

Projet de Maîtrise 2003-2004

Informatique Linguistique et Recherche d'Informations

Jacques Vergne

*GREYC
Université de Caen*

<http://www.info.unicaen.fr/~jvergne>

Contact

bureau S3-391, tél: 02 31 56 73 36, e-mail : Jacques.Vergne@info.unicaen.fr

(mise à jour le 1er novembre 2004)

Thème de travail du projet

Le thème de travail proposé se situe dans le domaine des outils de **navigation dans un site internet** :

Comment faciliter l'accès au contenu d'un site de presse

par une étude terminologique de ce contenu

et une interface graphique basée sur les termes

Sites web des organes de presse :

La plupart des organes de presse ont leur site web. Par exemple, pour la presse quotidienne française, vous pouvez consulter :

- <http://www.ouest-france.fr/>
- <http://www.liberation.fr/>
- <http://www.lemonde.fr/>

On trouve des adresses des sites web des principaux sites de presse du monde entier et des données sur chaque site sur le site de l'hebdomadaire "Courrier International" :

<http://www.courrierinternational.com/kiosk/kiosq.htm>

Les "Unes" des sites de presse :

Les "Unes" des sites de presse sont calquées sur la version sur papier des Unes des journaux. Elles font plusieurs écrans (d'où un scroll obligatoire) et contiennent une masse d'informations trop importante. Elles comportent souvent de 100 à 150 liens. Leur lisibilité est médiocre, et l'accès aux articles est difficile.

Thèmes et termes

Les **thèmes** abordés par la rédaction de l'organe de presse, et qui peuvent intéresser le lecteur, sont matérialisés et accessibles par des "**termes**" de un ou plusieurs mots.

Les termes sont le plus souvent des **expressions nominales répétées**.

On aura besoin d'une méthode d'extraction de termes, à partir des textes cliquables de la Une, et éventuellement à partir des textes des articles. Cette méthode devra être indépendante des langues des sites de presse, et donc ne pas utiliser des ressources linguistiques propres à une langue (de telles méthodes seront présentées durant les réunions de travail).

Termes et articles

Sur un site de presse, on a un ensemble d'articles, et on aura extrait un ensemble de termes.

Chaque article contient certains termes. Chaque terme est dans certains articles.

À partir de ces relations, des **graphes de termes** ou des **graphes d'articles** sont calculables, et pourraient servir à construire l'interface graphique.

Un petit groupe de termes peuvent tous se trouver dans un même petit groupe d'articles, ce qui pourrait permettre de grouper des articles thématiquement liés.

Recherche d'une autre interface d'accès au contenu du site de presse

On s'orientera vers une interface graphique, par exemple sous forme de graphe (comme www.kartoo.com), ou/et sous forme de rectangles emboîtés (comme [Map of the Market](#)). Tout autre idée est bienvenue. Plusieurs vues sont envisageables.

Lire aussi : [Les hypermédias graphiques explorateurs](#) de D. Bihanic.

L'interface ne devra pas présenter trop d'informations simultanément.

Elle devra permettre à l'utilisateur de "naviguer dans le contenu" de l'actualité, de voir rapidement les principaux thèmes abordés, et dans quels articles, enfin permettre l'accès aux articles.

Les difficultés du sujet

On pourra se servir seulement de certains textes cliquables de la Une (ceux vers des articles), ou/et aussi des documents pointés par les liens sortants de la Une.

Parmi les documents pointés par les liens sortants de la Une, certains sont des articles, d'autres pas. Une méthode de discrimination des articles sera nécessaire. Cette méthode devra être indépendante des sites, et donc ne pas utiliser des caractéristiques propres à un site (cf. réunion de travail).

L'interface nécessite des calculs préparatoires (graphes de termes, graphes d'articles, groupes de termes, groupes d'articles), puis du dessin proprement dit, dessin permettant une interactivité, une navigation dans le(s) graphe(s) et dans le site de presse.

Quels utilisateurs ?

Les utilisateurs sont des lecteurs qui veulent accéder rapidement aux contenus d'un site de presse, et à qui on veut proposer une navigation centrée sur les thèmes de l'actualité, rapide et agréable.

Aspects techniques

Les choix techniques vous appartiennent et devront être motivés. Vous pourrez bien sûr utiliser des outils Open Source et des composants logiciels existants.

Organisation du travail :

Ce projet est proposé à 2 binômes qui réaliseront ensemble le système, en se partageant le travail, par exemple :

- un binôme sera responsable du crawl et de l'étude terminologique,
- l'autre sera responsable de l'interface graphique.

La coordination entre les 2 binômes fait partie intégrante du travail.
