Classification de contextes de liens hypertextes

Moustafa Al-Hajj * Gilles Verley *

Université François-Rabelais de Tours
Laboratoire d'Informatique (EA 2101),
64, Avenue Jean Portalis,
37200 TOURS – France
* prenom.nom@univ-tours.fr

RÉSUMÉ. Les auteurs qui publient sur le Web des connaissances sous la forme de documents électroniques lisibles sur un écran utilisent de plus en plus la technologie des liens hypertextes pour améliorer l'ergonomie de leur sites et pour les enrichir par des informations provenant d'autres sites Web [1]. Nous nous intéressons à la sémantique des liens hypertextes, en termes d'extraction et d'exploitation, dans le but de faciliter le partage des connaissances sur le Web. Dans cet article, nous nous concentrons sur l'élaboration d'outils d'aide à l'analyse de la sémantique des liens hypertextes, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens.

ABSTRACT. The authors who publish knowledge on the Web in readable electronic documents on a screen use more and more the technology of the *hypertext links* to improve the ergonomics of their sites also to enrich it by information coming from other Web sites. We are interested in semantics of the *hypertext links*, in terms of extraction and exploitation, with the aim of facilitating the search of knowledge on the Web. In this article, we concentrate on the development of tools for the assistance using the analysis of hypertext links. We propose an automated tool for the literary pattern recognition of various contexts.

MOTS-CLÉS : Classification de contextes des liens hypertextes, Treillis de Galois, arbre de décision, K-means.

KEYWORDS: Classification of context of hypertext links, Galois lattice, Tree decision, K-means.

Introduction

Ce travail de classification de contextes de liens hypertextes fait partie d'une étude plus générale qui consiste à faire l'analyse sémantique des liens hypertextes. Pour faire l'analyse sémantique des liens, nous avons construit notre propre corpus et avec, comme domaine, les biographies d'hommes célèbres. Dans l'analyse de ces liens on a besoin de caractériser les formes littéraires des contextes appelants des liens et des contextes appelés par les liens.

On nomme « contexte appelant d'un lien » l'ensemble minimal de textes, caractères et objets, autour du lien et qui constituent une seule idée, concept ou sujet.

De même, on nomme « contexte appelé par un lien » l'ensemble minimal de textes, caractères et objets de la page ciblée par le lien et qui constituent un sujet en rapport avec le « contexte appelant du lien ».

Dans une première partie on s'intéressera aux travaux qui permettent de classifier des pages Web par leurs profils syntaxiques, et dans une deuxième on s'intéressera plus particulièrement aux profils syntaxiques des contextes appelants des liens et des contextes appelés par des liens, dans le reste on montrera qu'il est possible de classer ces contextes par de méthodes de reconnaissances de formes.

1. Classification de pages Web selon leurs formes littéraires

1.1. Généralités

Pour indexer les documents web, trois types d'information peuvent être utilisées :

- Le contenu lui-même des pages web : c'est-à-dire l'ensemble du code source de la page, le texte, les balises, les liens hypertextes, les liens vers les images ou d'autres ressources multimédias, la taille des fichiers, etc.

- Le graphe créé par les liens hypertextes reliant les pages les unes aux autres.

- Les données provenant de l'usage comme les fichiers de log, les "cookies", etc.

Cette classification est proposée par la communauté du « web mining » [2].

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 ces types de données.

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.

Les documents sur le Web sont hétérogènes (sites commerciaux, pages personnelles, livres, articles, annuaires), ne possèdent aucune véritable structure. Bélisle C., Zeiliger R. et Cerratto T. [3] 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, a visée commerciale provenant des entreprises.

- Information médiatique, provenant des organismes des presses.

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

Cette différence est floue car certains sites proposent plusieurs types d'informations. Les profils syntaxiques des sites peuvent varier d'un site à un autre par rapport aux objectifs de chaque site.

Papy F.et Bounai N. [4] proposent une approche fondée sur la classification de pages. Ils prennent en considération les balisages utilisés dans les pages Web pour élaborer des 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.

Ils distinguent 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.

Pour établir une catégorisation de classification automatique des pages, ils se sont appuyés sur les travaux d'Alain Lelu ([5], [6]) en utilisant l'algorithme de *K-means axiales*.

Une fois la méthode de *K-means* appliquée sur leur corpus, cinq types de pages ont été 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 : Le contenu de la page est un texte.

- Page informative avec texte illustré : Le contenu de la page est une illustration visuelle, ce peut être des images, des figures, des boutons, etc.

- Page carrefour interne au site : le contenu de la page est un ensemble de liens internes au site.

- Page carrefour externe au site : le contenu de la page est un ensemble des liens externes au site.

- Page interface à la saisie : le contenu de la page est un ensemble de champs de saisie.

2. Notre contribution

Nous nous sommes inspirés de ces travaux pour construire nos classes. Nous en avons retenu certaines et en avons rajouté d'autres spécifiques au domaine des biographies d'hommes célèbres.

Après une observation des formes littéraires des différents contextes de notre corpus, nous avons opté pour les classes suivantes :

- Classe sommaire : Le contenu du contexte est un résumé qui comporte les titres des parties des sites, c'est la même chose que la page carrefour interne. On les reconnaîtra principalement grâce à l'adjacence des liens.

- Classe illustration graphique : Le contenu du contexte est une illustration graphique par une image. On les reconnaîtra principalement grâce à la présence d'images de taille importante dans le contexte.

- Classe récit : Le contenu du contextes est en majorité du texte, on les reconnaîtra principalement grâce à la présence de texte en grand quantité dans le contexte.

- Classe citation : Le contenu du contexte est un texte qui fait référence directe à une oeuvre dans sa totalité ou en partie. On les reconnaîtra principalement grâce à la présence de texte en quantité moyenne et sans liens hypertextes.

- Classe liste : Le contenu du contexte est une suite d'articles inscrits les uns à la suite des autres. On les reconnaîtra principalement grâce à la présence des puces ou numéros aux débuts des articles.

3. Paramètres

En partant des caractéristiques citées auparavant, il est possible d'établir le profil d'un contexte en constituant un vecteur d'informations. Les données les plus significatives obtenues à partir de notre échantillon de contextes sont : *nbHref* : nombre de liens ; *nbImg* : nombre d'images ; *TGimg* : taille de la plus grande image ; *SMoyImg* : surface moyenne des images ; *nbMot* : nombre de mots hors balise ; *nbLEH* : nombre de lignes entre balises « a href » ; *nbLigne* : nombre de lignes hors balise ; *nbBListe* : nombre de balises qui définissent des listes et/ou listes avec puces et/ou les énumérations ; *nbBPg* : nombre des balises qui définissent les paragraphes ; *nbBSLigne* : nombre de balises de saut de lignes ; *cit* : prend 1 si des mots tels que « citation » figurent en balise méta name et 0 sinon ; *def* : prend 1 si des mots tels que « définition » figurent en balise méta name et 0 sinon ; *desc* : prend 1 si des mots tels que « description » figurent en balise 'méta name' et 0 sinon ; *sommaire* : prend 1 si des mots tels que « sommaire, résumé » figurent en balise méta name et 0 sinon.

L'agent Web recueille les indicateurs quantitatifs, et les stocke sous forme de matrice (cf. tableau 1), chaque ligne correspond à un contexte et chaque colonne correspond à l'un des paramètres.

nbHref	*nbImg*	*TGimg*	*SMoyImg*	*nbMot*	*nbLEH*	*nbLigne*	*nbBListe*	*nbBPg*	*nbBSLigne*	*cit*	*def*	*Desc*	*Sommaire*
10	1	4628	4628	2770	23	239	40	47	0	0	0	0	0

Tableau 1. *Une ligne de la matrice documents / attributs*

4. Découpage de la base de données

Le corpus de document sur lequel on travaille est composé de biographies d'hommes célèbres. Pour la phase d'expérimentation, nous avons annoté manuellement 1031 contextes appelants ou appelés de notre corpus par leurs formes littéraires. Deux tiers tirés au hasard de cet ensemble seront utilisés comme données d'apprentissage et le un tiers restant sera utilisé comme données de test. Le tableau 2 récapitule les effectifs des contextes dans la base d'apprentissage et de celle de test.

	Citation	Illustration	Liste	Sommaire	Récit
Base d'apprentissage	305	9	44	101	231
Base de test	151	7	29	47	107
% de classes dans les deux bases	44,3 %	1,6 %	7,1 %	14,4 %	32,6 %

Tableau 2. *Effectifs des formes littéraires dans les deux bases*

La classe citation est fortement représentée du fait du domaine d'application de biographies d'hommes célèbres.

5. Classification avec les treillis de Galois

L'analyse formelle des concepts (AFC) [7] offre un cadre théorique aux applications nombreuses et reconnues. Elle permet de représenter des données définies par une relation binaire entre deux ensembles, représentation encore appelée treillis de Galois [8].

5.1. Discrétisation

Notons que cette phase est un point très important pour l'efficacité du treillis de Galois. Nous avons utilisé la méthode de discrétisation utilisée par Quinlan dans le C4.5 [9]. Cette méthode discrétisation a permis de générer un ensemble de 54 intervalles pour tous les paramètres. Le tableau binaire utilisé comme entrée pour la construction du treillis de Galois est défini comme suit :

Chaque ligne du tableau représente un contexte de la base d'apprentissage. Chaque contexte de la base d'apprentissage est représenté dans ce tableau par 59 attributs binaires dont 54 sont obtenus par échantillonnage de chaque valeur de paramètre dans les intervalles obtenus par la méthode de discrétisation pour ce paramètre.

Les contextes de la base de test sont représentés de la même manière sauf les 5 derniers attributs de classes qui sont évidement tous des zéros. La classification des contextes de la base de test revient à leurs inférer des attributs de classe.

5.2. Résultats

L'application de la méthode « Validation Globale », que nous détaillons dans [10], sur les 341 contextes de la base de test a permis de classer 171 contextes dont 163 sont correctement classés. De même, l'application de la méthode « Validation Locale », aussi détaillée dans [10], sur l'ensemble de test a permis de classer 141 contextes dont 109 sont correctement classés. Le tableau 3 récapitule les résultats obtenus.

			Citation	Illustration	Liste	Sommaire	Récit
	Effectifs	341	151	7	29	47	107
VG	Classés	171	103	1	7	21	39
	Correctement classés	163	99	0	6	20	38
VL	Classés	141	60	0	17	17	47
	Correctement classés	109	52	0	10	13	34

Tableau 3. *Résultats obtenus avec les treillis de Galois*

6. Conclusion

Ce travail se situe dans un projet plus vaste d'analyse de la sémantique de liens hypertextes [1]. Nous avons présenté une expérience d'extraction, par les treillis de Galois, de la partie de la sémantique qui correspond aux formes littéraires des contextes

appelants des liens et des contextes appelés par des liens. Les méthodes de classification se basant sur les treillis de Galois peuvent inférer des attributs du contexte à classer, ce qui nous a permis d'adapter le treillis de Galois au problème d'apprentissage supervisé en intégrant les classes dans les attributs des contextes à classer. Les résultats sans être excellente sont encourageants. D'autres expériences de classification avec d'autres outils de classification supervisée sont en cours et une généralisation sur d'autres corpus est envisagée.

7. Références

[1] Gilles Verley, J.J. Rousselle, "*An evolved link-specification language for creating and sharing documents on the web*", CRIS 2000 Current Research Information Systems, 25-27 Mai 2000, Helsinky.

[2] Kosala R., Blockeel H., "*Web Mining Research: A Survey*", SIGKDD Explorations, vol. 2 (1) 2000, p. 1-15.

[3] 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.

[4] Papy F., Bounai N., « *Navigation et recherche par catégorisation floue des pages HTML*,», Actes des JET'2003, 2003.

[5] 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.

[6] Balpe J.P., Lelu A., Saleh I., Papy F., « *Techniques avancées pour l'hypertexte* », Editions Hermès, 1996.

[7] Ganter B. & Wille R. (1999). « *Formal concept analysis, Mathématical fondations* ». Springer Verlag, Berlin.

[8] Mephu Nguifo et Njiwoua, 2005, « *Treillis de concepts et classification supervisée : un état de l'art* ». CRIL rapport de recherche.

[9] QUINLAN J. R. , "*C4.5 : Programs for Machine Learning*", Morgan Kaufmann, 1993.

[10] M. Al-hajj, K. Bertet, J.Gay et J.-M. Ogier. « *Aide à la reconnaissance d'objets détériorés avec un treillis de Galois* » In Atelier Treillis, AFIA 2003, Laval, France, Juin 2003.