*Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., (1999) "Projet NeuroWeb : un moteur de recherche multilingue et cartographique", 5ème conférence Hypertextes et Hypermédias : Réalisations, Outils & Méthodes, H2PTM'99, Paris, 23 et 24 Septembre 1999.*

# Projet NeuroWeb : un moteur de recherche multilingue et cartographique.

**A. Lelu*, M. Hallab*, H. Rhissassi*, F. Papy**, S. Bouyahi*, N. Bouhaï*, H. He***, C. Qi****, I. Saleh***

** Université Paris 8 / Département Hypermédias.*
*** Université Paris 8 / Département Documentation.*
**** Central China Normal University – Wu Han, République Populaire de Chine.*
***** Université Paris 7 / Langues et civilisations orientales.*

RÉSUMÉ. *Le laboratoire Paragraphe a développé des méthodes d'exploration de corpus à partir de cartographies textuelles, qu'il a étendu à des langues et écritures différentes par la technique des n-grammes. Un prototype de moteur de recherche couplant approche multilingue par les n-grammes et approche linguistique est présenté, qui permet à la fois des requêtes fines utilisant une analyse linguistique, et des vues d'ensemble à la demande au moyen de cartographies.*

ABSTRACT. *The Paragraphe lab of Université Paris 8 has developed browsers for text databases based upon automatic information mapping, recently extended to languages other than french , and writings other than roman characters by means of a N-gram technique. We present here a prototype Web search engine grounded on these developments, which enables both precise queries, by means of a linguistic processing of the texts, and overviews of the contents by means of interactive information maps.*

MOTS-CLÉS : *moteur Web, cartographie de l'information, navigation hypertextuelle, indexation sémantique, multilinguisme.*

KEY WORDS : *information retrieval, Web browser, information mapping, semantic indexing, multilingual retrieval.*

## 1. Origine du projet et état de l'art

Les moteurs de recherche d'informations sur le Web présentent des limites que nous analysons comme suit :

1) Ils sont basés sur la notion de chaîne de caractères ; une chaîne représente en première approximation un mot, unité élémentaire de sens délimitée par des séparateurs triviaux, comme l'espace ou les signes de ponctuation.

2) L'absence d'analyse morpho-syntaxique qu'on y constate crée beaucoup de bruit :

— confusions entre mêmes formes graphiques de mots différents au sein d'une même langue,

— formes graphiques identiques au sein de langues ou écritures différentes (ex : EAU en français et en coréen...) (ceci sans compter les aspects techniques : volumes énormes de fichiers inverses pour accéder à l'ensemble des formes graphiques répertoriées, problème qui deviendra de plus en plus crucial quand le Web passera des 200 millions de pages actuelles à un milliard...).

3) L'approximation de la notion de mot par une chaîne de caractères, déjà problématique pour l'anglais, les langues latines et l'arabe, où en plus des conjugaisons et déclinaisons les fautes d'orthographe et variantes non stabilisées sont monnaie courante, est franchement inadaptée pour :

— les langues agglutinantes (allemand, russe, turc...) où les unités de sens élémentaires sont accolées pour former des mots composés parfois très longs,

— les écritures asiatiques (chinois, japonais, coréen, ...) où les mots sont accolés sans séparateurs, et décodés en fonction du contexte, comme on le fait dans nos pays pour lire les inscriptions latines. Le problème est compliqué par l'existence des deux systèmes de codages des caractères chinois (Big 5 et GB, l'Unicode restant pratiquement inutilisé), et leur mélange avec les caractères occidentaux.

### *1.1. La recherche d'informations sur le Web se fait selon deux modes principaux, mal et inégalement satisfaits à l'heure actuelle :*

*1.1.1. Un mode de recherche fine* d'une information précise, ciblée sur un objectif bien cerné : il s'opère à l'heure actuelle en la décrivant par un ou plusieurs mots dont on laisse au sort le soin de déterminer s'ils correspondent à une graphie recouvrant ce seul mot dans notre seule langue, ou bien s'il nous rapporte bien d'autres choses...

*1.1.2. Une démarche exploratoire*, où l'on cherche à "se faire une idée" de ce que le Web recèle sur un certain sujet d'intérêt : seuls AltaVista/Refine et Semiomap offrent à l'heure actuelle un début de réponse à ce besoin.

L'objectif général de notre projet est de fournir des réponses améliorées, cohérentes et complémentaires à chacun de ces besoins, validées par la mise en place d'un prototype de moteur de recherche donnant accès à quelques dizaines de milliers de pages Web en français et chinois, totalisant quelques centaines de mégaoctets de textes.

**2. Nos acquis :**

Le laboratoire Paragraphe travaille depuis de nombreuses années dans le domaine de la réalisation d'hypertextes et hypermédias avancés, de l'automatisation de la création de liens hypertextuels, de la génération automatique de textes, et des bases de données multimédias. Il a organisé plusieurs conférences internationales sur ce sujet, et est à l'origine de la revue "Hypertextes et hypermédias", ainsi que de plusieurs ouvrages de référence. Voici quelques-unes de ces réalisations sur lesquelles s'appuie notre projet :

1) HYPERMAP, environnement d'assistance à l'indexation de corpus volumineux encapsulant un module de lemmatisation, et réalisant entre autres les fonctions de :
- détection statistique des mots composés,
- filtrage statistique des lemmes trop rares, ou trop fréquents quand ils sont uniformément répartis (critère du "khi-deux+", cf. [RHI 97]),
- mise à jour incrémentale de l'indexation et de la base de documents.

2) NEURONAV+, interface convivial de navigation cartographique dans une base textuelle indexée manuellement ou par Hypermap, et de mise à jour dynamique de l'indexation et du corpus [LEL 97].

3) ENGRAMMES, logiciel de cartographie de corpus textuels, indépendant de la langue et de l'écriture, à partir du codage des documents sous forme de profils de fréquences de n-grammes, et de visualisation des résultats [HAL 97 ; LEL 98].

4) PROXILEX, logiciel de recherche rapide des mots, simples et composés, les plus proches lexicalement d'une chaîne de caractères fournie en entrée, ou incluant cette chaîne [HAL 99].

**3. Nos objectifs :**

En réponse à nos critiques adressées aux systèmes de recherche existants, nos objectifs sont les suivants :

- Concernant le besoin de recherche fine : compléter la recherche sur chaîne de caractères par la recherche sur lemmes et expressions composées lemmatisées, ainsi que par champ sémantique de ces lemmes et expressions (obtenu par « expansion

statistique »), pour la langue française. Il est clair que cette approche ne convient pas à une démarche d'exploration "pour voir", ou de demande de synthèse sur un sujet étendu donné.

- Satisfaire la demande exploratoire à partir de deux outils :

— La *cartographie de base de l'ensemble des pages Web* explorées par le moteur, offre des regroupements thématiques de ces pages pour les langues et écritures représentées dans le corpus de test, spécialement le français et le chinois, et permet de satisfaire les requêtes de proximité par rapport à une question détaillée en langage naturel ou une page Web spécifiée en tant que requête. Il est clair que, si cette cartographie peut donner un point de vue d'ensemble sur les thèmes traités dans les pages répertoriées par le moteur, elle ne saurait fournir la réponse précise à un sujet exprimé par un mot précis.

— La *cartographie à la demande* déclenchée à partir des résultats d'une requête telle que décrite ci-dessus, donc indépendante de la langue, ou d'une requête sur des lemmes pour le français. Un environnement graphique de navigation « triangulaire » (carte des thèmes, mots, documents) est alors téléchargé par l'utilisateur, qui peut l'explorer à loisir sur son ordinateur personnel.

## 4. Architecture d'ensemble

Elle s'établit autour de deux applications : l'application « Administration des données » créée et met à jour la bases de données et les fichiers nécessaires aux divers modules de l'application interactive ; ceux-ci sont chargés sur le serveur, et se trouvent en interaction avec les usagers finaux, à travers le navigateur de ces derniers, ainsi que via les scripts CGI et l'applet Java spécifiques de notre moteur de recherche.

Selon la langue et le corpus choisis (français ou chinois), les expansions statistiques de requête (sur un lemme ou sur un document) et les cartographies sont réalisées à partir des comparaisons de vecteurs de fréquences de lemmes, dans le cas du français, ou de vecteurs de fréquences de bigrammes, dans le cas du chinois. Nous avons testé actuellement, dans le cadre du projet Amaryllis d'évaluation de systèmes de recherche d'informations francophones, la comparaison des images de ces vecteurs dans un espace à nombre réduit de dimensions, défini par l'extraction de thèmes sur l'ensemble des pages par notre module Neuronad [LEL 94]. Cette réduction de l'information à l'« essentiel » rend ainsi proches des documents sans mots communs, mais traitant d'un même sujet – pour peu que ce sujet ressorte de la cartographie -, à l'instar de ce que réalise, d'une façon plus lourde informatiquement , et « en aveugle », sans possibilité d'interpréter le sens des composantes

élémentaires latentes, la procédure Latent Semantic Indexing [DUM 94]. Notre approche revient donc à constituer une indexation sémantique *patente*.

### *4.1. Application « administration des données »*

#### *4.1.1. Programme principal*

Il réalise les fonctions de collationnement initial et incrémental des pages Web, au fur et à mesure des campagnes de collecte, et d'appel des modules créant les fichiers à charger, sans oublier les vérifications de cohérence.

#### *4.1.2. Module Agent Web de collationnement et de filtrage des pages*

Ce module explore et rapatrie les pages d'une liste de sites donnée, avec une profondeur d'exploration paramétrable, et dans le respect de la déontologie définie sur le Web. Chaque page est alors débarrassée de ses balises HTML, puis filtrée et pré-traitée en vue de la suite des opérations (par ex. : les écritures asiatiques mélangent les codes-caractères sur un octet et sur deux octets…). Des statistiques sont recueillies sur chaque page : nombre de caractères hors-balises, nombre de liens externes, internes, …

#### *4.1.3. Module TypoWeb de catégorisation automatique des pages Web*

Il est d'observation courante que beaucoup de pages Web ne sont pas justiciables d'un traitement d'analyse de contenu : annuaires, listes de sites, pages « carrefour », listages de programmes informatiques, … Il est donc essentiel d'établir une typologie des pages Web, afin 1) d'informer l'utilisateur à l'avance du type des pages auxquelles aboutit sa requête, 2) d'éliminer de la cartographie les pages non-textuelles à proprement parler. Alors que [BOR 98] propose des critères issus d'une étude manuelle d'un nombre de pages limité, nous utilisons pour identifier de façon adaptative les « pages textuelles » notre module Neuronad, qui s'est révélé aussi efficace pour traiter des vecteurs quantitatifs à peu de dimensions qu'avec des données textuelles très multidimensionnelles. A partir d'indicateurs quantitatifs recueillis par l'agent Web (nombre de caractères après débalisage, nombre de liens (internes à la page, internes au site, externes) nombre de tableaux, nombre d'images, nombre de formulaires, nombre de balises « mail to »), cinq types de pages ont été ainsi distingués automatiquement, et leur degré de « typicité » visualisé par une échelle à trois degrés (*, **, ***) – en effet ces cinq catégories constituent des pôles flous, plus que des classes bien distinctes :

- page informative textuelle (20% des pages d'un échantillon représentatif du Web francophone),
- page informative avec texte illustré (27%),
- page carrefour interne au site (46%),

- page carrefour externe au site (8%),
- page interface de saisie.

*4.1.4. Module Hypermap-Web d'indexation lemmatisée par mots simples et composés*

Ce module fait d'abord appel à un lemmatiseur externe : nous avons choisi Nomino [PLA 91], qui propose 1) des candidats lemmes et termes composés de bonne qualité, 2) et ce sur une base à la fois linguistique et ouverte, mieux adaptée au caractère incrémental indispensable de tout stockage de pages Web, 3) la reconnaissance automatique des pages principalement en langue anglaise, qui parsèment bon nombre de sites francophones, et leur traitement. Notre module permet ensuite de réaliser des élimination de lemmes et termes composés par masses, à partir de leur nature grammaticale (mots grammaticaux, verbes, adjectifs, …), ou de critères statistiques (fréquence, uniformité de répartition, …).

*4.1.5. Module Engrammes-Web multilingue et multi-écritures*

Dans le cadre de notre expérience sur les sites chinois, chaque page Web, après avoir été récupérée par le robot de collecte, et filtrée (codes HTML, ...), est transformée en un profil d'occurrences de n-grammes hash-codés. Celui-ci est alors soumis à notre algorithme de classification. Les paramètres du filtrage, du hachage et de l'algorithme de classification sont cruciaux pour la qualité de la « quantification vectorielle du Web » obtenue, et font l'objet d'une mise au point soigneuse en interaction avec le travail d'évaluation.

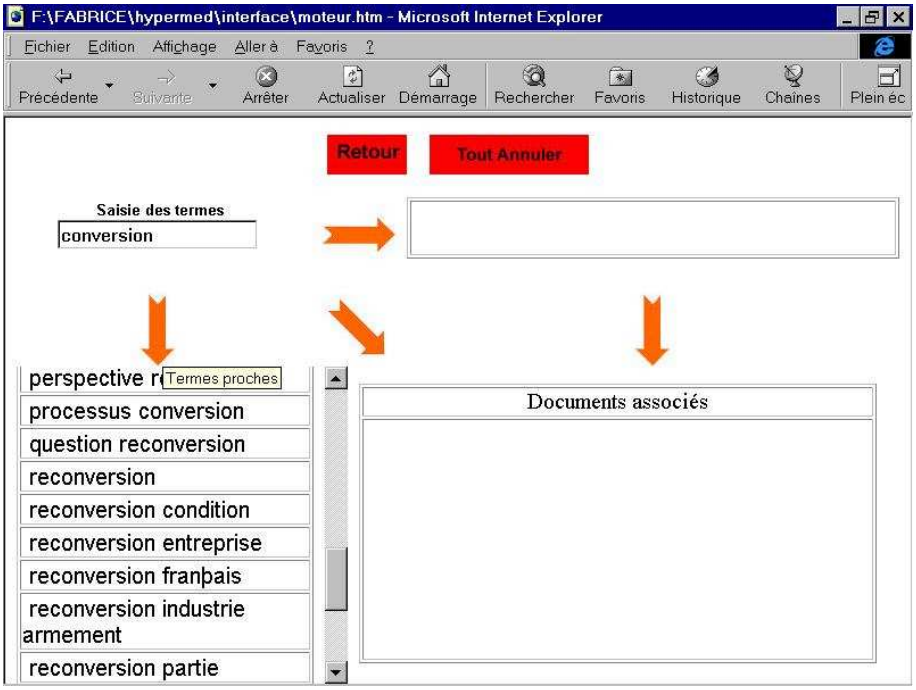



**Figure 1.** *Requête de proximité lexicale dans le moteur NeuroWeb*

Afin de donner une idée plus concrète de notre démarche, citons un exemple de résultats obtenus sur un corpus d'un millier de textes chinois courts, tirés de l'épopée mythologique et satirique du 16eme siècle « Le pèlerinage vers l'Ouest ». Notre logiciel Engrammes en a extrait une vingtaine de thèmes, qu'on peut interpréter comme autant de schèmes récurrents, de « motifs » [THO 55] élémentaires qui reviennent souvent dans ce récit, sous des formes et dans des contextes variables : le moine-héros Tripitaka et ses trois disciples arrivent dans un temple pour s'y reposer, le Roi-Singe s'arrache un poil qu'il transforme en un petit monstre, le Roi-Singe se bat contre un monstre et bondit dans les airs, … Une liste de bigrammes chinois caractérise chacun de ces thèmes ; certains de ces bigrammes combinés forment des mots de trois ou quatre caractères, ou des morceaux de phrases, qui constituent ainsi une pré-indexation automatique de ces textes, pas toujours correcte syntaxiquement, mais suffisante pour accéder au contenu [HAL 99b].

### *4.2. Application interactive*

*4.2.1. Programme client Web : interface utilisateur*

Compte tenu de la grande diversité des requêtes possibles sur notre moteur, nous nous appliquons à respecter au plus près le principe ergonomique consistant à ne présenter à l'utilisateur que les seules possibilités d'action qui ont un sens à l'instant *t*, munies de libellés les plus explicites possibles. Ce qui se traduit techniquement par la définition d'un automate de dialogue, décrivant à chaque instant l'état de l'écran, les actions possibles avec les transitions correspondantes vers d'autres états. Par exemple, la figure 1 montre l'écran à un moment du dialogue. A d'autres moments, le sens d'une flèche active aura pu changer, d'autres zones, libellés ou boutons auront pu apparaître ou disparaître.

Après une sollicitation, ici après la recherche des termes lexicalement proches du terme *conversion*, on obtient l'écran montré figure 1 ci dessus.

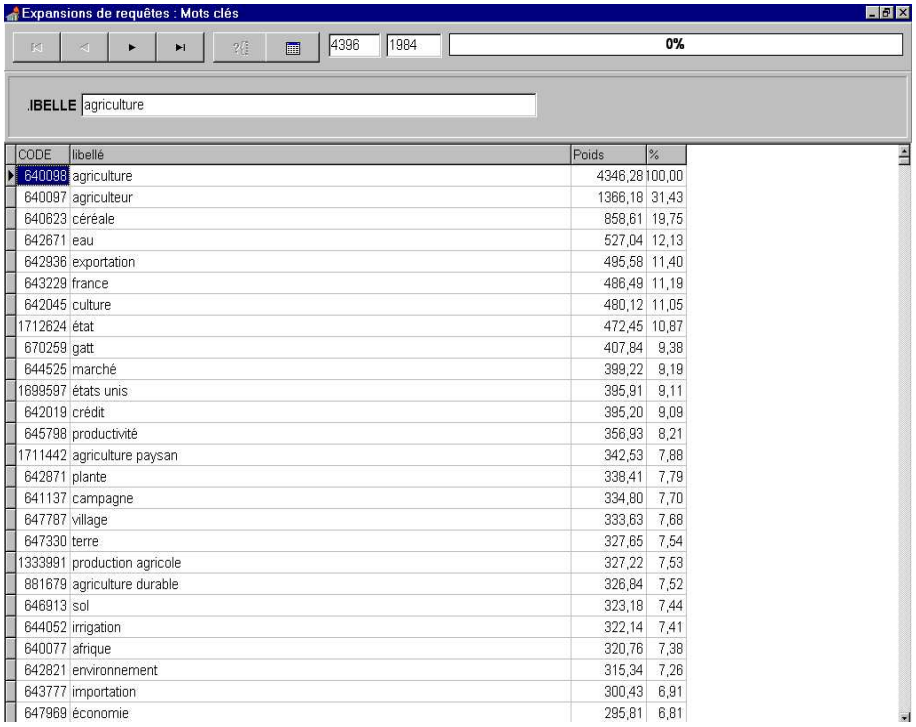
*4.2.2. Liste des requêtes :*

- *Requête de proximité lexicale* : l'utilisateur tape dans la zone « saisie » une chaîne de caractères, envoyée au module Proxilex [HAL 99] qui renvoie la liste des lemmes simples ou composés reconnus du système, les plus proches lexicalement de la chaîne en requête (cf. exemple fig. 1 plus haut)

- *Requête mot vers pages* : pour un lemme sélectionné et mis dans la zone « requête », le système renvoie les url qui contiennent une des formes de ce lemme.

- *Requête page vers mots* : pour une url sélectionnée et mise dans la zone « requête », le système renvoie les lemmes qui caractérisent cette url.

- *Requête d'expansion d'un mot* : pour un lemme sélectionné et mis dans la zone « requête », le module Expansion renvoie les lemmes simples ou composés les plus proches sémantiquement de ce lemme (ici : les plus souvent associés statistiquement). La figure 2 montre un exemple d'une telle expansion.




| CODE | libellé | Poids | % |
|---|---|---|---|
| 640098 | agriculture | 4346,28 | 100,00 |
| 640097 | agriculteur | 1366,18 | 31,43 |
| 640623 | céréale | 858,61 | 19,75 |
| 642671 | eau | 527,04 | 12,13 |
| 642936 | exportation | 495,58 | 11,40 |
| 643229 | france | 486,49 | 11,19 |
| 642045 | culture | 480,12 | 11,05 |
| 1712624 | état | 472,45 | 10,87 |
| 670259 | gatt | 407,84 | 9,38 |
| 644525 | marché | 399,22 | 9,19 |
| 1699597 | états unis | 395,91 | 9,11 |
| 642019 | crédit | 395,20 | 9,09 |
| 645798 | productivité | 356,93 | 8,21 |
| 1711442 | agriculture paysan | 342,53 | 7,88 |
| 642871 | plante | 338,41 | 7,79 |
| 641137 | campagne | 334,80 | 7,70 |
| 647787 | village | 333,63 | 7,68 |
| 647330 | terre | 327,65 | 7,54 |
| 1333991 | production agricole | 327,22 | 7,53 |
| 881679 | agriculture durable | 326,84 | 7,52 |
| 646913 | sol | 323,18 | 7,44 |
| 644052 | irrigation | 322,14 | 7,41 |
| 640077 | afrique | 320,76 | 7,38 |
| 642821 | environnement | 315,34 | 7,26 |
| 643777 | importation | 300,43 | 6,91 |
| 647969 | économie | 295,81 | 6,81 |

**Figure 2.** *Expansion statistique du terme « agriculture ».*

- *Requête d'expansion d'une page* : pour une url sélectionnée et mise dans la zone « requête », le module Expansion renvoie les url les plus proches sémantiquement de cette url, c'est à dire partageant un maximum de lemmes ou termes composés avec celle-ci, ou plus généralement un maximum de thèmes sémantiques (cf. fig. 3 pour un exemple).

- *Requête de cartographie* : pour les url de type « informatives » résultant d'une requête et remplissant la zone « pages », l'action sur le bouton *Cartographier* déclenche l'exécution du module Neuronad pour ces pages, et la transmission des résultats au module Java de navigation dans l'environnement « triangulaire » de navigation au sein des pages sélectionnées, et des thèmes et mots qui lui correspondent (cf. fig. 4 pour un exemple).

*Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., (1999) "Projet NeuroWeb : un moteur de recherche multilingue et cartographique", 5ème conférence Hypertextes et Hypermédias : Réalisations, Outils & Méthodes, H2PTM'99, Paris, 23 et 24 Septembre 1999.*

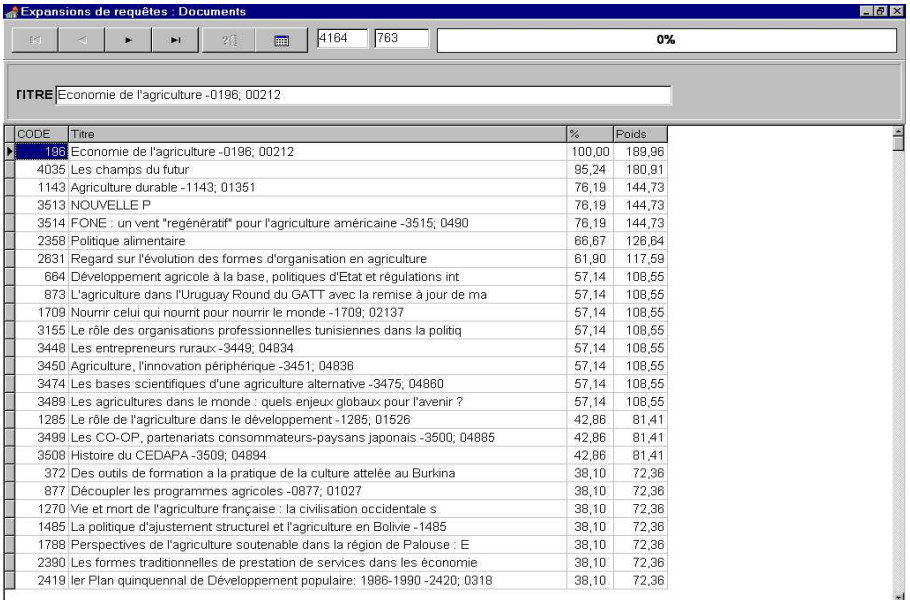



**Figure 3.** *Requête d'expansion du document « Economie de l'agriculture ».*

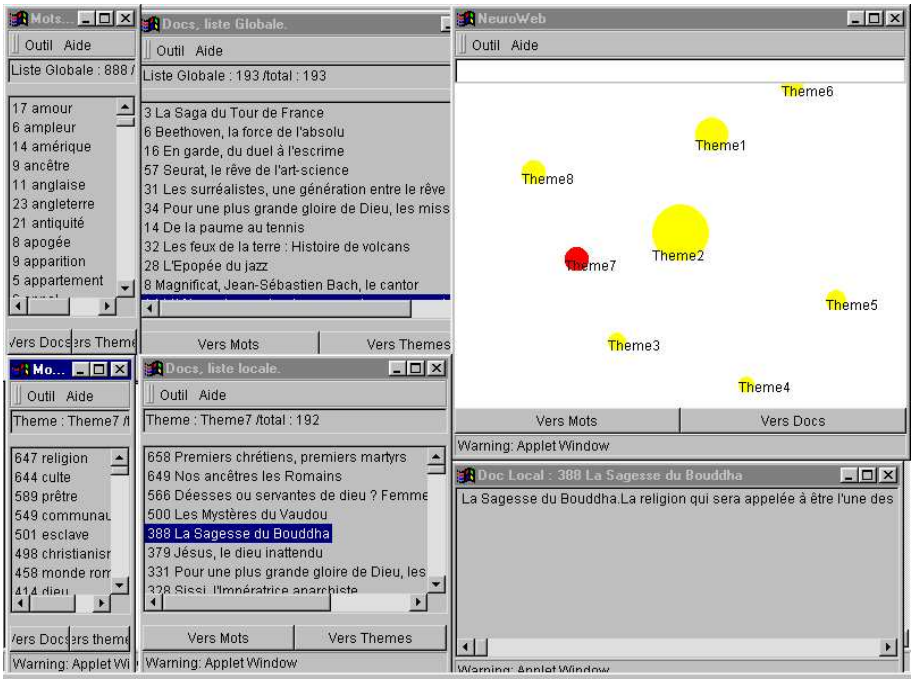


**Figure 4.** *Notre Applet Java de navigation dans une cartographie textuelle.*

— L'association d'un mot à des mots proches ou à des concepts rattachés, permet une recherche itérative, dans le corpus, de pages traitants de thèmes proches sans utiliser nécessairement le premier mot-clef entré : la recherche est à la fois ouverte et affinée, et fournit un ensemble de pages pour lequel on propose une cartographie.

— La présentation cartographique du corpus issu d'une requête, constitue une visualisation en 2D de son contenu sémantique, ainsi que des pages correspondant à chaque thème.

A la différence de la fonction Refine (proposée sur Alta-Vista), 1) seules les pages réellement informatives sont cartographiées, 2) les thèmes renvoient directement à un ensemble de pages associées, et 3) la carte des thèmes peut être "allumée" par les mots ou pages Web désignés par l'utilisateur, ce qui lui permet de situer ses centres d'intérêt dans l'espace sémantique de la carte.

### *4.3. Programme serveur : la « plaque tournante »*

#### *4.3.1. Communications Client / Serveur*

Le graphique suivant (fig. 5) illustre les mécanismes de communication entre le poste client et le moteur de recherches. Toute la partie transactionnelle est gérée par des formulaires qui après leur soumission par l'utilisateur sont transférés vers le serveur de requêtes HTTP, qui opère le traitement ad hoc au moyen du script CGI déterminé par le formulaire soumis. Le serveur Web reprend alors les éléments de réponse fournis par le moteur de recherches, les structure au format HTML puis les adresse au poste client.

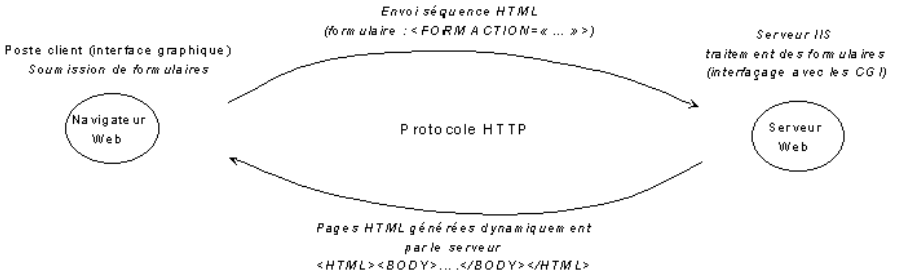



**Figure 5.** *Schéma des communications client-serveur.*

*4.3.2. Le serveur : un environnement hautement collaboratif*

Le dispositif rassemblant les différents éléments du moteur de recherches est hébergé dans sa totalité par le serveur de ressources. C'est le serveur de requêtes HTTP qui assure la distribution des demandes d'information transmises par les clients aux différents modules composant le moteur de recherches.

Le moteur est lui même constitué de modules indépendants qui peuvent dialoguer de manière privilégiée avec le serveur HTTP ou alors former un environnement collaboratif avec les autres modules en fonction du type de requêtes soumises. Les différents modules : approximation lexicale, expansion, cartographie dynamique, N-grammes, …sont présents sur le serveur de la façon présentée figure 6.

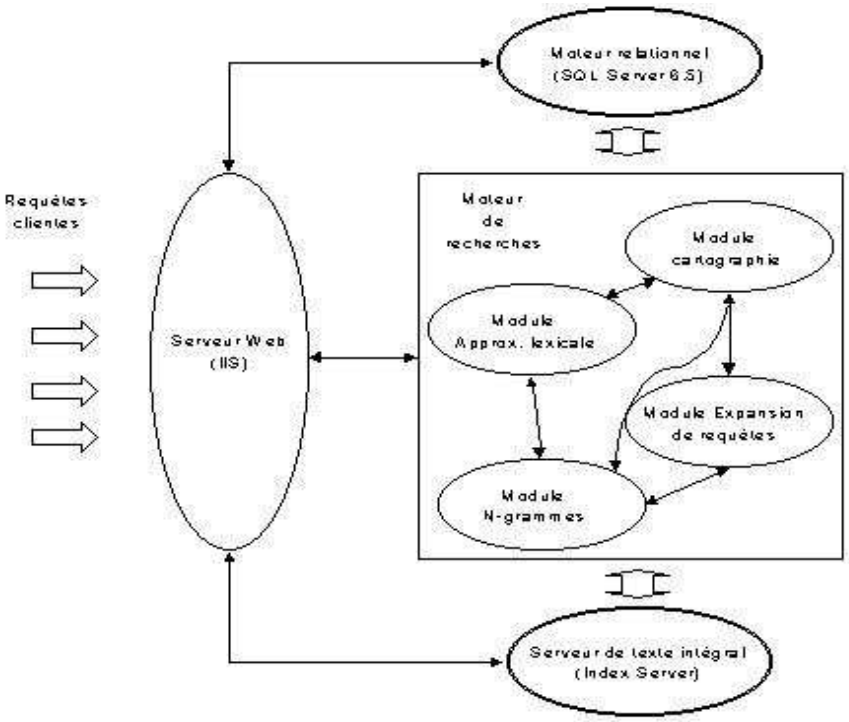



**Figure 6.** *Organisation modulaire du logiciel serveur.*

*4.3.3. Une organisation optimale de la mémoire*

Afin de garantir des délais de réponse courts, la collaboration des différents modules du moteur de recherche exploite les segments de mémoire partagée (*shared memory*) implémentés sous la forme : mappage par fichiers sous Windows NT (*file mapping*).

*4.3.4. Interfaçage avec les scripts CGI*

L'activation des différents modules du moteur de recherches sera commandée par les appels à des scripts CGI (Visual basic, Perl, API,…) gérés au moyen d'une file d'attente (*spooler*).

## 5. Conclusion

Pour résumer, nous dirons que notre moteur Web tente d'associer le meilleur des procédures de recherches d'information publiées aujourd'hui, par nous-mêmes et par d'autres :

. recherche monolingue sur la base d'une indexation par mots lemmatisés et expressions composées,

. recherche rapide dans la liste des termes à partir de proximités lexicales basées sur les N-grammes,

. proximités sémantiques dans l'espace de dimensions réduites, et à coordonnées obliques, issu du passage Neuronad sur l'ensemble des données (indexation sémantique patente, et non latente).

. recherche dans des corpus de langue et écriture exotiques à partir des N-grammes,

. cartographie des thèmes extraits d'un sous-ensemble de pages issues d'une requête de l'utilisateur, et exploration de cette carte en tous sens.

Cette réalisation donnera lieu à des démonstrations au moment de la conférence, et sera accessible sur notre site Web (<*www.hypermedia.univ-paris8.fr/equipe*>).

## Remerciements

Nous tenons à remercier le Ministère de l'Enseignement Supérieur, de la Recherche et de Technologie pour le financement de notre projet « Moteur de recherche multilingue et cartographique sur Internet », ainsi que l'action de coopération scientifique Franco-Québecoise « Outils de filtrage et de routage d'informations sur Internet ».

## Bibliographie :

[BOR 98] BORZIC B, « Un modèle de gestionnaire itératif de flux informationnel sur Internet », Thèse de doctorat, Information Scientifique et Technique, CNAM, Paris, Mars 1998

[DUM 94] DUMAIS S.T., « Latent Semantic Indexing (LSI) and TREC-2 », *NIST special publication*, N°500 215, pp.105-115, 1994.

[HAL 97] HALLAB M., LELU A., « Hypertextualisation multilingue à partir des fréquences de N-grammes », *4e conférence Hypertextes et hypermédias*, Université Paris 8, Saint Denis, 1997.

[HAL 98] M. HALLAB, A. LELU, B. DELPRAT, « Towards a New Type of Search Tool and Multilingual Text Cartography on the World Wide Web », *Actes de CESA'98 IMACS-IEEE Multiconference*, Hammamet, Tunisie, 1-4 Avril 1998

[HAL 99] HALLAB M., LELU A., « Proxilex : un outil d'approximation orthographique à partir des fréquences de N-grammes », dans ces mêmes actes, 1999.

[HAL 99b] HALLAB M., LELU A., « Indexer quelle que soit la langue et l'écriture : une approche combinant N-grammes et cartographie textuelle », communication acceptée au *2e colloque du chapitre français de l'ISKO*, Lyon, 21-22 octobre 1999

[LEL 94] LELU A., « Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets », *New Approaches in Classification and Data Analysis,* E. Diday, Y. Lechevallier & al. eds., Springer-Verlag, Berlin, 1994, pp.241-248

[LEL 97] LELU A., TISSEAU-PIROT A.G., ADNANI A., « Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation. », *Hypertextes et Hypermédias*, vol.1, N°1, éditions Hermès, Paris, 1997

[LEL 98] LELU A., HALLAB M., DELPRAT B., "Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes", *Actes des 4emes Journées Internationales d'Analyse Statistique des données Textuelles*, coord. S. Mellet, UPRESA « Bases, Corpus et Langage », Université de Nice, 1998

[LEL 99] LELU A., « Représentations cartographiques de corpus textuels : codages, algorithmes, ergonomie », Mémoire d'Habilitation à Diriger les Recherches, Université Paris 8, 1999

[PLA 91] PLANTE P., « La modélisation en faisceaux pour le parsage syntaxique », *Actes du colloque ILN'91*, Nantes, Janvier 1991. Repris aussi dans un Document Centre ATO.CI, Octobre 1992. Cf. <*www.ling.uqam.ca/nomino*>

[RHI 97] RHISSASSI H., LELU A., « De l'indexation à la navigation : vers un environnement complet d'hypertextualisation assistée », *4e conférence Hypertextes et hypermédias*, Université Paris 8, Saint Denis, 1997.

*Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., (1999) "Projet NeuroWeb : un moteur de recherche multilingue et cartographique", 5ème conférence Hypertextes et Hypermédias : Réalisations, Outils & Méthodes, H2PTM'99, Paris, 23 et 24 Septembre 1999.*

[RHI 98] RHISSASSI H., LELU A., « Indexation assistée et cartographie sémantique pour la génération automatique d'hypertexte », *Actes de la conférence CIDE'98*, coord. M. Mojahid, INPT, Rabat, Maroc, 15-17 Avril 1998, Europia Productions, pp.131-139.

[THO 55] THOMPSON S., *Motif-index of Folk-Literature. A Classification of narrative elements […]*, Indiana University Press, 1955.