherche les mots spécifiques dans les sections qui contiennent le mot choisi). A partir de cette recherche, on obtient une liste de mots qui portent soit un indice de spécificité positif soit un indice de spécificité négatif, dans le premier cas, on aboutit en gros à une liste des co-occurrents du mot choisi, dans le second cas on obtient des mots qui n'apparaissent pas avec ce mot.

4.8 Lectures complémentaires

Véronis, J. (2004). Hyperlex : lexical cartography for information retrieval. *Computer, Speech and Language, 18* (3), 223-252. [Lire]

Véronis , J. (2004). L'étiquetage sémantique des corpus. *Le Français Moderne. 2004/1,* 27-38. [Lire]

Véronis, J. (2003). Hyperlex : cartographie lexicale pour la recherche d'informations. *Actes de la Conférence Traitement Automatique des Langues (TALN'2003)* (pp. 265-274). Batz-sur-mer (France): ATALA. [Lire]

Véronis, J. (2003). *Hyperlex : cartographie lexicale pour la recherche d'informations.* Rapport interne. Equipe DELIC, Université de Provence. [Lire]

Ferret, Olivier Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales, TALN 2004 XIe conférence sur le traitement automatique des langues naturelles, 2004, Bernard Bel and Isabelle Marlien, Fès (Maroc), 19-22 avril, ATALA (Association pour le Traitement Automatique des LAngues [Lire]

Cédric Lamalle, William Martinez, Serge Fleury, André Salem, Andrea Kuncova, Aude Maisondieu (2001). "*Dix premiers pas avec Lexico3*", Manuel d'utilisation abrégé [Lire]

[11] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/BAO-master/bh-info-mutuelle/PresentationEtCalculInformationMutuelle.html

5 Références bibliographiques

[FLE 2005], Serge Fleury (EA2290 SYLED/CLA2T), "*Un corpus de veille : le journal Le Monde*", http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm.

[FLE 2006a], Serge Fleury (EA2290 SYLED/CLA2T), *Des nuages de mots (qui s'attirent) (1)* (Mars-Avril 2006) ou ici : *Des nuages de mots (qui s'attirent) (1)* (Mars-Avril 2006)

[FLE 2006b], Serge Fleury (EA2290 SYLED/CLA2T), *Des nuages de mots (qui s'attirent) (2)* (Juin 2006)

[FLE 2006c], Serge Fleury (EA2290 SYLED/CLA2T), *Des nuages de mots (qui s'attirent) (3)* (Septembre 2006)

[FLE 2006d] [*Fleury Serge, Salem André et co-auteurs*], *Explorations textométriques*, http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations-tdm.html

[GAN 2006] Jean-Gabriel Ganascia, Julien Bourdaillet, *Alignements unilingues avec MEDITE*, J, in Actes JADT2006, Besançon, 2006

[GAN 2002] Jean-Gabriel Ganascia, Irène Fénoglio, Jean-Louis Lebrave *Manuscrits, genèse et documents numérisés*, , Document numérique, Volume X, n°X/2002, pages 1-X

6 Annexes

6.1 RSS

L'illustration suivante présente de manière synthétique le rôle des fils RSS :

Suppose you have 50 sites and blogs that you like to visit regularly. Going to visit each website and blog everyday could take you hours. With RSS, you can "subscribe" to a website or blog, and get "fed" all the new headlines from all of these 50 sites and blogs in one list, and see what's going on in minutes instead of hours. What a time saver!

(Source : How to explain RSS the Oprah way)

6.1.1 Flux RSS

Source : http://www.commentcamarche.net/www/rss.php3

6.1.1.1 Introduction au RSS

Le standard RSS représente un moyen simple d'être tenu informé des nouveaux contenus d'un site web, sans avoir à le consulter.

*Le format « **RSS** » (traduisez « Really Simple Syndication ») permet ainsi de décrire de façon synthétique le contenu d'un site web, dans un fichier au format XML, afin de permettre son exploitation par des tiers. Le fichier RSS, appelé également **flux RSS**, **canal RSS** ou **fil RSS**, contenant les informations à diffuser, est maintenu à jour afin de constamment contenir les dernières informations à publier.*

Basiquement, un fil RSS est un fichier contenant le titre de l'information, une courte description et un lien vers une page décrivant plus en détail l'information. Cela permet à un site web de diffuser largement ses actualités tout en récupérant un grand nombre de visiteurs grâce au lien hypertexte permettant au lecteur de lire la suite de l'actualité en ligne.

Les sites proposant un ou plusieurs fils d'actualités au format RSS arborent parfois un des logos suivants :

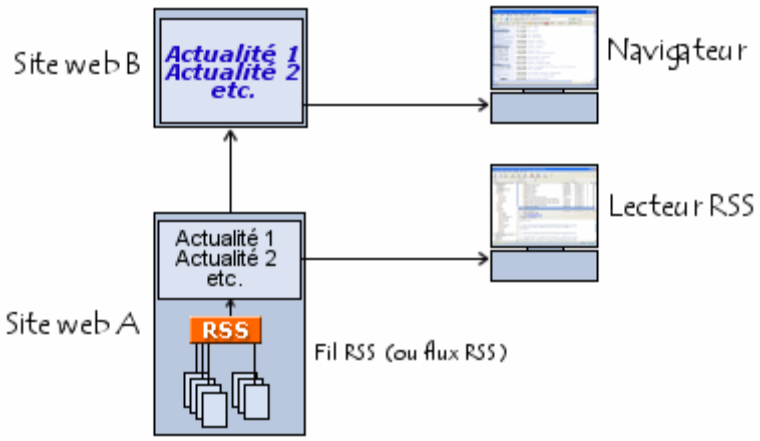
- RSS
- XML

Les blogs proposent ainsi généralement des outils natifs permettant de créer ou d'utiliser des fils RSS.

6.1.1.2 Utilisation de canaux RSS

Il existe typiquement deux façons d'utiliser RSS :

- *L'**utilisation des fils RSS par un particulier** pour son information personnelle. Il est alors nécessaire de disposer d'un outil spécifique, appelé « lecteur RSS » ou encore « agrégateur RSS », afin d'exploiter les fils RSS. Ainsi, l'utilisateur d'un lecteur RSS peut consulter en un seul endroit les dernières actualités de dizaines, et parfois de centaines de sites web, sans avoir à les visiter et sans avoir à communiquer d'informations personnelles.*
- *L'**utilisation des fils RSS par un webmaster** afin de syndiquer du contenu, c'est-à-dire publier automatique sur son propre site diverses informations émanant d'autres sites.*



6.1.1.3 Proposer un fil RSS

Pour proposer un flux RSS sur son site et mettre ainsi une partie de son contenu à disposition des autres webmasters, il suffit de créer un script chargé de récupérer les informations à inclure dans le flux RSS et de les écrire dans un fichier XML au format RSS.

6.1.1.4 Exploiter les fils RSS sur un site ?

*N'importe quel webmaster, pour peu qu'il dispose des outils adéquats, peut ainsi utiliser le flux RSS d'un autre site web afin d'afficher automatiquement sur son site les informations mises à sa disposition. Qui plus est, dans la mesure où les informations sont au format XML, il est possible de personnaliser l'affichage des données selon sa propre charte graphique et il est également possible d'agréger de multiples fils RSS au sein d'une même page : on parle ainsi de «**syndication de contenu**».*

Afin d'exploiter un fil RSS proposé par un site, il est nécessaire de disposer d'un outil capable d'analyser le XML (un parseur XML) afin de le convertir en XML. Il existe un grand nombre d'outils dans la plupart des langages permettant d'exploiter facilement des canaux RSS. L'outil MagPie RSS permet par exemple de parser les fils RSS, quelle que soit la version du standard utilisée, avec un simple script en langage PHP.

6.1.2 Pour aller plus loin

On pourra aussi se reporter à des présentations plus complètes :

1. Présentation très complète réalisée par des bibliothécaires de l'université de Montréal : *"La première partie se veut une introduction générale à RSS tandis que la seconde partie s'applique à démontrer des exemples d'usages courants et innovateurs de cette technologie par les bibliothèques, usages qui pourraient être amenés à se généraliser dans le futur."*

http://hdl.handle.net/1866/144

2. Support de cours réalisé par Stéphane Cottin pour la formation ADBS : « Utiliser les fils RSS ». Les points abordés au cours de cette formation :

- Définition(s) et usages du format RSS

- Comment s'y abonner ? Comment en trouver ?

- Quels outils ou services pour les exploiter (services en ligne, logiciels

indépendants, fonctionnalités des navigateurs)

- Comment chercher dans des fils RSS ?

- Comment syndiquer du contenu ?

http://www.servicedoc.info/rss/

6.2 Schéma[12] des états du monde (phase 1)

(le *Monde électronique et ses fils RSS*)

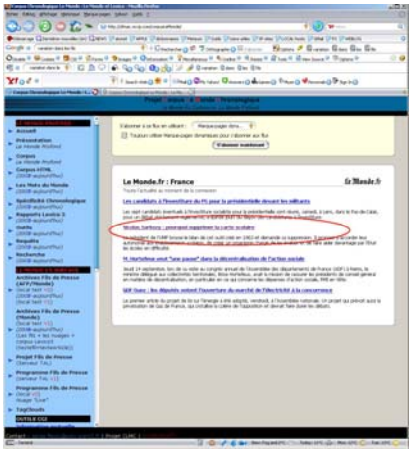
Ci-dessus la page[13] du site du journal le Monde

présentant l'ensemble des fils RSS disponibles.

(le *Monde en fil*)

Ci-dessus le contenu[14] d'un fil RSS (rubrique la Une)

(le *Monde en Surface*)



Ci-dessus, le contenu d'un fil RSS[15] archivé.

Nous avons surligné en rouge le contenu d'un élément de ce fil.

Ci dessous, le contenu textuel[16] du fil

(associé à l'élément *description* du fil) :

```
<filname="SURF-0,2-3224,1-0,0-2"> ¤
Le président de l&#38;#39;UMP brosse le bilan
de cet outil créé en 1963 et demande sa
suppression. Il propose d&#38;#39;accorder
leur autonomie aux établissements scolaires,
```

(le *Monde Semi-Profond*)

Ci-dessus, l'article long[17] (en ligne) associé

à l'élément surligné précédent.

Ci-dessous, un état du contenu textuel archivé

associé au contenu textuel du fil précédent :

```
<filname="PROF-0,2-3224,1-0,0-2"> ¤
Le président de l'UMP brosse le bilan de cet
outil créé en 1963 et demande sa suppression.
Il propose d'accorder leur autonomie aux
établissements scolaires, de créer un rganisme
```

---

[12] Dans le schéma suivant, chaque image est cliquable et donne à voir en taille réelle son contenu.

[13] http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html

[14] http://www.lemonde.fr/rss/sequence/0,2-3208,1-0,0.xml

[15] http://sfmac.no-ip.com/fils-presse-archivage/2006/Sep/17/00-00-00/0,2-3224,1-0,0.xml (accès restreint)

[16] http://sfmac.no-ip.com/fils-content-presse/2006/Sep/16/15-30-00/0,2-3224,1-0,0.txt (accès restreint)

[17] http://www.lemonde.fr/web/article/0,1-0@2-3232,36-813616,0.html?xtor=RSS-3224

```
de créer un organisme chargé de les évaluer
et de faire aider davantage par
l&#38;#39;Etat les écoles en difficulté.
```

```
chargé de les évaluer et de faire aider
davantage par l'Etat les écoles en difficulté.

En février 2006, lors de la convention
éducation de l'UMP, j'ai soulevé, parmi
d'autres questions, celle de la carte
scolaire. Plus de quarante ans après sa mise
en place, il n'est quand même pas incongru
d'en dresser le bilan.
```

Figure 1 : Les états du Monde (partie 1)

Pour terminer cette synthèse des états du monde, on présente ci-dessous l'article publié[18] dans la version électronique du journal (et archivé sous la forme dite « simplifiée »)

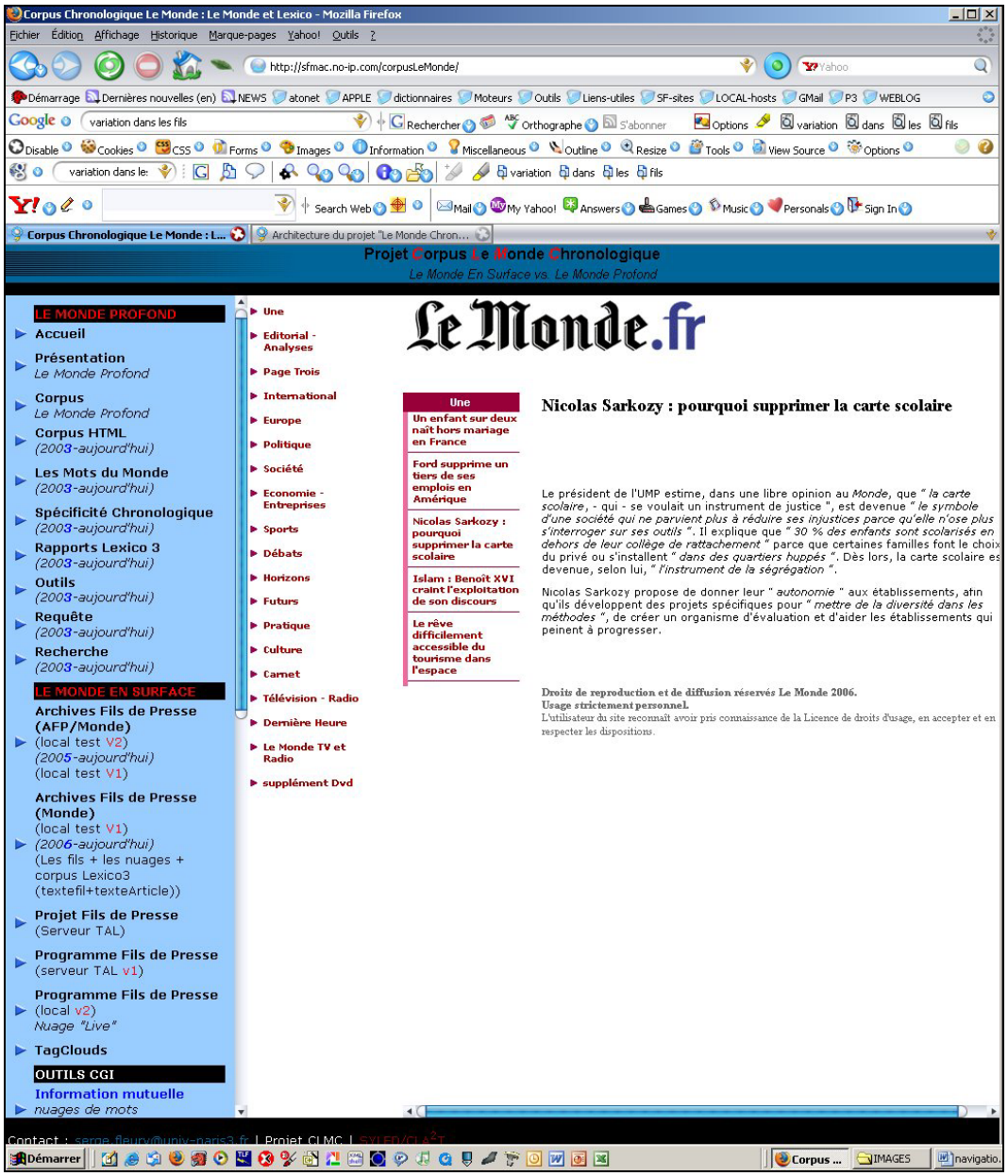


Figure 2 : Les états du Monde (partie 2)

[18] http://sfmac.no-ip.com/corpusLeMonde/HTML/060917/data/article_486884.html (accès restreint)

6.3 Information mutuelle : repérer les "mots" qui s'attirent

6.3.1 Bibliographie

Manning & Schütze, 1999
Manning, C. D. and Schütze, H. (1999).
Foundations of Statistical Natural Language Processing.
The MIT Press, Cambridge, Massachusetts.

6.3.2 Mesurer l'information mutuelle

Plusieurs mesures permettent de déceler les mots qui « s'attirent », c'est-à-dire qui tendent à apparaître en même temps. La mesure utilisée ici est l'*information mutuelle* [Manning & Schütze, 1999, p. 66-68] :

```
IM(x,y) = log(p(x,y) / p(x)p(y)).
```

C'est le rapport de la probabilité de la co-apparition des deux mots - p(x,y) - et du produit de la probabilité d'apparition de chacun d'eux : p(x)p(y). La probabilité p(x,y) est estimée comme : fréquence(x,y) / nombre total de mots. Il en va de même de p(x) et p(y). Le logarithme ajouté permet de "contracter" la dispersion des scores.

Il faut noter que l'information mutuelle mesure l'attirance au sein de couples (les mots dans un certain ordre) et non au sein de paires. La paire {*soin, bébé*} correspond à deux couples <*bébé, soin*> (*bébé* est le premier mot et *soin* apparaît plus à droite dans le texte) et <*soin, bébé*> (c'est cette fois *bébé* qui apparaît à droite de *bébé* dans le texte).

6.3.3 Mesurer l'information mutuelle : mise en œuvre

La chaîne de traitement utilisée est visible à cette adresse :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/BAO-master/bh-info-mutuelle/PresentationEtCalculInformationMutuelle.html