Identification et structuration hiérarchique des titres dans les documents HTML

Structuration hiérarchique des titres

Thierry Waszak*, — Claude de Loupy*,*** — Patrice Bellot****

* *Syllabs*
2, rue de Fontarabie, F-75020 Paris
{waszak, loupy}@syllabs.com

** *LIA / Université d'Avignon*
339, chemin des Mainajaries, Agroparc BP 1228, F-84911 Avignon

*** *MoDyCo / Université de Paris 10*
UMR 7114, 200, avenue de la République, F-92001 Nanterre

RÉSUMÉ. *Dans cet article, nous présentons une méthode pour automatiquement identifier et structurer hiérarchiquement les titres dans les documents HTML. Bien que la syntaxe HTML propose des balises de titres, l'usage de ces balises dans beaucoup de documents n'est pas correct ou ces balises ne sont pas utilisées. Notre méthode se base sur les propriétés visuelles, telles la taille ou la couleur de la police, obtenues grâce aux feuilles de style (CSS). L'hypothèse est que plus un élément est visible, plus son niveau dans la hiérarchie des titres est élevé. Nous avons extrait du Web un corpus de CSS que nous utilisons dans l'apprentissage d'un modèle de Markov caché. Les premiers résultats donnent une F-Mesure de 0,70 pour la structuration des titres et de 0,86 pour l'identification.*

ABSTRACT. *In this paper, we describe a method to automatically identify titles within Web pages. Although HTML syntax provides specific tags for titles, they are not always correctly used, and sometimes they do not even appear. We use visual clues like font size or colour provided by Cascading Style Sheets in order to retrieve the title hierarchy. The assumption is that the level of an element in the title hierarchy increases with its visibility. We automatically built a CSS corpus by crawling the Web and used it to learn a Hidden Markov Model which identifies titles and their hierarchy. Primary results give a F-Measure of 0.70 for titles structuring and 0.86 for titles identification.*

MOTS-CLÉS : *Hiérarchie des titres, Modèle de Markov Caché, Balises de visibilité, document HTML, Corpus Web.*

KEYWORDS: *Titles structuring, Hidden Markov Model, Visibility tags, HTML document, Web corpus.*

Thierry Waszak, Claude de Loupy et Patrice Bellot

1. Introduction[1]

Avec la grande quantité d'information en constante augmentation disponible sur le Web, la nécessité de développer des méthodes permettant l'extraction de l'information pertinente à l'intérieur même des documents HTML est évidente. Il s'agit de rechercher le contenu d'un paragraphe spécifique en faisant abstraction du bruit que représente certaine partie du document (publicité, pop-up…). Cet article traite du problème de l'identification et de la structuration hiérarchique des titres dans les pages Web. En effet, les titres sont de bons indicateurs des sujets développés dans les paragraphes qu'ils introduisent. Ils peuvent donc aider à repérer l'information pertinente. Non seulement ils organisent sémantiquement le texte en en permettant une meilleure compréhension, mais ils sont aussi révélateurs de l'organisation spatiale du texte (Ho-Dac *et al.*, 2004 ; Jacques *et al.*, 2006). Ainsi, les titres peuvent être utilisés dans l'optique d'améliorer les systèmes de résumé automatique (en se basant sur les titres – résumé thématique) (Edmundson, 1969 ; Marcu, 1997). En recherche d'information, Hu *et al.* prend en considération ces propriétés des titres en donnant un poids plus élevés aux mots faisant partie du titre principal des documents ; cela a pour effet d'améliorer les performances de leur système (Hu *et al.*, 2005).

Il n'existe pas, à notre connaissance, de précédents travaux portant sur l'identification et la structuration hiérarchique des titres dans les pages Web : Hu *et al.* recherche uniquement le titre principal des pages. D'autres études se sont intéressées à la segmentation des documents HTML : il s'agit de structurer les pages Web en identifiant les différentes zones des pages et en leurs affectant un label. Ainsi la structuration des titres peut s'apparenter à une segmentation de document. En effet, les principaux titres d'une page, peuvent être utilisés pour identifier les différentes zones de cette page. En segmentation de document HTML, Song *et al.* ainsi que Xue *et al.* utilisent la représentation DOM des documents HTML afin d'en extraire les différentes propriétés qui permettront une classification automatique des zones des pages Web à l'aide de machines à vecteur support (SVN) et de réseaux de neurones (Song *et al.*, 2004 ; Xue *et al.*, 2007). Ces propriétés sont des balises HTML (celles définissant la taille de police, la couleur, *etc.*), la position dans l'arbre DOM représentant le document HTML et des propriétés linguistiques comme le nombre de mots. Afin d'avoir une méthode plus portable et d'être moins dépendant du format spécifique aux différents sites Web, Hattori *et al.* ne considère que le nombre d'apparition des balises dans l'arbre DOM ainsi que leurs profondeurs relatives (Hattori *et al.*, 2007) alors que Mukherjee *et al.* essaye de rassembler les différentes zones des pages Web en considérant leurs similarités dans l'arbre DOM (Mukherjee *et al.*, 2003).

1. Ce travail a été effectué dans le cadre du projet ANR-RNTL TextCoop (www.textcoop.org).

Notre méthode se base sur les feuilles de style (Cascading Style Sheets – CSS) généralement associées aux documents HTML. En effet, ce sont les CSS qui contiennent les informations sur la mise en page des documents HTML (taille de police, couleur, *etc.*). Or, ces propriétés visuelles sont de première importance dans l'identification des titres (Hu *et al.*, 2005 ; Song *et al.*, 2004 ; Xue *et al.*, 2007 ; Yang *et al.*, 2001). L'idée est de classer les différents styles (qui sont appliqués a un ou plusieurs éléments dans l'arbre DOM) en fonction de leurs relatives visibilités et de décider si un style est lié à un titre (avec un certain niveau hiérarchique – plus le niveau est élevé plus la visibilité est importante) en utilisant un modèle de Markov caché appris sur un corpus extrait du Web. Contrairement à Hu *et al.*, 2005 ; Mukherjee *et al.*, 2003 ; Song *et al.*, 2004 ; Xue *et al.*, 2007 ; Yang *et al.*, 2001, nous ne considérons pas les balises HTML elles-mêmes, mais plutôt les propriétés visuelles associées à une ou plusieurs balises (*i.e.* une suite de balise) HTML. Dans un modèle étendu, nous utilisons (comme Hattori *et al.*, 2007) les statistiques sur les balises du document HTML.

Remarquons enfin que la syntaxe HTML possède des balises spécifiques pour les titres *(<H1>, <H2>, etc.*). Ces balises permettent une identification ainsi que la structuration hiérarchique des titres. Le problème est que l'utilisation de ces balises n'est pas toujours correcte. En effet, beaucoup de pages Web utilisent ces balises dans un ordre incohérent : on trouvera par exemple *<H5>*, *<H2>*, *<H3>* au lieu de *<H1>*, *<H2>*, *<H3>*. Notre étude de corpus montre qu'au moins 25% des pages Web sont ainsi mal structurées. De plus, il existe des cas ou ces balises n'apparaissent même pas. Par conséquent, nous ne considérerons pas ces balises, mais seulement les balises (ou suite de balises) pouvant être associées à un style (CSS) et portant donc une certaine information visuelle.

Dans la section 2, nous présentons notre méthode pour la structuration hiérarchique de titres ; dans la section 3 nous expliquons comment nous avons automatiquement constitué un corpus extrait du Web. Finalement, notre approche est évaluée dans la section 4.

2. Méthode

La principale hypothèse est de considérer qu'un titre à un niveau *n+1* (*<Hi>*) est plus visible qu'un titre à un niveau *n* (*<Hi+1>*). Le titre (*<H1>*) est généralement écrit d'une manière plus « visible » (en gras, avec une taille de police plus grande, *etc.*) qu'un sous-titre (*<H2>*). Notre système se base donc sur ces propriétés visuelles des documents (Hu *et al.*, 2005 ; Song *et al.*, 2004 ; Xue *et al.*, 2007 ; Yang *et al.*, 2001). Ces propriétés sont retrouvées en considérant les feuilles de style (CSS) dont l'utilisation est maintenant largement répandue sur le Web. Ces feuilles de style servent en effet à définir le formatage du texte (l'aspect visuel) alors que le document HTML en lui-même ne doit contenir que le texte. Afin de pouvoir utiliser les informations visuelles présentes dans les CSS, la première étape de notre

approche consiste à analyser ces CSS associées aux documents HTML et de créer ce qu'on défini comme des « *pseudo documents* ». Ces pseudo documents se trouvent au cœur de notre approche et sont notamment utilisés lors de la phase d'apprentissage et d'analyse.

2.1. *Pseudo document*

Les CSS sont composées de balises HTML T_i (ou de séquences de balises – le texte dans le document HTML délimité par ces balises sera formaté avec un certain style défini dans la CSS) associées à la définition d'un style défini à l'aide de propriétés p_i associées à des valeurs v_i. Ainsi une ligne dans une CSS aura la syntaxe suivante : $T_i :\{(p_{ij} :v_{ij} ;)\}^+$. Par exemple : *DIV :{font-color :red ;font-weight :bold ;}* signifie que le texte délimité par une balise DIV sera en rouge et en gras. Il est à noter que pour cet exemple, la balise <*DIV*> dans le document HTML pourrait être remplacée par la suite de balise suivante : <*FONT COLOR="red"*><*B*>. Nous appellerons par la suite ce type de balises des *balises de visibilité*. On a donc une balise *clé* T_i qui peut être associée avec une ou plusieurs balises de visibilité V_{ij}. La ligne CSS peut donc être convertie en une *association* : (T_i,V_{ij}). Le pseudo document est alors simplement constitué de ces associations. Il est à noter que tous les couples (p_{ij},v_{ij}) ne sont pas considérés. En effet, la fonction de transformation [1] utilisée, permet une réduction du nombre de propriétés de deux points de vue.

$$Tr_v(p_{ij}, v_{ij}) = V_{ij} \quad [1]$$

Tout d'abord, seuls les p_j pouvant être mis en relation avec des balises de visibilité sont conservés. Cette réduction est faite grâce à l'écriture de règles. Par exemple, des règles définissent qu'une indentation de texte sera transformée en * *, une marge ou un bord en <*BR/*> et toutes les propriétés sur les polices en différentes balises de visibilités : <*B*>, <*FONT SIZE="couleur"*>, *etc*. Deuxièmement, les tailles et couleurs sont normalisées. En effet, ces propriétés peuvent être numériques. Nous arrondissons donc ces valeurs afin de ne pas avoir plus de 19 tailles de police différentes et plus de 256 couleurs.

Remarquons que dans le cas où un document HTML n'est associé à aucune CSS, un pseudo document peut tout de même être généré. Dans ce cas, les balises de visibilités présentes dans le document HTML sont remplacées par des balises clés arbitraires. Par exemple, la suite de balises suivante : <*B*><*FONT SIZE="+2"*> sera remplacée par la balise clé <*T1*>.

Le tableau 1 est une illustration d'un pseudo document. Les niveaux de titre sont ce que nous cherchons et ce qui doit être étiqueté pour un apprentissage. On peut remarquer que le niveau de titre augmente alors que les <*Hi*> diminuent.

Puisque notre modèle de structuration hiérarchique des titres se veut statistique, un corpus annoté de pseudo documents est nécessaire. Ce corpus a été créé

automatiquement en récupérant des documents HTML ainsi que leur CSS en parcourant le Web. Ce corpus, le parcours du Web et l'étiquetage automatique des niveaux de titre sont présentés dans la section 3.

	Balise clé (T_i)	Balise de visibilité (V_{ij})	Niveau de titre
avec CSS	<P>	<B>	0
	<P>	<I>	0
	<H3>	<FONT SIZE=+2>	1
	<H2>	<FONT SIZE=+2><I>	1
	<H1>	<B><FONT SIZE=+3>	2
sans CSS	<T1>	<B>	0
	<T2>	<I>	0
	<T3>	<B><FONT SIZE=+2><I>	1
	<T4>	<B><FONT SIZE=+2>	2

Tableau 1. *Représentation de pseudo documents*

2.2. *Structuration hiérarchique des titres*

Afin de structurer les titres, il faut tout d'abord structurer le pseudo document. En effet, il s'agit d'ordonner le pseudo document de façon à avoir les balises les plus visibles en fin de document. Ainsi la dernière association du pseudo document représentera le titre de plus haut niveau. Il faut ensuite un modèle capable de reconnaitre un changement de niveau afin de savoir quand nous passons d'un titre de niveau *n* à un titre de niveau *n+1*. Pour cela nous utilisons un modèle de Markov caché (Hidden Markov Model – HMM) (Rabiner, 1990). Dans les prochaines sous-sections, nous introduisons le *score de visibilité* qui est utilisé pour ordonner les balises de visibilité dans le pseudo document puis le modèle HMM.

2.2.1. *Score de visibilité*

Cette fonction doit être capable de classer efficacement les balises en fonction de leur visibilité. On définit $S_i = V_{i1}\ldots V_{ij}$ (S_i est une suite de balises de visibilité). Soit la fonction S_v qui associe un score de visibilité à une suite de balises. S_v est défini comme la somme des probabilités qu'une balise de visibilité corresponde à un changement de niveau :

$$S_v(S_i) = \sum_{b_j \in S_i} P\left(changement | b_j\right) \quad [2]$$

où $P(changement|b_j)$ représente la probabilité que b_j (b_j est une balise de visibilité) corresponde à un changement de niveau. Ce score est calculé pour chaque b_j du corpus. Pour la phase d'analyse, on ne tient pas compte des b_j qui n'ont jamais été vu dans le corpus d'apprentissage.

2.2.2. *Modèle de Markov caché*

Les modèles de Markov cachés modélisent des séquences de données. Les HMM peuvent être vus comme une généralisation stochastique d'automates à état finis, où à la fois les transitions entre états et la génération des symboles sont gouvernés par des distributions de probabilité (Stolcke *et al.*, 1992). Un HMM peut être caractérisé comme suit : $HMM = \{W,V,A,B,\pi\}$ où W sont les différents états du modèle, V l'alphabet, A les probabilités de transition entre états, B les probabilités d'observation des symboles pour chaque état et π les distributions de probabilité initiales.

Pour notre problème de structuration des titres, nous voulons maximiser $P(L|D)$ où L représente la séquence des niveaux de titre dans le document D. (D fait référence au document HTML **et** à ses CSS associées.) D'après le théorème de Bayes, cela revient à maximiser :

$$P(L|D) = P(D|L).P(L) \quad [3]$$

Dans ce modèle, on décide de ne prendre en considération que la mise en forme (*i.e.* les suites de balises de visibilité S_j). On considère donc C (représentant cette mise en forme) comme une approximation de D. Ainsi [3] devient [4] :

$$P(L|D) = P(C|L).P(L) \quad [4]$$

Dans le modèle HMM, W représente les différents niveaux de titre : $W = \{0,\ldots,N\}$ où N est le plus haut niveau rencontré dans le corpus d'apprentissage ; $V = \{b_j\}_{(1\leq j\leq M)}$ avec M le nombre maximum de balises de visibilité b_j rencontrées dans le corpus ; les probabilités A et B sont obtenues grâce au corpus d'apprentissage. Remarquons qu'on doit avoir $\pi_0 = 1$ et $\pi_i = 0_{(1\leq i\leq N)}$. En effet, tous les pseudo documents, une fois ordonnés par leur score de visibilité, commencent par de balises de visibilité ne représentant pas un titre. La figure 1 illustre le HMM associé au pseudo document (avec CSS) présenté dans la table 1.

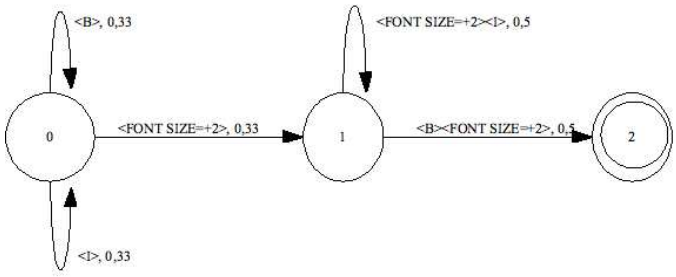

Figure 1. *Automate à états finis représentant un HMM*

Etant donnée une séquence d'observation de balises de visibilité, en utilisant l'algorithme de Viterbi, le modèle HMM nous donne la séquence optimale de niveaux de titre associée à cette observation.

2.3. *Modèle étendu*

Jusqu'ici uniquement les propriétés visuelles de mise en forme des documents HTML ont été utilisées. L'hypothèse du modèle étendu est que l'ajout d'autres propriétés issues non plus des CSS mais du document HTML lui-même peuvent améliorer les résultats. Comme évoqué dans (Hattori *et al.*, 2007), utiliser la grammaire des balises HTML peut s'avérer restrictif à un domaine (au corpus d'apprentissage). Nous décidons donc d'utiliser comme propriétés la fréquence d'apparition d'une balise dans le document HTML, ou encore la profondeur relative dans l'arbre DOM de la balise. L'hypothèse étant que les différents niveaux de titre sont moins fréquents que les paragraphes et qu'un titre ne peut pas avoir une profondeur plus importante qu'un sous-titre. Dans les sous-sections suivantes, nous présentons le modèle HMM étendu, puis ce même modèle auquel on applique certaines règles linguistiques.

2.3.1. *Description du modèle étendu*

D représentant le document HTML et ses CSS associées, on fait maintenant l'hypothèse que $D = H \cap C$ où *H* représente les propriétés extraites du document HTML et *C* celles extraites des CSS. En effet, le texte affiché dans un navigateur Web provient de la combinaison entre le texte lui-même et la mise en forme. On fait l'hypothèse que *H* et *C* sont indépendants et on a donc l'approximation suivante : $P(D|L) = P(C|L).P(H|L)$. L'équation [3] devient alors :

$$P(L|D) = P(C|L).P(H|L).P(L) \qquad [5]$$

A l'aide des propriétés provenant du document HTML, on modifie également le score de visibilité S_v en pondérant la première formule [2] avec la fréquence d'apparition f_i des balises clés (T_i) et leurs profondeurs relatives d_i dans l'arbre DOM représentant le document HTML. En effet, pour une balise T_i, le niveau de titre diminue lorsque la profondeur augmente et lorsque la fréquence d'apparition augmente. On obtient le score de visibilité suivant :

$$S_v(S_i) = \frac{1}{f_i.d_i}.\sum_{b_j \in S_i} P\left(changement|b_j\right) \qquad [6]$$

2.3.2. *Modèle étendu avec règles linguistiques*

A partir de l'observation du corpus, il apparaît que certaines règles linguistiques simples peuvent être ajoutées au précédent modèle afin d'améliorer le score de visibilité. Par exemple, si une balise clé apparaît moins de *x* fois (on fixe

empiriquement $x = 10$) dans le document HTML, elle a de forte chance de représenter un titre (dans la formule [7], on pose $\alpha = 1$ si $f_i \leq x$ et $\alpha = 0{,}9$ sinon). De plus, si la profondeur d'une balise dans l'arbre DOM est plus grande que celle de la balise précédente, cette précédente balise a de fortes chances d'être un titre de plus haut niveau ($\beta = 1$ sinon $\beta = 0{,}9$). Enfin, si une balise se trouve à la racine de l'arbre DOM ou à la profondeur maximale, elle ne représentera pas un titre ($\delta = 0$ sinon $\delta = 1$). On obtient pour le score de visibilité la formule suivante :

$$S_v(S_i) = \frac{\delta\alpha\beta}{f_i.d_i} . \sum_{b_j \in S_i} P\left(changement | b_j\right) \quad [7]$$

2.4. *Du pseudo document au document HTML restructuré*

A ce stade, seule la structuration des pseudo documents est faite. La dernière étape consiste à réaffecter aux balises clés les balises de titre identifiées dans les pseudo documents. Il suffit simplement de remplacer dans les documents HTML les balises clés en titres ou paragraphes. Le plus haut niveau de titre dans le pseudo document deviendra une balise *<H1>*, puis *<H2>*, *etc.* On peut encore rajouter ici quelques règles linguistiques pour corriger certaines erreurs que le modèle peut commettre. Ainsi, on remarquera que les titres sont souvent composés de peu de mots. On décide donc d'invalider la reconnaissance d'un titre de plus de *y* caractères (on fixe empiriquement $y = 80$). On fait aussi la correction dans le cas où un titre est identifié en fin de section : en effet, un titre est forcément suivit d'un paragraphe. Remarquons que ces règles portent uniquement sur les propriétés des documents HTML (*H*).

3. Corpus

Les méthodes d'apprentissage supervisées nécessitent un corpus d'apprentissage. Ce corpus doit être suffisamment important pour que l'apprentissage soit efficace ; ce qui demande du temps pour l'annotation. Afin de nous affranchir de ce problème, nous utilisons les propriétés de la syntaxe HTML qui prévoient des balises pour l'identification (et la structuration hiérarchique) des titres. Toute la difficulté est de parcourir le Web afin d'en récupérer seulement les documents *bien formés*.

3.1. *Documents bien formés*

Le principal problème des documents HTML est que l'utilisation des balises de titre, lorsqu'elles sont présentes, peut être faite dans un ordre incohérent. (Par exemple, *<H5>,<H2>,<H3>* au lieu de *<H1>,<H2>,<H3>*.) Ainsi, on considérera un document HTML comme correct, seulement si sa hiérarchie de titres est correctement suivit *i.e.* qu'il existe *<Hi>* dans le document HTML tel qu'il

existe j et n avec $1 \leq j,n \leq 6$ et $1 \leq j+n \leq 6$ tels que pour chaque i : $j \leq i \leq j+n$. (Le document HTML avec les balises <*H5*>,<*H2*>,<*H3*> ne sera donc pas considéré comme bien formé.) Cela ne garantit pas pour autant que l'ordre des titres dans les documents HTML restants soit correct. En effet, cette restriction ne nous assure pas que <*H2*> ne soit pas utilisée à la place de <*H1*> comme balise de titre de plus haut niveau. Nous décidons donc de ne conserver que les documents XHTML. En effet, il est raisonnable de faire l'hypothèse que l'ordre des balises de titre est respecté dans ces documents ; la syntaxe XHTML étant plus stricte que la syntaxe HTML, on peut donc penser qu'une attention particulière est donnée à la rédaction de ces documents XHTML.

Ainsi, afin d'obtenir un corpus exploitable, il nous suffit de parcourir automatiquement le Web à la recherche de ces documents bien formés (documents XHTML et leurs CSS associées) et d'en extraire les pseudo documents.

3.2. *Extraction du Web des documents bien formés*

La première étape consiste à poser une requête à un moteur de recherche[2] afin de retrouver des URL à partir desquelles on récupère les documents. Ainsi, on constitue des corpus en relation avec des requêtes : il s'agit en quelque sorte de corpus thématiques. La deuxième étape est le parcours du Web lui-même. Comme illustré dans la figure 2, on suit un algorithme de parcours en largeur.

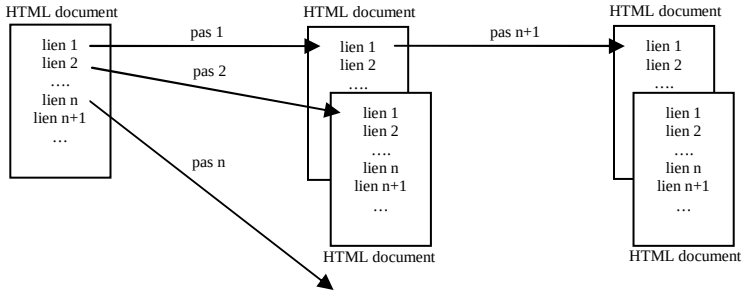

Figure 2. *Algorithme de parcours du Web*

En effet, comme le premier lien rapporté par un moteur de recherche doit être le lien le plus en relation avec la requête, ce lien doit d'abord être exploré. On explore ensuite le deuxième lien rapporté, *etc.* Ensuite, les liens des liens de la première page sont explorés et ainsi de suite. (On se limite au $n = 50$ premiers liens de chaque page. De plus, un maximum de $m = 7$ pages avec la même URL de base sont rapportées par soucis de couverture.) Cette stratégie de parcours du Web permet la

2. Nous avons utilisé Google (http://www.google.fr/)

création d'un corpus thématique dont la provenance n'est pas restreinte à un site particulier et donc à une certaine mise en forme. Il s'agit d'avoir la meilleure couverture de la grammaire HTML et CSS possible.

3.3. *Statistiques de corpus*

Dans le tableau 2 sont présentées les statistiques de corpus obtenues à la suite du processus présenté dans la section précédente. Ces statistiques sont calculées pour 5 différentes requêtes. Ces requêtes sont R1 : « CSS » ; R2 : « Histoire de France » ; R3 : « Politique » ; R4 : « Rugby » et R5 : « Afghanistan ». On récupère 200 documents pour chaque requête.

	R1	R2	R3	R4	R5	Moyenne
nombre de pages explorées	408	555	504	511	668	529
pourcentage de pages bien formées (pbf)	**0,49**	**0,36**	**0,40**	**0,39**	**0,30**	**0,39**
nombre de CSS par pbf	2,0	1,7	2,2	2,0	2,3	2,0
nombre de niveaux de titre par pages	2,7	2,7	2,3	2,7	2,6	2,6
pourcentage de pbf avec un ordre de titre cohérent	**0,78**	**0,77**	**0,84**	**0,70**	**0,55**	0,73
pourcentage de pbf contenant H1	0,88	0,90	0,87	0,81	0,81	**0,85**
pourcentage de pbf contenant H2	0,85	0,82	0,80	0,83	0,84	0,83
pourcentage de pbf contenant H3	0,69	0,57	0,51	0,60	0,56	0,59
pourcentage de pbf contenant H4	0,22	0,25	0,06	0,33	0,26	0,22
pourcentage de pbf contenant H5	0,06	0,07	0,02	0,11	0,13	0,08
pourcentage de pbf contenant H6	0,03	0,06	0,01	0,05	0,07	0,04

Tableau 2. *Statistiques des corpus*

On peut voir que seulement 39% des pages Web sont bien formées (de la façon dont nous l'avons défini). En moyenne, il y 2 CSS associées à une page bien formée et 2,6 niveaux de titre différents. De plus, seulement 73% des pages bien formées utilisent une hiérarchie de titre cohérente (comme présenté en section 3.1). On remarque aussi que 85% des pages utilisent la balise de plus au niveau <*H1*>. Cela sous entend que 15% des pages n'utilisent pas <*H1*> comme balise de titre de plus haut niveau ! Ces pages pourront tout de même nous servir dans l'apprentissage : on ne considère pas les balises elles-mêmes mais le style qui leur est associé ainsi que leur hiérarchie. En effet, dans le pseudo document, les balises de visibilité sont ordonnées selon leur hiérarchie et on pourra identifier le titre de plus haut niveau pour ce qu'il aurait du être : <*H1*>. En ordonnant les balises de visibilité suivant la hiérarchie suggérée par les balises de titre on peut ainsi obtenir un corpus d'apprentissage. Dans le souci d'avoir un corpus avec une meilleure couverture de la grammaire CSS et HTML, nous avons posé les mêmes requête et récupéré des corpus constitués de 1000 documents chacun. Les statistiques de corpus sont alors plus homogènes.

On peut aussi remarquer qu'il y a moins de pages bien formées parlant de l'histoire de France qu'il y en a propos des CSS. Cela pourrait s'expliquer par le fait que les pages parlant de CSS sont écrites par des informaticiens qui seront plus stricts dans l'utilisation de la syntaxe HTML.

Finalement, on voit qu'environ 27% de pages bien formées devraient être réordonnées. De plus, comme ces pages bien formées ne représentent que 39% des pages du Web, on voit bien qu'un outil de structuration hiérarchique des titres a tout son intérêt.

4. Evaluation

Dans cette section, nous présentons l'évaluation des différents modèles décrits précédemment. A cette fin, il nous faut tout d'abord définir les métriques que l'on va utiliser pour cette évaluation. Il faut noter que la première évaluation porte sur les pseudo documents ; on ne peut juger des performances de reconnaissance et de structuration des titres véritablement que sur la deuxième évaluation où les titres identifiés sont insérés dans les documents HTML.

4.1. *Métriques*

Nous effectuons une évaluation de nos modèles pour la tâche de structuration hiérarchique des titres mais aussi sur la tâche plus « simple » d'identification des titres. Dans ce cas, on cherche juste a savoir si un titre a été correctement reconnu. Pour cette tâche, on utilise les traditionnelles mesures de précision et de rappel. Dans le cas de la structuration hiérarchique des titres, on note l_i et l'_i respectivement le niveau de titre de référence et le niveau de titre supposé associés à une suite de balise de visibilité S_i. Précision et rappel sont définis de la manière suivante :

$$Précision = \sum_{l_i \neq 0} score(l_i) / nb(l'_i \neq 0) \quad [8]$$

$$Rappel = \sum_{l_i \neq 0} score(l_i) / nb(l_i \neq 0) \quad [9]$$

avec

$$score(l_i) = \sum_{l_j \neq l_i} ordre(l_j, l_i) / nb(l_j \neq l_i) \quad [10]$$

$$ordre(l_j, l_i) = \begin{cases} 1 \; si \; (l_j - l_i) > 0 \; et \; (l'_j - l'_i) > 0 \\ 1 \; si \; (l_j - l_i) < 0 \; et \; (l'_j - l'_i) < 0 \\ 0 \; sinon \end{cases} \quad [11]$$

où *ordre*(l_j,l_i) est égal à 1 si le niveau de titre l'_j est dans le même ordre par rapport à l'_i que l_j par rapport à l_i : il s'agit ici de donner plus de poids à un titre qui suit la hiérarchie de référence en vérifiant que quelque soit le niveau de titre l_j dans la hiérarchie de référence, le niveau de titre l'_j dans la hiérarchie supposée est dans le même ordre (plus grand ou plus petit) que le niveau du titre considéré (respectivement l_i et l'_i). *nb*($l_i \neq 0$) représente le nombre de titres dans la hiérarchie de référence et *nb*($l'_i \neq 0$) le nombre de titres identifiés par le modèle.

4.2. *Evaluation sur les pseudo documents*

A partir des cinq corpus extraits du Web (comme présenté en section 3.3), on produit cinq évaluations, une par corpus. A chaque fois, 80% du corpus est utilisé pour l'apprentissage et 20% pour l'évaluation. On procède à une évaluation croisée. Chaque corpus est constitué de 1000 documents. Les résultats obtenus, pour la structuration hiérarchique des titres puis pour la « simple » tâche d'identification des titres, sont présentés dans les tableaux 3 et 4. Dans ces tableaux, Modèle1 est le premier modèle « simple » que nous avons évoqué en section 2.2, Modèle2 est le modèle étendu présenté en section 2.3.1 et Modèle3 est le modèle étendu auquel on ajoute des règles linguistiques (section 2.3.2). (On note P pour la Précision et R pour le Rappel ; Ri sont les différentes requêtes *cf.* section 3.3.)

	R1		R2		R3		R4		R5		Moyenne	
	P	R	P	R	P	R	P	R	P	R	P	R
Modèle1	0,25	0,26	0,30	0,31	0,25	0,26	0,35	0,39	0,27	0,29	**0,28**	**0,32**
Modèle2	0,26	0,37	0,42	0,47	0,30	0,33	0,39	0,39	0,31	0,34	**0,34**	**0,38**
Modèle3	0,38	0,49	0,53	0,60	0,39	0,43	0,48	0,51	0,38	0,45	**0,43**	**0,50**

Tableau 3. *Structuration hiérarchique des titres dans les pseudo documents*

	R1		R2		R3		R4		R5		Moyenne	
	P	R	P	R	P	R	P	R	P	R	P	R
Modèle1	0,28	0,29	0,30	0,32	0,29	0,31	0,35	0,40	0,30	0,39	**0,30**	**0,34**
Modèle2	0,34	0,53	0,48	0,58	0,35	0,41	0,42	0,43	0,34	0,38	**0,39**	**0,47**
Modèle3	0,47	0,66	0,60	0,74	0,46	0,55	0,53	0,58	0,44	0,55	**0,50**	**0,62**

Tableau 4. *Identification des titres dans les pseudo documents*

On obtient de meilleurs résultats pour la tâche d'identification des titres. Cela n'est pas surprenant. La F-Mesure du Modèle1 au Modèle2 puis au Modèle3 passe de 0,30 à 0,36 et à **0,47** pour la structuration hiérarchique des titres et de 0,32 à 0,43 et à **0,56** pour l'identification de titres. On prouve ainsi l'importance des statistiques sur les balises HTML (*H*), qui peut s'apparenter à l'information structurelle des titres (Ho-Dac *et al.*, 2004). De plus, l'ajout de règles linguistiques permet un gain

de performance significatif par rapport au seul modèle statistique. Ces règles linguistiques modifiant uniquement le score de visibilité, cela montre également l'importance de la fonction de score de visibilité.

En différenciant les différents corpus, on remarque que celui traitant de l'histoire de France est celui qui obtient les meilleurs résultats. En effet, bien que les historiens écrivent moins de document bien formés, ils semblent utiliser moins de niveaux de titre (le plus souvent seulement <*H1*> et <*H2*>).

Il faut remarquer enfin que cette évaluation porte uniquement sur les pseudo documents. Or on peut s'apercevoir que tous les styles référencés dans une CSS, ne sont pas utilisés dans le document HTML. Ainsi, pour évaluer la structuration et l'identification des titres, il est préférable de le faire sur les documents HTML restructurés (comme décrit en section 2.4).

4.3. *Evaluation sur les documents HTML restructurés*

Cette évaluation à été menée sur une sélection aléatoire de 10 documents. Pour évaluer le système, il a fallu annoter ces 10 documents en ne considérant que la mise en forme de la page Web, ainsi que son contenu (on ne se base pas sur les balises HTML). Dans un premier temps ces documents ont été d'abord nettoyés : les parties non informatives (publicité, menus, *etc.*) ont été supprimées. Pour la tâche de structuration hiérarchique des titres, on obtient une précision de **0,66** et un rappel de **0,74** (soit une F-Mesure de 0,70) alors que pour la tâche d'identification, on obtient une précision de **0,92** et un rappel de **0,81** (soit une F-mesure de 0,86). Cette évaluation finale de notre système est bien meilleure que ce que laissait suggérer l'évaluation sur les pseudo documents. De plus ces résultats sont bien meilleurs que ce obtenus par Hu et al. sur l'identification du seul titre principal des pages Web (dans notre cas, on cherche à identifier tous les titres des pages).

5. Conclusion

Dans cet article, nous avons décrit une méthode pour automatiquement identifier et structurer hiérarchiquement les titres dans les documents HTML. Dans un premier temps, il s'agit d'utiliser un score de visibilité pouvant être attribué à une balise clé HTML en retrouvant dans la ou les feuilles de style associées le style à appliquer au texte encapsulé par cette balise clé (couleur du texte, taille de la police, *etc.*) Tout comme (Hu *et al.*, 2005 ; Song *et al.*, 2004 ; Xue *et al.*, 2007 ; Yang *et al.*, 2001), on se base sur les propriétés visuelles des documents. Dans un second temps, nous avons rajouté à ces propriétés des propriétés de la structure linéaire du texte que l'on peut extraire directement des documents HTML. Il s'agit de considérer seulement les statistiques des balises (fréquence, profondeur dans l'arbre DOM) et non la grammaire HTML afin de ne pas être dépendant d'un domaine (Hattori *et al.*, 2007).

L'évaluation a clairement montré l'importance de ces propriétés. En effet, les titres participent à cette organisation linéaire (Ho-Dac *et al.*, 2004 ; Jacques *et al.*, 2006). Notre modèle est basé sur un modèle de Markov caché qui identifie les changements de niveau de titre et leur hiérarchie. La principale difficulté a été de définir une fonction de score de visibilité la plus efficace possible. Enfin, l'ajout de certaines règles linguistiques simples (comme rejeter par exemple les titres de plus de 80 caractères) permettent d'améliorer sensiblement les résultats. On obtient une F-Mesure finale de 0,70 pour la structuration hiérarchique des titres et de 0,86 pour l'identification des titres.

Dans de futurs travaux, il nous faudra augmenter le nombre de documents pour l'évaluation sur les documents HTML restructurés (qui n'a été faite dans cet article que sur 10 documents). Nous pensons également intégrer une méthode de segmentation de document HTML comme présentée dans (Mukherjee *et al.*, 2003). En effet, cette technique tente d'identifier des séquences récursives de balises dans les arbres DOM. Identifier ces séquences nous permettra de reconnaître les sous-sections des documents et donc les sous-titres et titres plus facilement. On pourrait également étudier plus précisément les différences entre les différents domaines (issus des différentes requêtes posées au moteur de recherche) et évaluer les performances d'un modèle appris sur un domaine et appliqué à un autre. Un autre piste d'amélioration serait de combiner deux modèles de Markov caché appris d'une part sur les balises de visibilité et d'autre part sur la structure spatiale du texte.

6. Bibliographie

Edmundson H. P., « New Methods in Automatic Extracting », *Journal of the ACM*, vol. 16, 1969, p. 264-285.

Hattori G., Hoashi K., Matsumoto K., Sugaya F., « Robust web page segmentation for mobile terminal using content-distances and page layout information », *Proceedings of the 16th international conference on World Wide Web WWW'07*, 2007, p. 361-370.

Ho-Dac L., Jacques M., Rebeyrolle J., « Sur la fonction discursive des titres », *L'unité texte*, 2004, p. 125-152.

Hu Y., Xin G., Song R., Hu G., Shi S., Cao Y., Li H., « Title extraction from bodies of HTML documents and its application to web page retrieval », *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR'05*, 2005, p. 250-257.

Jacques M., Rebeyrolle J., « Titres et structuration des documents », *Actes International Symposium: Discourse and Document ISDD'06*, 2006, p. 1-12.

Marcu D., « From Local to Global Coherence: A Bottom-Up Approach to Text Planning », *Proceedings of the 14th National Conference on Artificial Intelligence AAAI'97*, 1997, p. 629-636.

Mukherjee S., Yang G., Tan W., Ramakrishnan I., « Automatic Discovery of Semantic Structures in HTML documents », *Proceedings of the Seventh International Conference on Document Analysis and Recognition ICDAR'03*, 2003, p. 669-679.

Rabiner L. R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Readings in speech recognition*, 1990, p. 267-296.

Song R., Liu H., Wen J., Ma W., « Learning block importance models for web pages », *Proceedings of the 13th international conference on World Wide Web WWW'04*, 2004, p. 203-211.

Stolcke A., Omohundro S. M., « Hidden Markov Model Induction by Bayesian Model Merging », *Advances in Neural Information Processing Systems 5*, 1992, p. 11-18.

Xue Y., Hu Y., Xin G., Song R., Shi S., Cao Y., Lin C., Li H., « Web page title extraction and its application », *Information Processing and Management: an International Journal*, vol. 43, 2007, p. 1332-1347.

Yang Y., Zhang H., « HTML Page Analysis Based on Visual Cues », *Proceedings of the Sixth International Conference on Document Analysis and Recognition ICDAR'01*, 2001, p. 859-864.

Zou J., Le D., Thoma G. R., « Structure and content analysis for html medical articles: a hidden markov model approach », *Proceedings of the 2007 ACM symposium on Document engineering*, 2007, p. 199-201.