HAL
open science

Navigation et recherche par catégorisation floue des pages HTML

F. Papy, N Bouhaï

To cite this version:

F. Papy, N Bouhaï. Navigation et recherche par catégorisation floue des pages HTML. JFT(Journée Francophones de la Toile) 2003, Jun 2003, Tours, France. hal-02091642

HAL Id: hal-02091642

https://univ-paris8.hal.science/hal-02091642v1

Submitted on 5 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Papy F., Bouhaï N., (2003), "Navigation et recherche par catégorisation floue des pages HTML", Actes des JFT'2003, 30 juin au 2 juillet 2003, Ecole Polytechnique de l'Université de Tours

Navigation et recherche par catégorisation floue des pages HTML

F. PAPY, N. BOUHAÏ
Laboratoire Paragraphe,
Université Paris VIII
2, rue de la Liberté ,
93526 Saint-Denis, FRANCE
Mail :fabrice.papy@univ-paris8.fr
nasreddine.bouhai@univ-paris8.fr
Tél : +33 1 49 40 67 58 Fax : +33 1 49 40 67 83

Résumé

Dans le but d'améliorer la recherche et la navigation sur le Web, nous proposons une approche inédite fondée sur la classification de pages. Nous prenons en considération les balisages utilisés dans les pages Web pour élaborer des profils.. Pour établir cette catégorisation de classification automatique des pages, nous nous sommes appuyés sur les travaux d'Alain Lelu en utilisant l'algorithme de K-means axiales. Utilisée dans le moteur de recherches NeuroWeb, cette méthode de sélection automatique de pages a été transposée dans notre dispositif de construction d'espaces de connaissances HyWebMap.

Abstract

In order to improve search and navigation through the Web, we propose an original approach based upon pages classification. We use tags embedded in web pages to build specific profiles. To produce this automatic categorization, we were inspired by Alain Lelu' s research on K-means axial algorithm. Initially used into the search engine Neuroweb, this method was adapted for our knowledge buildin tool ; HyWebMap.

1 Introduction

Internet est une source d'informations stratégiques que l'on peut aujourd'hui difficilement nier. Les entreprises, les organisations gouvernementales ou non sont ainsi confrontées à la pléthore d'informations, à l'abondance des processus nouveaux et à leur rapide obsolescence. Dans ce contexte, s'impose la nécessité d'une collecte et d'une transmission sélective de l'information, de dispositifs permettant de la trouver, de la filtrer et de la traiter automatiquement. Car, le meilleur côtoie aujourd'hui le pire : sites institutionnels des gouvernements, des universités, des centres de recherches sont proposés dans les listes d'URL des moteurs de recherches au même titre que les pages personnelles ou les sites commerciaux. Séparer le bon grain de l'ivraie est une tâche de plus en plus difficile malgré les fonctionnalités de plus en plus sophistiquées des systèmes de recherches d'informations en ligne qui outre le volume, les redondances…doivent répondre aux problèmes du multilinguisme. A l'heure de la surinformation sur les réseaux, la nécessité d'opérer une sélection s'impose donc comme un enjeu déterminant du traitement de l'information.

Les technologies de filtrage et de résumé apportent une réponse aux professionnels de l'information, cyber-documentalistes, courtiers, veilleurs, webmasters, responsables de portails d'information, etc. dans des secteurs aussi différents que le knowledge management, le e-commerce et l'intelligence économique.

La recherche d'informations repose assurément sur une relation étroite entre les possibilités opératoires des outils d'extraction de données et la capacité de l'utilisateur à s'impliquer dans sa recherche. Cette implication sous-entend une volonté de chercher qui s'exprime au travers de stratégies de recherches visant à évaluer, comparer et confronter les résultats.

Développé dans le cadre du laboratoire Paragraphe de l'Université Paris 8, le projet de moteur de recherches NeuroWeb *(http://neuroweb.univ-paris8.fr)* avait pour objectif un dispositif d'exploration et de recherche fine sur le Web en couplant une approche multilingue (utilisant les N-grammes) avec une approche linguistique / sémantique.

Ce projet, retenu à la suite d'un appel d'offres lancé par le Ministère de l'Éducation Nationale et l'Agence Nationale pour la Recherche Technologique, se proposait d'intégrer dans un prototype de moteur de recherche permettant requêtes fines et cartographies à la demande, deux approches complémentaires ; la première exploitant les méthodes d'exploration de corpus à partir de cartographies textuelles utilisant les N-grammes, la seconde utilisant la lemmatisation de textes et les réseaux sémantiques.

Ce moteur de recherche alimenté par un robot de téléchargement de sites développé spécifiquement, est doté d'une série de modules déclenchés circonstanciellement en fonction des actions opératoires lancées par l'utilisateur à partir du navigateur (texte intégral, approximation lexicale, interrogation sur lemme, expansion de requêtes lemme ⇒ lemmes et document ⇒ documents, N-grammes dotés de fonctions de proximité lexicale, cartographie dynamique)

Dans le cadre de ce projet NeuroWeb, nous avons sélectionné les sites d'information devant alimenter le moteur en utilisant une méthode originale pour réduire l'espace de recherche sur le web en classifiant automatiquement les pages HTML. Nous proposons de prendre en considération les balisages utilisés dans les pages pour construire les profils des pages Web. Cette approche est fondée sur les caractéristiques de pages HTML. Cette catégorisation permet alors :

- d'améliorer les navigations en réduisant l'espace de recherche en montrant seulement les pages pertinentes par rapport aux souhaits de l'usager,
- d'éviter la situation de surcharge cognitive à laquelle l'usager est souvent confronté au fil de ses lectures,

- de signaler à l'usager les types de pages auxquels aboutit sa requête,
- de donner des possibilités à l'usager de filtrer et de choisir les types de pages qu'il désire consulter.

Utilisée par les agents arpenteurs afin de collecter des sites "homogènes", nous avons transposé cette méthode dans notre logiciel de construction d'espaces de connaissances virtuels HyWebMap.

2 Type d'informations sur le Web

Il existe plusieurs approches pour aider l'utilisateur à naviguer sur le Web mais aucune ne prend en considération la notion de profil syntaxique des documents. Pourtant ces profils permettent d'identifier les types de données qu'ils contiennent. Les balisages utilisés dans les documents écrits par exemple en HTML, fournissent explicitement ces types de données.

Nous proposons de prendre en considération les balises des pages pour construire les profils des pages Web. Pour établir une catégorisation de classification automatique des pages Web, nous nous sommes appuyés sur les travaux d'Alain Lelu en utilisant l'algorithme de K-means axiales [Lelu 99], [Balpe & al. 96].

Les documents sur le Web sont hétérogènes (sites commerciaux, pages personnelles, livres, articles, annuaires), ne possèdent aucune véritable structure. Les sources d'informations sont diverses, ainsi que leurs types. [Bélisle et al., 99] distinguent plusieurs grands types d'information :

· Information publique de référence, provenant des gouvernements, d'organismes professionnels, de bibliothèques, d'associations, ou de sociétés privées.

· Information scientifique et éducative (disciplinaire), dont les banques de données traditionnelles, provenant de laboratoires de recherche, d'universités, ou de sociétés de services.

· Information publicitaire, visée commerciale provenant des entreprises.

· Information médiatique, provenant des organismes des presses.

· Information personnelle, provenant des individus ayant leur propre site.

Cette distinction est floue car certains sites proposent plusieurs types d'informations. Les contenus des sites peuvent varier d'un site à un autre par rapport aux objectifs de chaque site. Nous distinguons trois catégories de sites Web par rapport à leurs contenus :

· Les sites textuels privilégient les contenus textuels avec plusieurs liens internes et des liens externes car leur objectif est de diffuser les informations auprès des utilisateurs (les sites institutionnels, bibliothèques, universitaires, entreprises). Dans ceux-ci, les images ou les illustrations offrent des informations complémentaires et n'interviennent le plus souvent qu'à un deuxième niveau de recherche.

· Les sites visuels : privilégient les contenus visuels (images, graphiques d'illustration, etc.). Ainsi, ils intègrent souvent des formulaires (champs de saisies), par exemple les sites commerciaux, publicitaires, commerces électroniques, musées. L'image joue un rôle important, elle participe à l'attractivité du site et pour les commerciaux, elle est une valeur ajoutée indispensable. Pour les sites "plus techniques", l'image a une fonction différente. Elle permet à l'utilisateur de mettre rapidement ses attentes en correspondance avec l'information présentée. Dans ces sites les textes offrent des informations complémentaires et n'interviennent qu'à un deuxième niveau de recherche.

· Les sites portails (annuaires) : privilégient plutôt les liens externes.

3 Les attentes des utilisateurs

Le Web est un service d'information et de communication d'un contenu selon certaines modalités. L'un et l'autre répondent à des besoins précis des utilisateurs, dans un contexte donné. La qualité du service dépend de façon cruciale de l'identification correcte de ces besoins qui doivent demeurer centraux.

Les besoins changent par rapport aux objectifs de chaque utilisateur, certains souhaitent visiter des pages contenant seulement des images lorsqu'ils visitent un site de musée, ou un catalogue de produit, d'autres souhaitent visiter des pages textuelles lorsqu'ils visitent un site institutionnel, universitaire, ou d'autres souhaitent visiter des pages contenant textes et images. Cela peut se comprendre comme une demande de maîtrise de la recherche de contenus plus qu'une demande brute portant sur le simple accès à l'information.

3.1 Typologies des utilisateurs sur le Web

Après des données recueillies et analysées au centre Georgie Institut of technology les chercheurs Catledge et Pitlow [Catledge & Pitlow 95] proposent trois classes d'utilisateurs, cette analyse illustre la tension entre recherche d'information (query en anglais) et navigation :

· Les utilisateurs appelés "*searchers*" reprennent épisodiquement des séquences courtes mais s'engagent souvent dans des séquences longues.

· Les "*general purpose browser*" n'ont en moyenne qu'une probabilité sur 4 de répéter des séquences complexes (inertie moyenne des usagers).

· Les "*serendipitous browser*" évitent systématiquement de s'engager dans de longues séquences de navigation, ils visitent superficiellement les sites.

Cette analyse montre d'une part que l'utilisateur, explore une zone restreinte à l'intérieur d'un site visité et d'autre part que les utilisateurs ne traversent que rarement plus de deux couches d'hypertexte avant de retourner à leur point d'entrée.

Cette étude suggère que les pages personnelles sont utilisées préférentiellement comme relais dans la navigation et jouent ainsi un rôle d'index vers les autres sites. Ainsi cela illustre la nécessité de bien connaître et comprendre les stratégies navigationnelles des utilisateurs comme base pour la conception de nouveaux logiciels navigateurs.

4 Classification automatique : choix algorithmique

Historiquement, la classification de documents a été utilisée pour améliorer les performances des systèmes de recherche d'informations. L'hypothèse formulée par [Rijsbergen 79] est la suivante : "*…closely associated documents tend to be relevant to the same requests…*". Au lieu de comparer une requête à chaque document d'une base documentaire, le système ne compare la requête qu'avec le modèle représentant chacun des groupes de documents. Les algorithmes actuels autorisent la comparaison de la requête avec chaque document. Aujourd'hui, les méthodes de classification automatique sont utilisées dans plusieurs domaines pour visualiser un ensemble de documents retournés par un moteur de recherche Web.

Les deux méthodes les plus utilisées pour classer les documents sont :

1. **La classification hiérarchique ascendante** : elle possède deux propriétés intéressantes. Tout d'abord, l'utilisateur doit définir le nombre de groupes à obtenir. Ensuite, la méthode induit naturellement une hiérarchie entre les groupes de documents.

Cette propriété est intéressante à condition que la hiérarchie ne soit pas trop profonde. Une hiérarchie trop profonde nuit en effet à la recherche d'informations. [Maarek & Schaul 96] a

proposé de simplifier la hiérarchie en la coupant arbitrairement à des seuils de similarité de 10%, assurant donc une profondeur maximale de 10. Une autre propriété importante concerne la représentation des groupes de documents [Cutting 92]. Les noms de groupe permettent en effet aux utilisateurs de décider quelle branche de classification explorer. Généralement, le nom d'un groupe reflète le contenu thématique des documents qu'il contient. L'approche traditionnelle consiste à sélectionner quelques mots dont l'importance est calculée en combinant la fréquence des mots et le nombre de documents où ils apparaissent. Lorsque le texte intégral est utilisé, des étapes de filtrage sont nécessaires pour choisir un nombre réduit de termes et les présenter à l'utilisateur.

2. **La classification des K-means vers les K-means axiales** : Cette méthode de classification automatique est très ancienne. Connue sous d'autres noms (quantification vectorielle, méthode des centres mobiles, etc.), souvent réinventée et dotée de nombreuses variantes (méthode des nuées dynamiques, ISODATA, etc.), elle est de fonctionnement très intuitif :

· On choisit le nombre K de classes désirées.

· On sème au hasard K points représentatifs des futures classes dans l'espace où sont représentés les objets à classer (par exemple, on détermine au hasard K profils de fréquences de mots, s'il s'agit de classer des documents).

· On présente le premier objet à classer - géométriquement, c'est un point dans cet espace - et on détermine à partir de ses distances aux K centres de classes quel centre est le plus proche : comme la classe que représente ce centre ne contient aucun élément, ce centre est mis à la place du premier objet et symbolise la classe 1.

· On présente le deuxième objet, et on calcule encore ses distances aux K classes. Si la classe la plus proche est la classe 1, le point représentatif de celle-ci est déplacé, dans la direction de l'objet, de la moitié de la distance à celui-ci (la classe contient déjà un élément). Si la classe la plus proche n'est pas la classe 1, le deuxième objet devient le premier élément d'une nouvelle classe.

· On procède de la même façon pour tous les autres objets : si la classe la plus proche d'un objet en contient déjà n, elle est rapprochée en direction de celui-ci d'un n-ième de sa distance [Lelu 93].

En fin de compte, après épuisement des objets à classer, chaque centre de classe représente la position moyenne des points affectés à cette classe. D'où le nom de K-means, signifiant moyennes en anglais.

Alain Lelu [Lelu 93] propose une extension de cette méthode, appelé K-means axiales, le principe est très proche de celui des K-means. La différence réside dans le fait que les K-means axiales contraignent les objets et les centres de classes à se trouver sur une hypersphère de rayon 1, c'est-à-dire à être représentés par des vecteurs normalisés, de longueur 1. Tout se passe comme si on obtenait des axes de classes, et non des points représentatifs de ces classes, au moyen de corrections angulaires successives. En définitive, on obtient un axe par classe, sur lequel on projette les éléments de cette classe : les éléments les plus élevés, dont les projections sont les plus proches de la valeur 1, sont les plus centraux et typiques de la classe. Il est possible de projeter également les éléments qui n'en font pas partie : c'est de cette façon que cette méthode de classification stricte au départ peut nous fournir la représentation " floue" qu'illustre "l'allumage" nuancé des diverses classes par un objet présenté au système.

Autres différences avec les K-means originelles :

· Les K-means axiales représentent les K classes par leurs positions sur une carte globale. Celle-ci est obtenue par une analyse des données "au deuxième degré", par exemple

par une analyse en composantes principales sur le nuage des K points représentatifs des classes.

· Des résultats plus stables et indépendants de l'ordre d'entrée des données sont obtenus par des variantes itératives, non adaptatives, des K-means. La méthode K-means axiales comporte en option une telle variante. Pour la description précise de cet algorithme, nous renvoyons le lecteur à [Balpe & al. 96, annexe 2].

Nous nous sommes arrêtés sur la méthode K-means axiales pour la catégorisation automatique des pages Web parce que :

· de par sa variante, elle présente l'avantage de cumuler - pour une fois - une exécution très rapide avec une occupation très faible d'espace mémoire, ce qui la rend apte à analyser nos colossales bases documentaires avec nos moyens de calcul actuels.

· elle fournit une représentation floue de classes (par exemple, une page Web peut-être à la fois, textuelle et image, etc.).

5 Objectifs

Nous avons défini comme objectif de :

· Permettre une catégorisation rapide d'un ensemble de documents Web issus des sites en ligne.

· Donner la possibilité à un utilisateur lors de la consultation d'un document Web de connaître sa catégorie (texte, graphique, navigation, etc.).

· Mettre en place des outils graphiques de consultation et de navigation permettant d'exploiter cette catégorisation.

5.1 Les étapes méthodologiques

Notre démarche consiste à retenir préalablement des sources d'informations (sites ou pages Web). Nous constituons alors un échantillon de documents représentatifs (articles de journaux, publications, documents techniques, documents amateurs, etc.).

La présence d'un nombre important d'outils de recherche francophones (annuaires et moteurs de recherche) sur Internet rend la tâche plus délicate dans la mesure où les uns et les autres ne proposent pas les mêmes fonctionnalités. Notre choix s'est porté sur l'annuaire Lokace (désormais Nomade), un choix basé sur les données statistiques annoncées par ce dernier.

5.1.1 Recherche et sélection des sources

Le guide francophone Lokace/Nomade proposait 12 catégories générales (principales) au départ, elles-même décomposées en sous-catégories. Ces sous-catégories sont découpées en un nombre variable de sous-catégories. Le chiffre annoncé était de 3000. Le nombre des sites pointés dépasse 20000, couvrant les grands domaines.

L'utilisation de la méthode de classification K-means axiales avec le choix de K classes désirées, nécessite l'utilisation d'une quantité de données très importante pour obtenir un résultat "interprétable". A partir de cette collecte, nous nous sommes fixé de réunir 20000 profils de pages sur l'ensemble des catégories du guide (3000). Pour arriver à ce résultat, nous nous sommes basés sur les données annoncées (nombre de catégories et nombre de sites du guide). Une moyenne de 6 sites par catégorie (20000 sites/3000 catégories), cette moyenne a été vérifiée sur l'une des catégories.

Le choix aléatoire d'une catégorie sur six nous donne environ 500 catégories. En prenant aléatoirement un site sur six par catégorie, on aura environ 500 sites. Il faut prendre en compte un certain nombre d'URLs obsolètes. Le nombre moyen de pages par site est de 70.

5.1.2 Le recueil des données

Les premières données à recueillir sont les URLs des pages HTML qui vont constituer l'échantillonnage documentaire. Cette opération consiste à effectuer un sondage de l'annuaire retenu en appliquant la stratégie de sélection expliquée auparavant. Nous avons utilisé un agent Web pour explorer la structure catégorique de l'annuaire Lokace/Nomade. Les points de départ de cette exploration sont les URLs des 12 catégories principales de l'annuaire. L'agent Web devait sonder la totalité de la structure arborescente.

6 Profil de documents Web

HTML définit un ensemble de balises de base. On cite les balises de structure, puis celles qui permettent d'agencer et de composer du texte. L'autre catégorie de balises est celle qui permet de mettre en place des hyperliens. Une page Web peut être définie par un ensemble de caractéristiques (domaine du site, structure (frames, etc.), liens internes, liens externes, quantité et poids des images intégrées, rapport balise/contenu, …)

On part de l'idée qu'une page HTML peut être intéressante par sa forme descriptive et par son aspect. Celle-ci est intéressante si elle contient des liens vers le site lui même, des liens externes vers d'autres serveurs. Une page Web peut contenir des formulaires ce qui permet de comprendre qu'il s'agit d'une interface de saisie.

Il est aussi important de signaler que le poids d'une page est un élément très significatif car il peut permettre de déduire l'importance du contenu de la page quantitativement. La présence d'images dans une page est un élément qui permet aussi de dégager une idée sur la dimension esthétique de la page.

Une page HTML peut être considérée du point de vue de son contenu réel (contenu et balises HTML) ou de son rendu-écran (dans un navigateur).

En partant des caractéristiques citées auparavant et en observant une page Web sous ces deux angles, Il est possible d'établir le profil d'une page HTML en constituant un vecteur d'informations.

Le profil est construit par une analyse et un traitement statistique de balises HTML. Nous avons sélectionné les données les plus significatives (cf. tableau 6.1.) obtenues à partir de notre échantillon documentaire initial.

Liens locaux	Liens internes	Liens externes	Taille Hors balises	images	tableaux	formul.	mail
0	12	5	24096	7	1	1	0

Tableau 6.1. Exemple de vecteur de données (profil de page Web)

7 Processus de catégorisation

7.1 Catégorisation de base

Les indicateurs quantitatifs recueillis par l'agent Web sont stockés sous forme de matrice (la relation Profil) (cf. tableau 6.1.). Le processus débute par une catégorisation de

l'échantillon de documents collectés auparavant, cela passe d'abord par l'analyse et l'extraction des profils des documents. Les profils sont stockés sous forme de matrice, chaque ligne correspond à un document et chaque colonne correspond à l'un des attributs cités précédemment.

La méthode des classification K-means axiales nécessite le choix de K classes de sorties, pour cela nous avons effectué une catégorisation avec 10 classes. Les résultats obtenus étaient difficilement interprétables. Cette difficulté résulte d'un éparpillement sur plusieurs classes dont certaines sont très proches. L'autre choix consiste à donner 5 classes. Ce choix se justifie par les résultats de la catégorisation manuelle [Borzic 98] sur un échantillon d'une centaine de documents.

Nous avons validé les résultats obtenus lors de cette deuxième catégorisation par une vérification manuelle sur le Web. Une centaine de documents apparus au début, au milieu et à la fin de chaque classe ont été vérifiés, la quantité importante des documents ne permettant pas une vérification complète. Les données statistiques obtenues avancent un pourcentage de 72% de classification pertinente.

L'algorithme des K-means axiales [Lelu 93] permet une projection des éléments d'une classe sur un axe, les éléments les plus élevés, dont les projections sont les plus proches de la valeur 1, sont les plus centraux et typiques de la classe. Concrètement, on obtient une matrice de 5 colonnes, chacune de ces colonnes correspond à une classe. L'identification des classes (texte, graphique, etc.) se fait par l'étude des premières valeurs et les documents qui leurs sont associés.

La figure 5.1 présente le fonctionnement de l'agent Web et du module de classification. 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** (représenté par la lettre T) : Le contenu de la page est un texte.

· **Page informative avec texte illustré** (représenté par les lettres I) : Le contenu de la page est une illustration visuelle, ce peut être des images, des figures, des boutons, etc.

· **Page carrefour interne au site** (représenté par les lettres CI) : le contenu de la page est un ensemble de liens internes au site.

· **Page carrefour externe au site** (représenté par les lettres CE) : le contenu de la page est un ensemble des liens externes au site.

· **Page interface à la saisie** (représenté par la lettre S) : le contenu de la page est un ensemble de champs de saisie.

On peut constater que cette classification est floue car on peut avoir une page entrant dans deux ou trois catégories. Par exemple une page peut être Page informative textuelle alors elle est représentée par l'échelle T*, ou bien une Page informative textuelle avec des liens externes, alors elle est représentée par TCI*, etc.

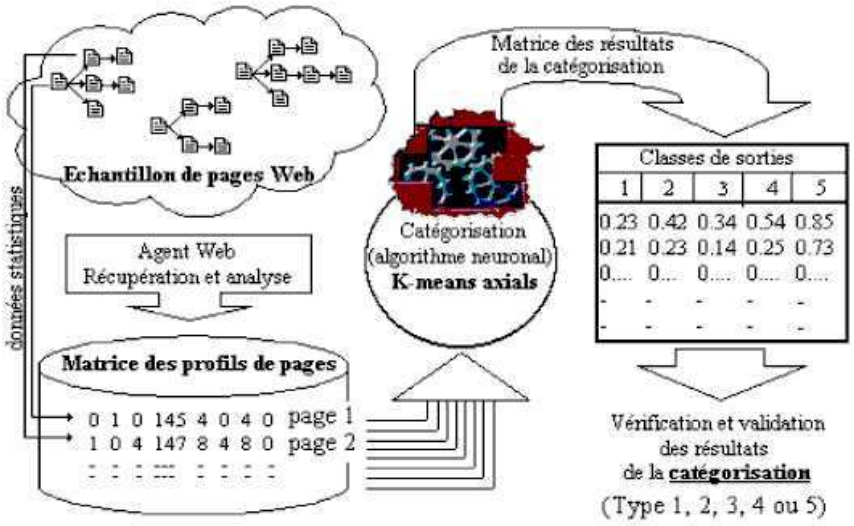

Figure 5.1 : Catégorisation automatique d'un échantillon de pages Web

8 Usage et exploitation des résultats

8.1 Comme module HyWebMap

La catégorisation automatique des pages Web au niveau du système HyWebMap permet à l'utilisateur de filtrer les pages Web référencées par les différents nœuds d'un réseau HyWebMap. Cela lui permettra de naviguer par catégorie donc par type de pages. Il est possible d'indiquer la ou les classe(s) à visualiser.

8.2 Au sein d'un moteur de recherche

En tant que filtre des pages HTML entre l'utilisateur et les moteurs de recherches sur le Web : la figure 5.4. présente l'interface d'un prototype de méta-chercheur, l'utilisateur choisit un ou plusieurs mots-clés, ainsi que la catégorie de pages qu'il souhaite visiter.

Les agents Web sollicitent le moteur de recherche Altavista , et les résultats obtenus suite à la requête sont analysés et un profil de chaque document est construit.. Le module de catégorisation utilise une matrice intermédiaire pour insérer le nouveau profil dans la catégorisation de base validée. De cette manière, il est plus facile de dresser une typologie d'un document Web sans le visualiser.

Dans l'exemple de la figure 5.2, l'utilisateur ne souhaite visiter que les pages de catégorie texte, il sélectionne alors la catégorie Texte, où se trouve les mots-clés "guide Internet". Après récupération, analyse, débalisage et comparaison, les agents affichent les résultats avec un degré d'échelle : T* (une page contient du texte), T** (une page moyennement textuelle), T***(une page faiblement textuelle). En tant que module de filtrage pour un moteur de recherche : cette classification peut être exploitée dans le cadre d'un moteur de recherche pour éviter l'indexation de pages qui contiennent peu de texte et qui sont donc peu significatives.

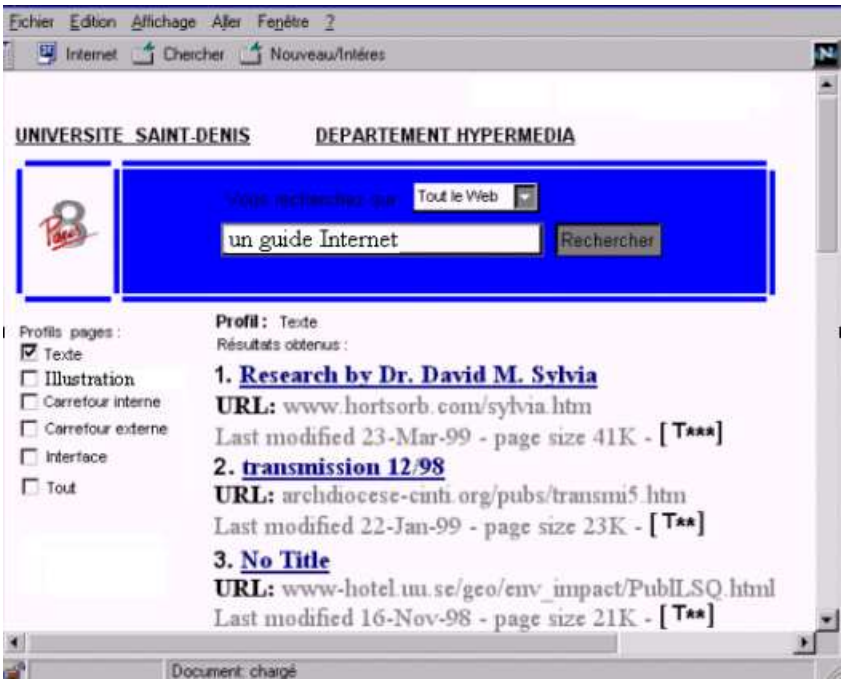

Figure 5.23 : Utilisation du module de catégorisation au sein d'un moteur de recherche

9 Conclusion

Les moteurs de recherches, les agents intelligents, les portails, les solutions d'E-learning, etc…au-delà de leur spécificités fonctionnelles, témoignent de la présence d'informations conçues et élaborées pour le Web. Ce sont désormais des informations ciblées, que ces dispositifs grâce à leurs puissants mécanismes de filtrage transmettent aux internautes. Ceux-ci ne sont plus confrontés à la recherche de «LA» bonne information mais à la nécessité de traiter les multiples bonnes réponses qui leur sont envoyées. Compiler les URL, les colationner, les ré-organiser fréquemment, les compléter le cas échéant représente un mode de fonctionnement (presque) banalisé chez les internautes les plus concernés par l'exploration du Web. La classification automatique de pages HTML que nous avons présentée, s'insère dans cette logique d'assistance à la recherche et la navigation dans l'espace virtuel du Web. Cette technique algorithmique est opérationnelle sous la forme d'un module utilisable au sein d'un environnement comme le système HyWebMap ou comme un module spécifique de moteur de recherche pour le filtrage de document à destination d'environnement d'indexation ou comme indicateur informatif sur la typologie de documents Web. Il reste à étudier l'amélioration de l'ergonomie des interfaces d'affichage : les techniques 3D peuvent être utiles pour présenter les résultats et s'y mouvoir.

Références

[Balpe et al. 96] Balpe J.P., Lelu A., Saleh I., Papy F., "Techniques avancées pour l'hypertexte", Editions Hermès, 1996.

[Bélisle et al. 99] Bélisle C., Zeiliger R., Cerratto T., "S'orienter sur le Web en construisant des cartes interactives : le navigateur NESTOR", in Hypertextes hypermedias et Internet H2PTM'99 Balpe, Natkin, Lelu, Saleh, Hermes Science Publications, Paris, pp. 101-117, 1999.

[Borzic 98] Borzic B., "Un modèle de gestionnaire itératif de flux informationnel sur Internet", Thèse de doctorat, Information Scientifique et Technique, CNAM, Paris, mars 1998.

[Catledge 95] Catledge L.D., Pikow J.E., "Characterizing browsing startegies in the World Wide Web", Proc. of the 3th International Conference on the World Wide Web, Darmstadt, Germany, 1995.

[Cutting 92] Cutting D.R. et al., "Scatter/Gather : A cluster-based approch to browsing large document collections", Actes de la 15th Conférence ACM/SIGIR Research and developpement in information retrievel, Danemark, 1992.

[Lelu & al 99] Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., "Projet NeuroWeb : un moteur de recherche multilingue et cartographique", 5e conf. Int. H2PTM'99, Paris, France, septembre 1999.

[Lelu 93] Lelu A., "Modèles neuronaux pour l'analyse de données documentaires et textuelles", Doctorat de l'Université Paris VI, mars 1993.

[Maarek & Schaul 96] Maarek Y.S., Schaul I., "Automatically organizing bookmarks per content", 5th International World Wide Web Conference, Paris, 1996.

[Rijsbergen 79] Rijsbergen C.J., "Information retrievel", Butterwords, London, 1979.